Does Threshold Imprecision in Semantic Entropy Carry Information? A Credal-Set Analysis

Gregory Wheeler
Frankfurt School of Finance & Management
g.wheeler@fs.de

Preprint, November 2026

Abstract

Large language models often answer fluently and wrongly. Semantic entropy flags likely errors by sampling answers, grouping them by meaning, and measuring their spread. The grouping needs a similarity threshold, however, and nothing in the problem fixes it. Imprecise probability can leave such choices open with a set of distributions, a *credal set*. Pairing the answer distribution with each threshold gives one member of a credal set, under which semantic entropy is expected surprisal. Over all members, the least and greatest are coherent lower and upper expectations. Credal decision rules select admissible answers: the cautious and optimistic rules are plurality votes at the strictest and loosest thresholds, and E-admissibility keeps any answer that wins at some threshold. We ask whether admissible answers are more often right, and whether leaving the threshold open improves on agreement at one threshold. On the first, yes: admissible answers are more accurate than rejected ones by 2.98× on TriviaQA, our discovery set, and a pre-registered 1.62× on NQ-Open, short of its 2.0× target. On the second, mostly no (post hoc). The optimistic rule alone has a higher ratio (3.15× and 1.86×, at higher rejection). On five benchmarks, using every threshold changes correctness AUROC by −.004 to +.007 against one threshold chosen on other labeled questions. Most free-form cases (55–59%) cluster identically at every threshold. The plurality can shift when partial matches merge only at loose thresholds, which the answers alone reveal. There an open threshold gains .054 AUROC on TriviaQA (16% of cases), .026 on ARC-Challenge under a lenient scorer, about zero elsewhere. Entropy-range width does not flag free-form errors. The strictest threshold lost .05 to .25 AUROC against 0.50, which did as well as the labeled choice on all but TruthfulQA. An open threshold still needs endpoints, which also matter.

Keywords: Semantic Entropy, Imprecise Probability, Credal Sets, Choice Functions, Uncertainty Quantification, Large Language Models

1 Introduction

Semantic entropy (Farquhar et al., 2024; Kuhn et al., 2023) estimates an LLM's uncertainty by sampling several responses, clustering them by meaning, and taking the entropy of the cluster distribution. The clustering needs a criterion for when two responses mean the same thing. Kuhn et al. (2023) use bidirectional entailment: a natural language inference model classifies each response against the other, and two responses are equivalent when each is classified as entailing the other. A classification reports the most probable of three classes (entailment, neutral, contradiction) and discards the probabilities. We keep them. The similarity of two responses is the mean of the two entailment probabilities, one for each direction. Two responses are linked when their similarity is at

least a threshold, and a cluster is a connected component of the graph of links, so two responses share a cluster when a chain of linked responses joins them. Nothing in the problem fixes the threshold, however. Different thresholds give different clusterings and different entropies, and being linked is not transitive, so no single clustering is canonical (§2.3.3). Two later methods dispense with clustering and aggregate a pairwise similarity measure directly, but each has a free parameter of its own. Kernel Language Entropy (Nikitin et al., 2024) weights a graph on the responses by a natural language inference model's outputs on each pair, forms a kernel on that graph, and takes the kernel's von Neumann entropy, so the kernel and its hyperparameters, such as a heat kernel's lengthscale, must be chosen. Its authors report that default hyperparameters do about as well as hyperparameters selected on a validation set. Semantic Nearest-Neighbor Entropy (Nguyen et al., 2025) divides pairwise similarities by a scale factor before aggregating them, and the scale decides whether the most similar pairs dominate the aggregate or all pairs count about equally. Its authors report that their results are not sensitive to the scale factor at the four values they tried: 0.1, 1, 10, and 100.

Imprecise probability is a standard tool when the available information does not single out a determinate value for a model parameter (Troffaes and de Cooman, 2014; Walley, 1991; Wheeler, 2022). The semantic-entropy threshold is such a parameter: neither the task nor the sampled responses determine it. Rather than commit to a single arbitrary value, we keep a model for every threshold in a range and reason with the whole set, calculating lower and upper expectations that bound the expectation under each model. Decision rules defined on that set then select among the sampled responses. This paper works out that model and tests whether it helps separate correct from incorrect responses.

The construction is simple. Each threshold clusters the same sampled responses, but different thresholds produce different clusters, so their cluster distributions are defined on different outcome spaces. We put them on one space by taking an outcome to be a pair: a sampled response and the threshold at which "same answer" is read. We then consider every joint probability mass function that pairs the response distribution with some distribution over thresholds, with the threshold independent of the response. These form a credal set, a closed convex set of probability mass functions, that is precise on responses, since every member gives the responses their sampled weights, and vacuous on thresholds, since any distribution over thresholds is allowed (§3). Semantic entropy at one threshold is the expected surprisal of a response's cluster under the member that puts all its weight on that threshold, and every other member averages these entropies across thresholds. Loosening the threshold only merges clusters, and merging cannot raise entropy, so the least and greatest expected surprisal over the set are the entropies at the loosest and the strictest thresholds. They are coherent lower and upper previsions (expectations), and we call the interval between them the entropy range.

The standard decision rules of imprecise probability then select among the sampled responses. At each threshold, a response's agreement share is the weight of its cluster: how much of the sample agrees with it. Because loosening only merges clusters, each response's agreement share is greatest at the loosest threshold and least at the strictest. Plurality at a threshold keeps the responses in a largest cluster there, ties included, and we call those responses the plurality. So Γ-maximax, the optimistic rule, which keeps the responses whose best case is greatest, is plurality at the loosest threshold, and Γ-maximin, the cautious rule, which keeps those whose worst case is greatest, is plurality at the strictest. E-admissibility keeps a response if its expected agreement share is greatest under at least one member of the credal set. Under the member that puts all its weight on one threshold, a response's expected agreement share is its agreement share at that threshold, and the responses with the greatest share are those in a largest cluster there. So, taken over these members alone, one for each threshold in the range, E-admissibility keeps exactly the

responses that lie in a largest cluster at some threshold: the union of the pluralities. Taken over the whole credal set, mixtures of thresholds included, it can keep more, but with six equally weighted responses, as in our experiments, the two coincide (§3). We call the responses in the union of the pluralities admissible and the rest rejected. The partition of the sampled responses into admissible and rejected is the one we test. The rule that keeps the admissible responses is a coherent choice function when the agreement shares computed on the full sample are held fixed. Re-running the rule on a subset of the responses re-clusters them and renormalizes their weights, which changes their agreement shares, and there the rule fails Sen's contraction condition (a response admissible in a set stays admissible in any subset that contains it). Plain plurality fails it too, so the failure is not owed to threshold imprecision.

The construction also sets a limit on what the experiments can show. Because the credal set is precise on responses, its members differ only in how they weight the thresholds, so each of these decision rules depends on a response only through its agreement shares across thresholds, its *agreement profile*. And because the partition is a union of pluralities, it can differ from plurality at a single threshold only where the largest cluster changes with the threshold (§3.6). The empirical question about threshold imprecision is therefore narrow: does leaving open what counts as agreement tell us anything about which responses are correct, beyond agreement at one threshold?

We take up two questions on five QA benchmarks, three free-form (TriviaQA, NQ-Open, TruthfulQA) and two multiple-choice (MMLU, ARC-Challenge), with four open base models of 7 to 9 billion parameters (§4). The first is whether the partition separates correct from incorrect responses, measured by the ratio of admissible to rejected accuracy and reported with the share of responses rejected. Three pre-registered studies test it on four of the benchmarks; TriviaQA, on which the effect was first observed, is the discovery set and is not registered. The second is the question about threshold imprecision above, and the comparisons that bear on it are post hoc, on the same data (on MMLU, for one of the four models).

- **The partition separates correct from incorrect responses.** On free-form factual QA the ratio is 2.98× on TriviaQA, the discovery set, with 5.2% of responses rejected, and a registered 1.62× on NQ-Open with 19.9% rejected, short of its 2.0× criterion. On multiple choice the registered ratios are small (1.18× on MMLU and 1.10× on ARC-Challenge): accuracy there is high, in part because the registered scorer credits the gold letter inside any word, so even rejected responses are often scored correct and the ratio has little room above 1. Under a stricter rule ARC-Challenge's ratio is 1.72×, above NQ-Open's, so these data do not establish that free-form benchmarks separate better than multiple-choice ones. As odds ratios the four registered benchmarks are alike, from 1.77 to 2.06, and the discovery set is the outlier at 9.93; its runs used a different prompt. The stricter rule and the odds ratios are post hoc. Of the 20 registered hypotheses, 12 are supported, though three of these, on TruthfulQA, hold only under its substring scorer: rescored post hoc by an LLM judge, TruthfulQA's registered 1.88× falls to 1.08×. Tables 4 and 5 give all twenty with their outcomes.

- **The choice of threshold matters.** Plurality at the loosest threshold, Γ-maximax, separates better than plurality at the strictest, Γ-maximin (post hoc): 3.15× against 2.00× on TriviaQA and 1.86× against 1.44× on NQ-Open, at similar rejection rates. Intervals on the difference exclude zero on four of five benchmarks. On three of the four the two rules reject similar shares of responses; on ARC-Challenge Γ-maximax also rejects more, 15.8% against 9.6%. On MMLU the interval is too wide to separate them. Separation is lost mainly at the strict end, where the similarity of two matching answers often falls below the threshold, so that they do not share a cluster: under the scorers of the registered runs, the AUROC of the agreement share as a predictor of correctness moves by at most .03 between 0.50 and 0.85 on free-form

QA and falls by .05 to .14 at 0.95, and on multiple choice it already falls by .08 to .11 at 0.80 and by .23 to .25 at 0.95, all against 0.50 (post hoc).

- **Leaving it open adds little.** On held-out questions, a response's agreement profile predicts its correctness no better than its agreement share at one threshold chosen on other labeled questions: the gain in AUROC runs across benchmarks from $-.004$ to $+.007$, with no interval reaching above .023, and gradient-boosted models on the profile gain no more, from $-.024$ to $+.013$. Often the sweep, clustering at every threshold in the range, changes nothing: on 55–59% of free-form question-model pairs every threshold gives the same clustering. The plurality can shift only where responses whose similarity lies inside the range merge at loose thresholds but not at strict ones, and the sweep reveals this without labels. On the pairs where it does shift, the profile beats the threshold chosen for all pairs by .054 in AUROC on TriviaQA (16% of its pairs), by .026 on ARC-Challenge, under the lenient registered letter scorer, and by about zero elsewhere.

- **The imprecise outputs add little on their own.** The partition has a lower ratio than Γ-maximax on all five benchmarks, at a lower rejection rate, mostly because of ties: under equal weights, when every response is a singleton at some threshold, all tie for the plurality and all are admitted. Without these tied singletons it matches Γ-maximax on TriviaQA and NQ-Open, and on the other three benchmarks it has a higher ratio at a lower rejection rate (point estimates, on matrices that only partly reproduce the registered partitions). The width of the entropy range does not flag errors on free-form QA (AUROC .479 for majority-vote error on NQ-Open). On the multiple-choice benchmarks it falls with difficulty, against a registered prediction: almost every response is a singleton at the strictest threshold, so the upper end of the entropy range sits at its largest possible value and the width shrinks as the lower end rises. All of this is post hoc except that prediction.

So the threshold matters, but treating it imprecisely adds little: the whole range and a single threshold chosen on other labeled questions predict correctness about equally well. That threshold in turn beat a fixed 0.50 only on TruthfulQA, where every fold chose 0.90 (AUROC .703 against .670); on the other benchmarks the two differ by at most .006 (post hoc). Leaving the threshold open does not remove the choice, because the range's endpoints are hyperparameters too. Across 35 endpoint pairs examined after the fact, the NQ-Open ratio stays above 1 but runs from $1.52\times$ to $2.93\times$; the registered range gives $1.62\times$ at 19.9% rejection, and starting it at 0.40 instead gives $2.78\times$ at 13.3%. These are findings about these data: a range chosen on them would need registering before it is tested on fresh data. The structure of §3.6 explains why there was little to gain: the rules see a response only through its agreement profile, and wherever every threshold gives the same clustering, as on most free-form pairs, that profile is flat. A null over thresholds does not show that useful imprecision lies elsewhere, but it makes the weights on responses a natural next place to look, since imprecise weights would let a rule depart from plurality (§4.7).

1.1 Contributions

A threshold credal set for semantic entropy (§3). The threshold stops being a setting to tune and becomes part of the model, with three consequences. First, existing practice becomes a special case, so what it gives up by fixing the threshold can be asked exactly: semantic entropy at a fixed threshold is the expectation under one extreme point of the credal set, and plurality at the loosest and at the strictest threshold are Γ-maximax and Γ-maximin. Second, the model delivers what a single entropy does not: coherent bounds on entropy, the entropy range; a coherent interval

for each response's agreement share; and a label for each sampled response, admissible or rejected, from a rule that is a coherent choice function when agreement shares are held fixed. Third, the model says in advance what it can show, which is what makes the empirical question of Section 4 narrow. Every rule acts on a response only through its agreement profile. As many thresholds as there are sampled responses, one for each distinct clustering, recover the partition over the whole range, so the partition depends on the range and the data and not on how finely the range is sampled, though where those thresholds fall depends on the data. And with six equally weighted responses, as in our experiments, the partition, E-admissibility over the whole set, and maximality coincide on every family of nested clusterings, by exhaustive enumeration, so our experiments cannot tell these rules apart; with seven they need not coincide. To our knowledge this is the first treatment of the semantic-entropy threshold as a credal set. Other recent work places the imprecision elsewhere: in an ensemble of adapters, with interval dominance deciding when to commit to an answer (Kudukkil Manchingal et al., 2026), or in probability intervals elicited by prompting (Yang et al., 2026).

Registered tests and post hoc comparisons (§4). The evidence for each empirical claim has a stated status, so a reader can tell what was predicted before the data were seen from what was found afterwards. Whether the partition separates correct from incorrect responses was registered: three public registrations, every registered hypothesis reported with its outcome (Tables 4 and 5), 12 of 20 supported (three only under TruthfulQA's substring scorer), and the deviations from the registrations reported, with their effects where they can be measured. Every other analysis in Section 4 is labeled as discovery (TriviaQA), post hoc, or exploratory. The post hoc ones include the LLM-judge rescoring of TruthfulQA, which shows how far the registered result there depends on its scorer, and every analysis of whether threshold imprecision adds anything: the comparisons of decision rules, the tie mechanism behind the partition's low rejection rate, the analyses of how often the clustering changes across the range, the 35 threshold ranges, and a cross-validated test of the agreement profile against a single threshold. These analyses answer the paper's second question and explain the answer, and because they were found on these data, we name the three that should be registered before they are tested on fresh data (§6.5).

One vocabulary for other methods (§5). Current methods report different kinds of object, a distribution, a stated number, a set of models, a prediction set, which makes them hard to compare. Read as imprecise-probability constructs, token probabilities, verbalized confidence, and ensembles can be compared on two counts: what each constructs, and whether the result is coherent. Conformal prediction is set beside them as a complement, since coverage is a different property from coherence. The reading also marks where imprecise probability adds something and where it does not. For these methods it is mostly definitional: it fixes terms, separates coherence from calibration, and locates one label-free check. That check, that verbalized confidences in a claim and its negation should not sum above one, follows from precise coherence as well and is already in use (Burns et al., 2023; Fluri et al., 2023). In our data a link between passing it and calibration appears in one model, at a size that depends on the prompt (Appendix D). The threshold credal set is the one place where the reading yields a new construction.

Other predictions, and how early effects arose (Appendices D and E). We state three further predictions, none registered and two restated after the first results. They say that instruction tuning costs more accuracy where a base model's answers are more diverse, that incoherent verbalized confidences go with worse calibration, and that the temperature sensitivity of accuracy varies with base-model diversity. The first and third are not supported; the evidence on the second

is inconclusive. That limits what the reading can claim beyond the threshold construction. We report the three in full, with the harness defects that first made their effects look real, because these are ordinary faults that other evaluations may share: unmatched arms, a missing chat template, doubled begin-of-text tokens, a base-instruct pair that also differs in pretraining, a truncating scorer, and a prompt that stated the answer before asking for confidence.

1.2 Paper Structure

Section 2 gives the background on semantic entropy and imprecise probability. Section 3 develops the threshold credal set and its decision rules, and Section 4 reports the tests and what their null result says. Section 5 places the other methods in the same vocabulary. Section 6 discusses decisions on credal sets, a diagnostic checklist, limitations, related work, and future directions, and Section 7 concludes. The appendices give proofs (Appendix A), the theory of choice functions (Appendix B), further background (Appendix C), the other predictions (Appendix D), and experimental details (Appendix E).

2 Background

This section sets out semantic entropy and the parts of imprecise-probability theory that the construction of Section 3 uses. Appendix C surveys other LLM uncertainty methods and gives a fuller introduction to imprecise probability; Shorinwa et al. (2025) review LLM uncertainty quantification more broadly.

2.1 Semantic Entropy

Token-level measures, such as the entropy of a model's distribution over token sequences, conflate linguistic variation with uncertainty about the answer: a model that gives the same answer but phrased differently spreads its probability over different token sequences, so entropy is high even though the answer never changes (Appendix C.1.1). Semantic uncertainty methods address this by shifting attention from token sequences to semantic meanings: what matters for uncertainty is how probability distributes across *distinct* answers, not across surface syntactic forms.

The semantic entropy (SE) framework of Kuhn et al. (2023) does this in three steps. First, sample multiple responses from the model. Second, use a natural language inference (NLI) model, a classifier trained to predict whether annotators would label one sentence as entailing, contradicting, or neutral toward another, to decide which responses are semantically equivalent, grouping responses into clusters $C_1, \ldots, C_K$ based on bidirectional entailment. Third, compute the entropy of the distribution over clusters rather than over individual responses:

$$\mathrm{SE} = -\sum_{j=1}^{K} P(C_j) \log P(C_j),$$

where $P(C_j) = \sum_{r \in C_j} p_r$ and the weights p_r on the sampled responses sum to one. Kuhn et al. define semantic entropy over all meanings a model can generate. The form above is the estimate of Farquhar et al. (2024) from the sample, with p_r proportional to the response's sequence probability or, in their discrete variant and in our experiments, equal for every response (§3).

With clusters in place of token sequences, variation in wording no longer counts: if all sampled responses fall in one cluster despite different phrasings, semantic entropy is zero, and it rises only as responses fall into different clusters.

Farquhar et al. (2024) show that semantic entropy detects confabulations, errors where the model gives different wrong answers when resampled, and that it does so better than three baselines: the entropy of token sequences, the model's own answer when asked whether its response is true, and a classifier trained on its hidden states. It cannot detect systematic errors, where the model gives the same wrong answer every time.

Lin et al. (2024) measure semantic uncertainty from a weighted graph on the sampled responses instead of a clustering. The weight on a pair of responses is their similarity, computed from word overlap or from an NLI model. Three measures are read from the graph. The first sums $\max(0, 1 - \lambda_k)$ over the eigenvalues λ_k of the normalized graph Laplacian. Since the multiplicity of the eigenvalue zero is the number of connected components of the graph, the sum is a continuous count of the distinct answers, and it scores the prompt. The other two also score each response: by its degree, the sum of its similarities to the sampled responses, and by its eccentricity, its distance from the centroid of the responses when each is given coordinates from the Laplacian's eigenvectors with the smallest eigenvalues.

Two later methods also work without a clustering. Kernel Language Entropy (KLE) computes the von Neumann entropy of a kernel built from pairwise semantic similarities (Nikitin et al., 2024). Semantic Nearest-Neighbor Entropy (SNNE) aggregates pairwise similarities with a LogSumExp (Nguyen et al., 2025). Semantic entropy, KLE and SNNE return one score per prompt. The degree and eccentricity measures of Lin et al. (2024) rank the individual responses, but a ranking divides the sample into responses to keep and responses to reject only once a cutoff is chosen, and none of these methods supplies one.

Methods that cluster need a criterion for when two responses count as the same answer, and nothing in the problem fixes one. The released implementation of semantic entropy (Farquhar et al., 2024) uses the NLI model's classifications and compares each response with a single representative of each earlier cluster, so a response joins a cluster only if it is equivalent to that representative. We proceed differently. The similarity of two responses is the mean of the NLI model's two entailment probabilities, one for each direction, as in the entailment graph of Lin et al. (2024); two responses are linked when their similarity is at least a threshold θ; and the clusters are the connected components of the graph of links (§3). KLE and SNNE have no threshold, and must choose a kernel and its hyperparameters, or a scale factor, in its place. Section 2.3.3 shows on an example how the clusters change with the threshold and why, since being linked is not transitive, no single clustering is canonical.

2.2 Imprecise Probability in Brief

Imprecise probability covers mathematical models of uncertainty that do not require sharp numerical probabilities (Augustin et al., 2014; Troffaes and de Cooman, 2014; Walley, 1991). On its behavioral interpretation, which generalizes the Bayesian theory of fair prices, the most one would pay for a gamble may fall below the least one would sell it for (Wheeler, 2022). A common construction is in terms of a *credal set*, a closed convex set of probability distributions, and that is the construction we use. Section 3 uses three objects of the theory: credal sets, coherent lower previsions, and choice functions. This section defines the first two and Section 2.3 the third; Appendix C.2 gives the history.

2.2.1 Credal Sets and Lower and Upper Previsions

Definition 1 (Credal Set)**.** *A* credal set $\mathcal{M}$ *is a non-empty, closed, convex set of probability mass functions on a finite outcome space* Ω*.*

A *gamble* f is a bounded real-valued function on Ω, read as an uncertain reward: if ω obtains, the reward is $f(\omega)$. A credal set bounds the expectation of every gamble.

Definition 2 (Lower and Upper Prevision)**.** *For a gamble f and a credal set $\mathcal{M}$,*

$$\underline{P}(f) = \inf_{P \in \mathcal{M}} \mathbb{E}_P[f], \qquad \overline{P}(f) = \sup_{P \in \mathcal{M}} \mathbb{E}_P[f]$$

are the lower prevision *and the* upper prevision *of f under $\mathcal{M}$.*

Lower and upper previsions are lower and upper expectations, and each determines the other: $\overline{P}(f) = -\underline{P}(-f)$. An event $A \subseteq \Omega$ enters through its indicator $\mathbf{1}_A$, the gamble that pays 1 if A occurs and 0 otherwise. The lower and upper previsions of $\mathbf{1}_A$ are the *lower and upper probability* of A, written $\underline{P}(A)$ and $\overline{P}(A)$. Events are a special case, and not one from which the rest can be recovered: two credal sets can give every event the same lower and upper probability and still give some gamble different lower previsions (Figure 3 in Appendix C). So the theory is stated for gambles, and so is our construction, which needs the lower and upper previsions of a gamble that is not an indicator, the surprisal of a response's cluster (Proposition 9). The *imprecision*, or *width*, of a gamble f is $\overline{P}(f) - \underline{P}(f)$.

2.2.2 Coherence and Desirable Gambles

A credal set is one way to arrive at lower and upper previsions. They can also be assessed directly, as the prices at which one would buy and sell gambles, and then not every assessment is consistent. Some assessments can be combined into transactions that lose whatever happens, and some imply sharper prices than they state. *Coherence* rules out both. A lower prevision $\underline{P}$ defined on all gambles is *coherent* if, for all gambles f and g and every $\lambda > 0$ (Walley, 1991, Theorem 2.5.5):

LP1. $\underline{P}(f) \geq \inf f$. (Bounds)

LP2. $\underline{P}(f + g) \geq \underline{P}(f) + \underline{P}(g)$. (Super-additivity)

LP3. $\underline{P}(\lambda f) = \lambda\, \underline{P}(f)$. (Non-negative homogeneity)

Section 3 also needs the object from which these conditions are derived. A *desirable gamble* is one an agent would accept in exchange for the status quo, and the set $\mathcal{D}$ of gambles the agent finds desirable represents the agent's assessments. Axioms of the following kind for sets of desirable gambles go back to Williams (1975, 1976); Walley (1991) later developed the comprehensive theory of coherent lower previsions on these foundations, and de Cooman et al. (2023) and Walley (2000) develop desirability further. Gambles are compared pointwise: $f \leq 0$ means $f(\omega) \leq 0$ for every outcome ω. For gambles f, g and $\lambda > 0$:

A1. If $f \leq 0$, then $f \notin \mathcal{D}$. (Avoid partial loss)

A2. If $f \geq 0$ and $f \neq 0$, then $f \in \mathcal{D}$. (Accept partial gain)

A3. If $f \in \mathcal{D}$, then $\lambda f \in \mathcal{D}$. (Positive scaling)

A4. If $f, g \in \mathcal{D}$, then $f + g \in \mathcal{D}$. (Combination)

Axioms A1 and A2 are rationality conditions; A3 and A4 are closure operations. A3 extends each desirable gamble to a ray and A4 combines the rays, so together they generate a convex cone. A set $\mathcal{D}$ satisfying A1–A4 is *coherent*.

Lower and upper previsions are derived from $\mathcal{D}$. The lower prevision of a gamble f is the supremum price the agent would pay for it,

$$\underline{P}(f) = \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}\},$$

and the upper prevision is the infimum price at which the agent would sell it, $\overline{P}(f) = -\underline{P}(-f)$. The agent accepts $f - \mu$ for every μ below $\underline{P}(f)$ and need not accept it at $\underline{P}(f)$ itself, which is why these are a supremum and an infimum and not a maximum and a minimum. A lower prevision derived from a coherent $\mathcal{D}$ satisfies LP1–LP3.

Theorem 3 (Lower Envelope Theorem (Walley, 1991, Theorem 3.3.3))**.** *A lower prevision $\underline{P}$ defined on all gambles is coherent if and only if there exists a closed, convex credal set $\mathcal{M}$ such that for all gambles f:*

$$\underline{P}(f) = \inf_{P \in \mathcal{M}} \mathbb{E}_P[f]$$

Equivalently, $\overline{P}(f) = \sup_{P \in \mathcal{M}} \mathbb{E}_P[f]$.

The theorem joins the two routes. The lower prevision of Definition 2, computed over a credal set, is coherent; and every coherent lower prevision, however it was assessed, is the lower prevision of some credal set.

Section 5.2 tests a model's stated confidences against coherence in the case of events. Restricted to events A and B, coherent lower and upper probabilities satisfy:

1. **Bounds:** $0 \leq \underline{P}(A) \leq 1$ and $0 \leq \overline{P}(A) \leq 1$
2. **Conjugacy:** $\underline{P}(A) + \overline{P}(\neg A) = 1$
3. **Superadditivity:** For disjoint A, B: $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$

Miranda (2008) surveys the theory.

2.2.3 Extreme Cases

Credal sets range from maximally precise to maximally imprecise:

- **Precise:** $\mathcal{M} = \{P\}$ contains exactly one distribution. Then $\underline{P}(f) = \overline{P}(f) = \mathbb{E}_P[f]$ for all gambles f. Standard probability is a special case of imprecise probability.
- **Vacuous:** $\mathcal{M}$ contains every distribution. Membership in that set holds in all cases, so it constrains nothing: $\underline{P}(f) = \inf f$ and $\overline{P}(f) = \sup f$ for every gamble f, and every event other than the impossible and the certain one has lower probability 0 and upper probability 1.

The threshold credal set of Section 3 combines the two: it is precise on responses and vacuous on thresholds.

Current LLM uncertainty methods mostly report points. Token-level softmax outputs are single distributions, and Bayesian deep learning (Gal and Ghahramani, 2016) maintains uncertainty over model parameters but marginalizes to a single posterior predictive distribution. Both are precise. We call a credal set with one member *degenerate*. Whether the width lost by committing to a point carries information is an empirical question, which Section 4 takes up for the threshold credal set.

2.3 Choice Functions and the Threshold Problem

Lower and upper previsions bound the expectation of a gamble, but they do not settle which of several gambles to choose. When the bounds of two gambles overlap, neither need be preferred to the other, and several options may stand unbeaten. Choice is therefore modeled by a separate object. A *choice function* C assigns to each finite set of options A a non-empty subset $C(A) \subseteq A$, the options that are admissible in A. (The options in $A \setminus C(A)$ are rejected.) A choice function need not pick out a single best option but instead records which options survive comparison, leaving the choice among them open.

Section 3 uses choice functions for the following reason. Semantic entropy is computed from clusters of responses, and the clusters depend on a threshold that nothing in the problem fixes. A credal set gives lower and upper previsions that bound the entropy at every threshold in a range. Those bounds are a single pair of numbers for the whole sample of responses to a question, and they do not say which of the responses to keep and which to reject. For that, we take the sampled responses as the options of a choice function.

This subsection gives the definitions that Section 3 uses and then states the threshold problem on an example. The general theory of choice functions is developed in de Bock and de Cooman (2019), Seidenfeld et al. (2010), and Van Camp (2018), and Appendix B gives the axioms.

2.3.1 Choice Functions and Decision Criteria

Imprecise probability supplies criteria for producing the admissible subset $C(A)$ from a credal set $\mathcal{M}$ (Troffaes, 2007). The options are gambles, so the expected utility of an option under P is its expectation, and its lower and upper expected utilities are its lower and upper previsions. For an option $a \in A$:

- **E-admissibility** (Levi, 1980): a is admissible if it maximizes expected utility under at least one $P \in \mathcal{M}$.
- **Maximality** (Walley, 1991): a is admissible if no single $b \in A$ has higher expected utility than a under every $P \in \mathcal{M}$.
- **Interval dominance**: a is admissible if no $b \in A$ has a lower expected utility above the upper expected utility of a.
- **Γ-maximin** (Berger, 1985, §4.7.6; Gilboa and Schmeidler, 1989): select the options with the greatest lower expected utility.
- **Γ-maximax** (Satia and Lave, 1973): select the options with the greatest upper expected utility.

Every E-admissible option is maximal, and every maximal option survives interval dominance. E-admissibility rejects a when, under each $P \in \mathcal{M}$, some alternative has higher expected utility, even if the alternative varies with P. Maximality rejects a only when one alternative has higher expected utility under every P. Γ-maximin ranks options by their worst case over $\mathcal{M}$ and Γ-maximax by their best.

2.3.2 Desirable Option Sets and Representation

The criteria above differ in the comparisons they need. Maximality compares two options at a time: it rejects a from A only when some one alternative is better than a under every $P \in \mathcal{M}$. A

set of desirable gambles $\mathcal{D}$ can express such a comparison, since b is preferred to a exactly when the gamble $b - a$ belongs to $\mathcal{D}$. E-admissibility needs more. It rejects a when under each P some alternative is better, although no one alternative need be better under every P. The judgment behind that rejection is a disjunction: at least one of the gambles $b - a$, with $b \in A$, is desirable, and which one is left open. A set of desirable gambles cannot express a disjunction, however, because it says of each gamble separately whether it is desirable.

To address this shortcoming, de Bock and de Cooman (2019) take sets of gambles as the objects of assessment. An *option set* is a finite set of gambles, and an option set is *desirable* when at least one of its members is desirable. A set of desirable option sets $\mathcal{K}$ collects the option sets so judged. The set $\mathcal{K}$ determines a choice function: an option a is rejected from A when the option set $\{b-a : b \in A\}$ belongs to $\mathcal{K}$. A set of desirable gambles $\mathcal{D}$ gives the special case $\mathcal{K}_{\mathcal{D}} = \{A : A \cap \mathcal{D} \neq \emptyset\}$, in which an option set is desirable exactly when it contains a gamble in $\mathcal{D}$. Coherence for a set $\mathcal{K}$ is defined by axioms that extend A1–A4 (Appendix B), and a choice function is coherent when a coherent $\mathcal{K}$ determines it.

Theorem 4 (Representation (de Bock and de Cooman, 2019, Theorem 9))**.** *A set of desirable option sets $\mathcal{K}$ is coherent if and only if*

$$\mathcal{K} = \bigcap_i \mathcal{K}_{\mathcal{D}_i}$$

for some non-empty family of coherent sets of desirable gambles $\mathcal{D}_i$.

The theorem says that a coherent $\mathcal{K}$ behaves as a family of coherent sets of desirable gambles consulted together. An option set is desirable when every member of the family finds some gamble in it desirable, and the gamble may differ from one member to the next. In Section 3 the family has one member for each threshold (Theorem 14).

Coherent choice functions satisfy Sen's contraction condition α (de Bock and de Cooman, 2019; Sen, 1971): an option that is admissible in a set remains admissible in every subset of it that contains the option. Section 3 uses the condition twice. Its rule for selecting responses satisfies the condition when the quantities it compares are held fixed (Theorem 14), and violates it when the rule is run again on a subset (Proposition 15).

2.3.3 The Threshold Problem

We return now to semantic entropy. Its clusters depend on a threshold (§2.1): two responses are linked when their similarity is at least θ, and the clusters are the connected components of the graph of links. An example shows what changes with θ and what does not.

Consider an LLM answering "What causes the seasons?" with five distinct sampled responses:

- r_A: "The tilt of Earth's axis"
- r_B: "Earth's axial tilt of 23.5 degrees"
- r_C: "The angle of sunlight throughout the year"
- r_D: "Earth's distance from the Sun"
- r_E: "The elliptical orbit around the Sun"

Write $s(r, r')$ for the similarity of responses r and r', and suppose that $s(r_A, r_B) = 0.92$, $s(r_B, r_C) = 0.83$, $s(r_A, r_C) = 0.71$, $s(r_D, r_E) = 0.88$, and that every other pair has similarity near zero.

The clusters change with the threshold. At $\theta = 0.90$ only r_A and r_B are linked, and the responses form four clusters: $\{r_A, r_B\}$, $\{r_C\}$, $\{r_D\}$, and $\{r_E\}$. At $\theta = 0.85$, r_D and r_E are linked as well, which leaves three clusters: $\{r_A, r_B\}$, $\{r_C\}$, and $\{r_D, r_E\}$. At $\theta = 0.75$, r_B and r_C are linked too, which leaves two: $\{r_A, r_B, r_C\}$ and $\{r_D, r_E\}$. Semantic entropy is the entropy of a distribution over clusters, so each of these thresholds gives a distribution over a different set of outcomes. Distributions over different sets of outcomes cannot be averaged, or gathered into one credal set, as they stand.

Being linked is not transitive. At $\theta = 0.80$, r_A is linked to r_B and r_B to r_C, but r_A is not linked to r_C, since $s(r_A, r_C) = 0.71 < 0.80$. A clustering is a partition, however, and a partition treats sharing a cluster as transitive. Taking connected components therefore puts r_A and r_C in one cluster at $\theta = 0.80$, although the two are not linked there. The clustering at one threshold does not record that they were joined through r_B. The clusterings at all thresholds together record more, though not everything: they give, for each pair of responses, the greatest threshold at which the two share a cluster. For r_A and r_C that threshold is 0.83, the smaller of the two similarities on the chain through r_B, and not their own similarity of 0.71.

What varies, and what does not. The clusters differ across thresholds, but the weights on the responses do not. Each threshold coarsens the same distribution over responses in a different way: it changes which responses count as the same answer, and it does not change how probable any response is. Section 3 uses this fact to place the distributions at all thresholds on one fixed outcome space.

A question about responses. Semantic entropy asks how the sample is distributed over meanings, and its answer changes with the threshold. A choice function allows a question about the individual responses whose answer does not depend on any one threshold: which responses lie outside the largest cluster at every threshold in a stated range? Those responses are rejected, and the rest are admissible.

Definition 5 (Admissibility)**.** *A response is* admissible *if, for at least one threshold in a stated range, it lies in a largest (highest-probability) cluster, ties included. It is* rejected *if at every threshold in the range its cluster has smaller probability than the largest cluster. The cluster that beats it may vary with the threshold.*

In the example, take the range $[0.70, 0.90]$, and suppose that ten samples contain r_A three times, r_B, r_C, and r_D twice each, and r_E once, each sample with weight $1/10$. The highest-probability cluster is $\{r_A, r_B, r_C\}$ at $\theta = 0.70$ and 0.80 (weight $7/10$) and $\{r_A, r_B\}$ at $\theta = 0.85$ and 0.90 (weight $5/10$). The admissible set is $\{r_A, r_B, r_C\}$, and r_D and r_E are rejected. Section 3 shows that the rule of Definition 5 is E-admissibility over the extreme points of a credal set on one outcome space. We call the resulting division of the responses into admissible and rejected the admissible partition, and Table 1 sets it beside fixed-threshold semantic entropy and the entropy range.

Approach	Threshold uncertainty	Clusterings	Output
Fixed-θ SE	Not represented	One	Point entropy
SE range over θ	Within a stated range	A nested family	Entropy range
Admissible partition	Within a stated range	A nested family	Response-level labels

Table 1: Comparison of approaches to semantic uncertainty; §3 develops the last row.

3 A Threshold Credal Set for Semantic Entropy

Semantic entropy depends on a threshold that nothing in the problem fixes, and the clusterings at different thresholds cannot be compared as they stand, because each gives a distribution over a different set of clusters (§2.3.3). This section removes that obstacle and then leaves the threshold open. We first place the cluster distributions at all thresholds on one outcome space. There they form a credal set, and fixing a threshold amounts to selecting one of its members. From the credal set we obtain lower and upper previsions, which bound semantic entropy and each response's agreement with the rest of the sample across the range of thresholds, and decision rules, which divide the sampled responses into admissible and rejected without committing to one threshold. We then ask what the construction guarantees: how many thresholds it needs, in what sense its rule for selecting responses is coherent, and how far its output can differ from what a single threshold already gives.

3.1 Review of Semantic Entropy

We restate the three steps of §2.1 in the notation used below:

1. Sample M responses $\{r_1, \ldots, r_M\}$ from the LLM
2. Group the responses into clusters $\{C_1, \ldots, C_K\}$ by meaning
3. Compute the cluster probabilities $P(C_j) = \sum_{r \in C_j} p_r$ and return their entropy, $\mathrm{SE} = -\sum_j P(C_j) \log P(C_j)$

The weights p_r are strictly positive and sum to one. Our experiments use *equal-per-sample (Monte Carlo)* weights, $p_r = 1/M$, over the $M = 6$ retained responses. This does not assume that answers are equally probable. A frequent answer is sampled often, so its cluster is large. Our retained samples come from the two lowest temperatures (Appendix E.3), so the estimate is of that temperature mixture, not of the distribution at one temperature. Normalized generation probabilities can replace $1/M$ where they are available.

3.2 The Threshold Credal Set

The threshold enters at the second step. Two responses r and r' are linked at threshold θ when $s(r, r') \geq \theta$, where s is the similarity of §2.1, which is symmetric, and the clusters at θ are the connected components of the graph of links. Write P_θ for the distribution over those clusters, in which the probability of a cluster is the sum of the weights of its responses, and $C^\theta(r)$ for the cluster that contains r. The *agreement share* of r at θ is $u_\theta(r) = P_\theta(C^\theta(r))$, the probability of its cluster. It is the share of the sample, by weight, that gives the same answer as r when "same answer" is read at θ. Lowering θ only adds links, so clusters can only merge. Four consequences follow.

Proposition 6 (Nesting)**.** *For $\theta < \theta'$, the θ-clustering coarsens the θ'-clustering. Hence $u_\theta(r) \geq u_{\theta'}(r)$ for every response r, the largest cluster probability is non-increasing in θ, $SE(\theta) \leq SE(\theta')$, and at most M distinct clusterings occur over any range of thresholds.*

As §2.3.3 shows, the distributions P_θ over clusters have different outcome spaces and cannot be mixed directly. But they are the same distribution on responses, coarsened in different ways. So the family $\{P_\theta\}$ can be stated on one fixed outcome space, by taking an outcome to be a response together with a threshold.

Definition 7 (Threshold Credal Set)**.** *Fix responses $\mathcal{R} = \{r_1, \ldots, r_M\}$ with weights $\{p_r\}$ (repeated samples may be kept separately, or merged with summed weight if their similarity is at least $\theta_{\max}$), a symmetric similarity s, and a non-empty finite set of thresholds Θ with least element $\theta_{\min}$ and greatest element $\theta_{\max}$. An outcome is a pair $(\omega, \theta) \in \mathcal{R} \times \Theta$: a sampled response together with the reading of "same answer" fixed by θ. Each response r defines a gamble*

$$g_r(\omega, \theta) = \mathbf{1}[\omega \in C^\theta(r)],$$

which pays 1 when the sampled response ω shares a cluster with r at θ and 0 otherwise. For a distribution λ on Θ, write $p \otimes \lambda$ for the product distribution that gives the outcome (ω, θ) probability $p_\omega \lambda(\theta)$. The threshold credal set *is $\mathcal{M}_\Theta = \{p \otimes \lambda : \lambda$ a distribution on $\Theta\}$.*

Three features of $\mathcal{M}_\Theta$ follow from its definition. Every member gives the responses their weights p, so the set is precise on responses. Any distribution λ over thresholds is allowed, so it is vacuous on thresholds. And every member is a product, so the threshold is taken to be independent of the sampled response. Write δ_θ for the distribution that puts all its weight on θ. Under the member $p \otimes \delta_\theta$ the expectation of g_r is the agreement share $u_\theta(r)$, the probability that a sampled response agrees with r at θ, and under a general member it is the mixture $\mathbb{E}_{p\otimes\lambda}[g_r] = \sum_\theta \lambda(\theta) u_\theta(r)$. The imprecision concerns what counts as agreement, not how probable any response is.

The first use of the credal set is to bound, for each response, how much of the sample agrees with it.

Proposition 8 (Response Intervals)**.** *For each response r, the lower and upper previsions of g_r under $\mathcal{M}_\Theta$ are*

$$\underline{P}(g_r) = \min_\theta u_\theta(r) = u_{\theta_{\max}}(r), \qquad \overline{P}(g_r) = \max_\theta u_\theta(r) = u_{\theta_{\min}}(r),$$

and they are coherent.

We call $[\underline{P}(g_r), \overline{P}(g_r)]$ the *response interval* of r. Its width is how far the agreement share of r depends on the threshold. Committing to one threshold replaces $\mathcal{M}_\Theta$ by the singleton $\{p \otimes \delta_\theta\}$, a degenerate credal set. The singleton has the same status as the single softmax distribution of §5.1: a precise commitment where nothing in the problem fixes one. We do not claim that $\mathcal{M}_\Theta$ is the only representation of the problem. It suffices to place semantic entropy in the same pattern as the methods of Section 5.

3.3 Semantic Entropy Bounds

The same credal set bounds semantic entropy. The gamble whose expectation is semantic entropy is the surprisal of the cluster of the sampled response.

Proposition 9 (Semantic Entropy Bounds)**.** *Let $h(\omega, \theta) = -\log u_\theta(\omega)$ be the surprisal gamble. Then $\mathbb{E}_{p\otimes\delta_\theta}[h] = SE(\theta)$, and the lower and upper previsions of h under $\mathcal{M}_\Theta$ are $\underline{SE} = SE(\theta_{\min})$ and $\overline{SE} = SE(\theta_{\max})$.*

So semantic entropy at a fixed threshold is the expected surprisal under one member of $\mathcal{M}_\Theta$, and the least and greatest expected surprisal over the set are the entropies at the loosest and the strictest threshold. We call $[\underline{\mathrm{SE}}, \overline{\mathrm{SE}}]$ the *entropy range*. Its endpoints are coherent lower and upper previsions, its width is how far semantic entropy depends on the threshold, and two thresholds suffice to compute it. Entropy is not linear in the distribution, but that causes no difficulty here:

the weights on responses are precise, so the surprisal gamble is fixed and only its expectation varies. It would matter if the weights were imprecise as well.

A related scalar is $1 - \max_\theta \max_k P_\theta(C_k^\theta)$, where C_k^θ ranges over the clusters at θ: one minus the greatest cluster probability at any threshold. It is one minus the Γ-maximax value $\max_r \overline{P}(g_r)$, and by Proposition 6 it equals $1 - \max_k P_{\theta_{\min}}(C_k^{\theta_{\min}})$. So it depends on $\theta_{\min}$ alone and does not combine information across thresholds. Section 4.4 reports it as a secondary diagnostic, not as a measure of uncertainty in its own right.

3.4 Decision Rules and the Admissible Partition

The response intervals and the entropy range are bounds, and bounds do not select among the responses. For that we apply the decision rules of §2.3.1, with the sampled responses as the options and the gamble g_r as the reward for choosing r. We begin with the rule that this paper tests, which keeps every response that lies in a largest cluster at some threshold.

Definition 10 (Semantic Choice Function)**.** *For $\theta \in \Theta$, let $O_\theta^* = \{r : u_\theta(r) = \max_{r'} u_\theta(r')\}$ be the plurality at θ: the responses in a largest cluster there, all tied clusters included. The* admissible set *is*

$$C_{sem}(\mathcal{R}) = \bigcup_{\theta \in \Theta} O_\theta^*,$$

and $\mathcal{R} \setminus C_{sem}(\mathcal{R})$ is the rejected *set.*

The admissible set is the union of the pluralities across the range, and with the rejected set it divides the sampled responses in two. This is the admissible partition of §2.3.3. It is a different kind of output from semantic entropy, KLE, and SNNE, each of which returns one score for the whole sample and labels no response. Plurality at one threshold does label responses, and the admissible partition extends it: a response is kept if it is in the plurality at any threshold in the range. Figure 1 gives the algorithm, and Figure 2 shows it on the example of §2.3.3. The first property says what admission and rejection amount to.

Proposition 11 (Pointwise Dominance)**.** *Every $r \in C_{sem}(\mathcal{R})$ attains $\max_{r'} u_\theta(r')$ at some $\theta \in \Theta$. Every $r \notin C_{sem}(\mathcal{R})$ satisfies $u_\theta(r) < \max_{r'} u_\theta(r')$ at every $\theta \in \Theta$, and the response attaining the maximum can be taken in $C_{sem}(\mathcal{R})$.*

Both claims follow directly from the definition. Note what the second does not say. The response that beats a rejected response may be a different one at each threshold, so a rejected response need not be beaten by any one alternative throughout the range. The next proposition places the admissible set among the standard rules of §2.3.1, where that difference separates E-admissibility from maximality.

Proposition 12 (Decision Rules)**.** *Under $\mathcal{M}_\Theta$, let* EA *contain each r that maximizes $\mathbb{E}_P[g_r]$ for some $P \in \mathcal{M}_\Theta$, let* Max *contain each r for which no single r' satisfies $u_\theta(r') > u_\theta(r)$ for all $\theta \in \Theta$, and let* ID *contain each r for which no r' satisfies $\min_\theta u_\theta(r') > \max_\theta u_\theta(r)$. Then Γ-maximax selects $O_{\theta_{\min}}^*$, Γ-maximin selects $O_{\theta_{\max}}^*$, and*

$$O_{\theta_{\min}}^* \cup O_{\theta_{\max}}^* \subseteq C_{sem}(\mathcal{R}) \subseteq \text{EA} \subseteq \text{Max} \subseteq \text{ID}.$$

The proposition identifies three familiar procedures with standard rules on $\mathcal{M}_\Theta$. Plurality at the loosest threshold is Γ-maximax, and plurality at the strictest is Γ-maximin, because the upper prevision of each g_r is the agreement share of r at $\theta_{\min}$ and its lower prevision is the share at $\theta_{\max}$

(Proposition 8). The admissible set C_{sem} is E-admissibility restricted to the extreme points $p \otimes \delta_\theta$ of $\mathcal{M}_\Theta$, since under $p \otimes \delta_\theta$ the responses of greatest expectation are the plurality at θ. E-admissibility as Levi defines it ranges over the whole convex set, mixtures of thresholds included, and can admit more. One might expect the admissible set to be the Pareto frontier of the agreement shares $\{u_\theta\}$, taken as objectives. It is not: C_{sem} and EA can each contain a Pareto-dominated response and omit a Pareto-optimal one. The set Max is the weak Pareto frontier (Appendix A.2).

How far these rules can differ depends on the number of responses. With six equally weighted responses, as in our experiments, C_{sem}, E-admissibility over the whole set, and Max coincide: over all 36,047 families of nested clusterings of six responses, C_{sem} equals Max, and E-admissibility lies between them (exhaustive enumeration). With seven responses they can differ: if one threshold merges only a pair and a looser one splits the responses four and three, with the pair inside the three, the third member of that block is maximal but in no largest cluster. So the agreement of these rules in §4.3 is forced by the sample size, not found in the data.

The admissible set is defined on a finite grid Θ, and one may ask whether it depends on how finely the range is sampled. It does not, provided the grid meets every clustering. To state the result, assume $\theta_{\min} < \theta_{\max}$. Sort $\theta_{\min}$, $\theta_{\max}$, and the pairwise similarities in $[\theta_{\min}, \theta_{\max}]$ as $b_0 < \cdots < b_L$. Call the sets $(b_k, b_{k+1}]$ the *cells* of the range, and add $\theta_{\min}$ to the first cell, or make $\{\theta_{\min}\}$ a separate cell if some similarity equals $\theta_{\min}$.

Theorem 13 (Grid Invariance)**.** *The clustering, O^*_θ, and $SE(\theta)$ are constant on each cell. Any two finite grids in $[\theta_{\min}, \theta_{\max}]$ that meet every clustering yield the same C_{sem}. There are at most $\binom{M}{2}+1$ cells, and at most M distinct clusterings among them, so M thresholds, one per clustering, recover the admissible set over the whole interval $[\theta_{\min}, \theta_{\max}]$. A grid that misses a clustering admits a subset of what a grid meeting every clustering admits.*

A fixed grid carries this guarantee only if it meets every clustering, and the thresholds at which the clustering changes depend on the data. The theorem also does not remove dependence on the range. If $\theta_{\min}$ is low enough that all responses merge, everything is admissible; under equal weights the same holds if all responses are singletons at $\theta_{\max}$, since they tie. The range is therefore a hyperparameter, and endpoints of either kind should be avoided. Even inside a fixed range, an all-singleton clustering can occur for a particular question, and the rule as registered does not exclude it; Table 8 shows the consequence.

The last two results concern coherence. Section 2.3.2 gave the test: a choice function is coherent when a coherent set of desirable option sets determines it, and a coherent choice function satisfies Sen's α. The admissible set passes when the agreement shares are computed once, on the full sample. It fails when they are computed again on each subset.

Theorem 14 (Coherence with Fixed Agreement Shares)**.** *For $\theta \in \Theta$, let $\mathcal{D}_\theta = \{f : \mathbb{E}_{p\otimes\delta_\theta}[f] > 0\} \cup \{f : f \geq 0, f \neq 0\}$, for gambles on $\mathcal{R} \times \Theta$. Each $\mathcal{D}_\theta$ is coherent, so $\mathcal{K}_\Theta = \bigcap_{\theta\in\Theta} \mathcal{K}_{\mathcal{D}_\theta}$ is a coherent set of desirable option sets. On sets $A \subseteq \mathcal{R}$ of responses, each offered as its gamble g_r and with r admissible when g_r is, the choice function it induces is*

$$C_\Theta(A) = \bigcup_{\theta\in\Theta} \arg\max_{r\in A} u_\theta(r),$$

with the agreement shares u_θ computed on all of $\mathcal{R}$. Thus $C_\Theta(\mathcal{R}) = C_{sem}(\mathcal{R})$, and C_Θ satisfies Sen's α.

Proposition 15 (Contraction Failure on Re-Running)**.** *For $A \subseteq \mathcal{R}$, let $C_{sem}(A)$ be the construction run on A alone, with the same s and Θ and with weights renormalized over A. This map violates Sen's α (Sen, 1971): there exist $\mathcal{R}$, s, Θ, and $A \subseteq \mathcal{R}$ with $r \in A$ such that $r \in C_{sem}(\mathcal{R})$ but $r \notin C_{sem}(A)$.*

Algorithm: Admissible Partition

1. **Input:** Responses $\{r_1, \ldots, r_M\}$, weights $\{p_r\}$, similarity function s, grid Θ with least element $\theta_{\min}$ and greatest element $\theta_{\max}$
2. **Initialize:** Admissible $\leftarrow \emptyset$; compute $S_{ij} = s(r_i, r_j)$ once
3. **For each** $\theta \in \Theta$:
   (a) Build a graph with edge (i, j) if $S_{ij} \geq \theta$
   (b) Find connected components $\{C_1, \ldots, C_K\}$
   (c) Compute $P_\theta(C_k) = \sum_{r \in C_k} p_r$
   (d) Add every response in a maximum-probability cluster to Admissible, including ties
4. **Output:** Admissible set; Rejected $= \{r_1, \ldots, r_M\} \setminus$ Admissible

Complexity: $O(|\Theta| \cdot M^2)$ after the similarity matrix.

Figure 1: Algorithm for the admissible partition. Step (b) imposes transitivity at each threshold.

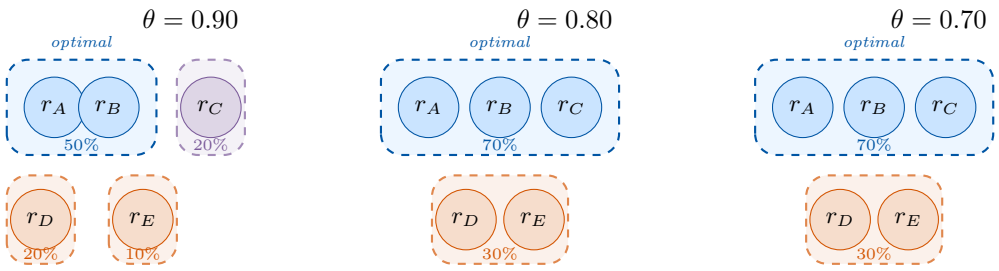
Status of the construction. Theorem 14 and Proposition 15 differ in what changes. With agreement shares fixed, the set of options shrinks and the evidence stays, and the rule is coherent. Re-running on a subset shrinks the evidence as well: removing a response can shrink another response's cluster, and α fails. Since coherent choice functions satisfy α, no coherent set of desirable option sets reproduces the re-run map, under any assignment of options to responses. That failure already occurs at $|\Theta| = 1$, so it belongs to plain plurality as well and not to threshold imprecision. For filtering a fixed sample, the fixed-share reading is the relevant one; re-running asks what would follow had the other samples not been drawn. Sen's α also holds for the re-run map on subsets that are unions of $\theta_{\min}$-clusters (Appendix A.2).

The coherence in Theorem 14 has two limits. It does not favor using every threshold: the same argument makes any sub-grid coherent, including plain plurality at one threshold. And it is axiomatic, not behavioral: no procedure reveals θ, so these gambles are never settled, and the betting reading of §2.2.2 does not carry over directly.

Existing practice as the degenerate case. A credal set with one member is the common case in current methods. Plurality at a fixed θ keeps the largest cluster and rejects the rest, which is Definition 10 with $|\Theta| = 1$ and the Bayes rule under $p \otimes \delta_\theta$. A self-consistency partition (Wang et al., 2023) does the same with exact string groups in place of NLI clusters. These methods make the threshold commitment implicitly. The threshold credal set makes it explicit, and Section 4 tests what relaxing it does.

3.5 Worked Example

We return to the example of §2.3.3: the question "What causes the seasons?", its five responses r_A to r_E with their similarities, and the ten samples in which r_A occurs three times, r_B, r_C, and r_D twice each, and r_E once. Table 2 gives the number of clusters, the largest cluster probability, and the entropy at four thresholds.



Result: Admissible = $\{r_A, r_B, r_C\}$ (optimal at some θ) Rejected = $\{r_D, r_E\}$ (below the maximum at all θ)

Figure 2: Admissible partition for "What causes the seasons?" Five responses: r_A–r_C give the axial-tilt explanation (correct); r_D and r_E give the distance explanation (a misconception). Clustering changes with θ. A response is admissible if it lies in a highest-probability cluster at *some* threshold. r_D and r_E are rejected because their cluster is below the maximum at *every* threshold shown.

Threshold θ	Clusters	Max cluster prob	SE (bits)
0.70	2	0.70	0.881
0.80	2	0.70	0.881
0.85	3	0.50	1.485
0.90	4	0.50	1.761

Table 2: Semantic entropy varies with clustering threshold

The entropy range is $[\underline{\mathrm{SE}}, \overline{\mathrm{SE}}] = [0.881, 1.761]$, a width of 0.880 bits, and its ends are the entropies at the endpoint thresholds, as Proposition 9 states.

The response intervals over this range are $[0.50, 0.70]$ for r_A, $[0.20, 0.70]$ for r_C, and $[0.20, 0.30]$ for r_D. The width for r_C shows how far its agreement share depends on the threshold.

The admissible set is $\{r_A, r_B, r_C\}$, as in §2.3.3. The responses r_D and r_E are below the maximum at all thresholds in the table. Γ-maximin selects r_A and r_B, and Γ-maximax selects all three admissible responses. Extending the range to $\theta = 0.95$ makes all five responses singletons, with r_A alone at the maximum, and leaves the admissible set unchanged in this example. In a variant with tied singletons, all tied responses would become admissible; Table 8 shows how much this matters in practice.

3.6 What the Construction Can and Cannot Do

Proposition 12 fixes how far the construction can depart from plurality at a single threshold. Under $\mathcal{M}_\Theta$ every expectation of a response gamble is a mixture $\sum_\theta \lambda(\theta)\, u_\theta(r)$ of its agreement shares, so every rule in the proposition depends on a response only through its agreement profile $(u_\theta(r))_{\theta\in\Theta}$: how much of the sample agrees with it at each threshold. The admissible set is a union of pluralities, one per threshold. So the partition can differ from plurality at a single threshold only on questions where the largest cluster changes with θ.

Dependence on the agreement profile alone is what precise weights on responses impose. Maximality can admit a response that is the plurality at no threshold, as the seven-response example after Proposition 12 shows, but only through the profile. If the weights were themselves imprecise,

from prompt perturbation (Definition 16) or from an ensemble (Definition 20), each $u_\theta(r)$ would become an interval, and a rule could respond to how agreement varies with the weights, not only with the threshold. The construction here leaves the weights precise. The empirical question of Section 4 is therefore narrow: whether the agreement profile across thresholds carries information about correctness that its value at one threshold does not.

3.7 When to Use the Partition

As §3.6 shows, the partition can differ from plurality at one threshold only where the largest cluster changes with θ, which requires some similarities to lie inside the range. It could add value where responses vary in paraphrase, so that similarities spread across the range and any single threshold is arbitrary, and where decisions attach to individual responses; these data do not test either. A fixed threshold suffices when domain knowledge dictates one or when similarities are bimodal, and on free-form short-answer QA they mostly are (§4.3). Even on the multiple-choice benchmarks, where more similarities lie inside the range, a response's agreement shares at all thresholds added little to its share at one (§4.3). Entropy bounds suffice when only a per-prompt signal is needed.

3.8 Summary of the Construction

For a fixed clustering, semantic entropy uses standard probability and is coherent, at the cost of an unargued commitment to one threshold. The threshold credal set replaces that commitment with a commitment to a stated range, and fixed-threshold semantic entropy becomes the expectation under one extreme point of the set, which puts semantic entropy on the same footing as the methods of Section 5. The set supports an interval for each response, an entropy range, and a family of decision rules: fixed-threshold plurality is the Bayes rule under one member, the Γ rules are plurality at the endpoints, and E-admissibility over its extreme points gives the admissible partition, which labels individual responses as per-prompt scores cannot (Figure 2). The response intervals are coherent, and the partition is a coherent choice function with agreement shares fixed on the full sample. Re-run on subsets it is not, and neither is plain plurality.

4 Does Threshold Imprecision Carry Information?

Section 3 built a credal set and its decision rules. Whether they are of any use is an empirical matter. Semantic entropy and self-consistency share a premise: a model's errors are less consistent than its correct answers, so an answer it gives repeatedly is more often right than one it gives rarely. If the premise holds, a response that falls outside the largest cluster at every threshold in a range should be wrong more often than a response that does not, and the admissible partition should separate correct from incorrect responses. We call this Prediction 3, the third of the paper's four predictions; Appendix D reports the other three.

We take two questions in turn. The first is whether the partition separates correct from incorrect responses (§4.2). The second is whether leaving the threshold open adds anything to a single threshold, and it has two parts: whether the sweep, clustering at every threshold of the grid instead of one, adds anything to plurality at one threshold (§4.3), and whether the width of the entropy range, which is zero at any single threshold, flags errors (§4.4). The two questions differ in the standing of their evidence. The first was registered on four benchmarks. The analyses of the second are post hoc, with one exception: the correlation of width with difficulty, the share of the six responses to a question that are scored incorrect, was registered (H4). Post hoc analyses are labeled as such throughout.

Benchmark	Status	n_{rej}	Rejected	Adm. Acc	Rej. Acc	Ratio [response CI]	Pair CI
TriviaQA	Discovery	627	5.2%	77.9%	26.2%	2.98× [2.64, 3.42]	[2.53, 3.60]
MMLU	Registered*	618	5.1%	77.9%	65.9%	1.18× [1.12, 1.25]	[1.11, 1.27]
ARC-Challenge	Registered*	636	5.3%	86.9%	78.9%	1.10× [1.06, 1.15]	[1.05, 1.16]
TruthfulQA	Registered	3847	19.6%	17.6%	9.4%	1.88× [1.70, 2.09]	[1.65, 2.15]
TruthfulQA (judge)	Post hoc	3847	19.6%	27.5%	25.4%	1.08× [1.02, 1.15]	[1.01, 1.17]
NQ-Open	Registered	2389	19.9%	31.0%	19.2%	1.62× [1.49, 1.77]	[1.45, 1.81]

Table 3: Admissible versus rejected accuracy, pooled over four base models (Qwen2.5-7B, Mistral-7B, Llama-3.1-8B, Gemma-2-9B) with similarity the mean of the two directed entailment probabilities (§4.1). The ratio is admissible accuracy over rejected accuracy. Response CI: bootstrap over responses, the registered method (10,000 resamples). Pair CI: bootstrap over question-model pairs (2,000 resamples, post hoc), which respects dependence among responses to one question. *With deviations from the registered prompt format, retention rule, and tie rule, and a lenient letter rule in the scorer; see the text. Per-model values: Table 25.

4.1 Protocol and Evidence Status

Prediction: Admissible responses should be more accurate than rejected responses. We measure this with the *admissibility ratio*, admissible accuracy divided by rejected accuracy, and report it with the *rejection rate*, the share of responses rejected, since a high ratio on a very small rejected pool has little practical use.

Protocol: For each question and each of four base models (Qwen2.5-7B, Mistral-7B, Llama-3.1-8B, Gemma-2-9B), we sample 15 responses, three at each of five temperatures, from a prompt fixed within each benchmark. We cut each response at its first newline and then its first period (Appendix E), and retain the first six in generation order, which are the three at $\tau = 0.3$ and the three at $\tau = 0.5$. We then compute the similarity of each pair of retained responses, the mean of its two directed entailment probabilities from DeBERTa-large-MNLI. We apply Definition 10 with equal weights on the six responses and $\Theta = \{0.50, 0.55, \ldots, 0.95\}$. Each response is scored against the benchmark's gold answers by substring match, and on the multiple-choice benchmarks also by the letter of the gold answer (Appendix E.3). We call these the original scorers, since they are the ones the registered runs used. Registered tests are two-sided Fisher exact tests, and registered intervals are 95% percentile bootstrap intervals over responses. Responses to one question are dependent, so those tests overstate significance and those intervals are too narrow; Table 3 also gives intervals that resample question-model pairs, a pair being the six responses of one model to one question (post hoc). Appendix E.3 gives details.

Evidence status: TriviaQA is the discovery set: the effect was first observed on it, and it was not registered. MMLU (Hendrycks et al., 2021) and ARC-Challenge (Clark et al., 2018) (registration Wheeler 2026a), TruthfulQA (Lin et al., 2022) (registration Wheeler 2026c), and NQ-Open (Kwiatkowski et al., 2019; Roberts et al., 2020) (registration Wheeler 2026b) were preregistered before data collection. Analyses labeled post hoc below are new analyses of data already collected, and runs labeled exploratory were never registered.

4.2 Does the Partition Separate Correct from Incorrect Responses?

Results: Table 3 gives pooled results and Tables 4 and 5 give every hypothesis outcome.

The answer to the first question is yes in direction and mixed in size. Pooled over the four models, the ratio is above 1 on all five benchmarks under the original scorers, and its size varies with benchmark and scorer. Each registration also set a size criterion, a pooled ratio that its benchmark

was predicted to reach, and these criteria are met only on TruthfulQA under substring scoring. On free-form factual QA the separation is clear: 2.98× [2.64, 3.42] on TriviaQA at 5.2% rejection, and a registered 1.62× [1.49, 1.77] on NQ-Open at 19.9%, short of its registered 2.0× criterion. By that registration's decision rule this is partial support: a positive effect smaller than TriviaQA would predict. NQ-Open also misses H6, the registered test that free-form benchmarks separate better than multiple-choice ones (pooled ratio at least 1.71×), though it lies nearer TruthfulQA's 1.88× than the multiple-choice values.

Multiple-choice benchmarks show small effects (1.18× MMLU, 1.10× ARC-Challenge), and the registered criterion of 1.5× fails on both. Those two results carry four caveats. Prompts included the answer options, contrary to the registered format. The retained samples were the first six in generation order, not the registered mix of temperatures. Ties for the largest cluster were kept, where the registration breaks them by lowest cluster index. Under the registered rule ARC-Challenge gives 1.29× [1.24, 1.34] at 16.0% rejection (post hoc), with no registered ARC-Challenge outcome changed. And the registered scorer accepted the gold letter as a substring, so a letter inside any word earned credit: over all stored samples (post hoc), accuracy for Mistral on ARC-Challenge is 0.745 under this rule and 0.452 under a stricter one (Appendix E.3). Rescoring the stored partitions under the stricter rule (post hoc) gives 1.72× [1.55, 1.93] on ARC-Challenge and 1.24× [0.88, 2.50] for Gemma on MMLU, the only MMLU model whose samples are stored. So the weak multiple-choice effect is in part a scorer effect, and on ARC-Challenge largely one. Under the stricter scorer, ARC-Challenge's 1.72× exceeds NQ-Open's registered 1.62×, so these data do not establish the format effect that H6 tests.

TruthfulQA shows the role of the scorer: the registered substring scorer gives 1.88×, and post hoc rescoring of the same partition by an LLM judge (Claude Haiku 4.5; Appendix E.3) gives 1.08× [1.02, 1.15]. The rejected set there contains truthful debunkings that substring matching does not credit.

Individual models fall at or below 1 in some cells: Qwen on MMLU (0.97×), Gemma on ARC-Challenge (0.91×, with 7 rejected responses). When the model's most frequent answer is wrong, filtering by plurality rejects the correct minority. Stratifying by the number of correct responses shows this directly: on the two free-form registered benchmarks the 1–2/6 bin, the question-model pairs on which one or two of the six responses are correct, inverts (0.51× TruthfulQA, 0.48× NQ-Open), and the 3–4/6 bin peaks (2.97×, 2.66×; Table 26). Binning conditions on the outcome. Under stratifiers that do not (Appendix E.3; post hoc, TriviaQA), the peak is smaller and the inversion is absent. Both registrations bound the inversion (H5), and both bounds fail. As registered, we treat the failure of both bounds as evidence that this limitation, the rejection of a correct minority, is more severe than we had characterized. The post hoc stratifiers qualify that reading and do not remove it.

What the partition delivers. The partition is a response-level filter, not an estimator of predictive uncertainty. On TriviaQA, the discovery set, 26.2% of rejected responses are correct, so it discards some correct responses. It barely changes the majority-vote answer, the most frequent of the sampled answers after normalization, since every response in a largest cluster at some threshold is admissible: majority-vote accuracy is 75.9% over all responses and 75.8% over admissible ones, a difference of 2 of 2,000 question-model pairs (post hoc). Its use is abstention and flagging at the level of individual responses. We do not offer it as a general method for improving answers.

Ratios and odds ratios (post hoc). The accuracy ratio depends on base accuracy, which is high on the multiple-choice benchmarks and low on TruthfulQA and NQ-Open. As odds ratios computed

Label	Criterion	Result	Outcome
Not pre-registered			
P1	Accuracy loss grows with base-model diversity; restated after the data as non-linear, with location set by tuning style (TriviaQA)	Matched reruns: Qwen Q3 −9.0 becomes +1.4, Llama Q1 −8.9 becomes −1.4, Mistral Q1 −13.4 becomes −6.9 and is reproduced by another base model; Qwen's Q1 loss under its template remains (§D.5.1)	Not supported
P2	Calibration error (ECE) higher for incoherent responses (MMLU statements)	Llama 2.21×, 1.46× without both-1.0 pairs, 1.26× [0.61, 2.22] under a neutral prompt ($n = 400$); Gemma reflects selection; Mistral $n_{\text{coh}} = 6$	Inconclusive: one model, prompt-dependent
P3	Admissible more accurate than rejected (TriviaQA)	2.98× [2.64, 3.42]	Holds (discovery set)
P4	Optimal τ varies with difficulty; restated after the data as sensitivity falling with diversity (TriviaQA, one model)	Peaks at low τ in every quartile; standard deviation falls with diversity, peak-normalized does not; retention runs opposite (§D.5)	Not supported
MMLU, registration osf.io/ady3k, registered scorer; prompt, retention, and tie-rule deviations			
H1a	Pooled ratio > 1	1.18× [1.12, 1.25], $< 10^{-10}$	Supported
H1b	Pooled ratio ≥ 1.5	1.18×	Not supported
H1c	Ratio > 1 for ≥ 3 of 4 models	3 of 4	Supported
ARC-Challenge, registration osf.io/ady3k, registered scorer; prompt, retention, and tie-rule deviations			
H2a	Pooled ratio > 1	1.10× [1.06, 1.15], 7.1×10^{-8}	Supported
H2b	Pooled ratio ≥ 1.5	1.10×	Not supported
H2c	Ratio > 1 for ≥ 3 of 4 models	3 of 4	Supported
MMLU and ARC-Challenge, secondary, registration osf.io/ady3k			
H3	String admissibility beats frequency and consistency filters, rejection rates within 3 pp	MMLU 1.48× vs. 1.23×, 1.23×; rejection gap 3.9, 3.9 pp. ARC 1.54× vs. 1.19×, 1.19×; rejection gap 5.4, 5.3 pp	Not supported (rates not comparable)
H4	Entropy-range width correlates positively with difficulty	Spearman $\rho = -0.54$ MMLU, -0.46 ARC	Not supported

Table 4: All hypotheses in this paper. P1–P4 are this paper's predictions and were not registered; P1 and P4 were restated after their data were seen (§D.5). H-labels follow each registration. Implications: the direction holds on free-form factual QA; the registered format test (H6) fails on NQ-Open; registered pooled size criteria are met only on TruthfulQA under substring scoring; the bin peak recurs on two registered benchmarks; the bounded-inversion criterion fails on both. Bins count correct responses among six.

Label	Criterion	Result	Outcome
TruthfulQA, registration osf.io/dqbmg, substring scorer			
H1	Pooled ratio > 1, $p < 0.05$	1.88× [1.70, 2.09], $< 10^{-10}$	Supported
H2	Pooled ratio ≥ 1.5	1.88×	Supported
H3	Ratio > 1 for ≥ 3 of 4 models, $p < 0.05$ for ≥ 2	4 of 4; 4 of 4	Supported
H4	3–4/6 bin: lower CI ≥ 2.0	2.97× [2.38, 3.89]	Supported
H5	1–2/6 bin: lower CI ≥ 0.5	0.51× [0.45, 0.59]	Not supported
H6	Pooled ratio ≥ 1.71	1.88×	Supported
NQ-Open, registration osf.io/78pb6, substring scorer			
H1	Pooled ratio > 1, $p < 0.05$	1.62× [1.49, 1.77], $< 10^{-10}$	Supported
H2	Pooled ratio ≥ 2.0	1.62×	Not supported
H3	Ratio > 1 for ≥ 3 of 4 models, $p < 0.05$ for ≥ 2	4 of 4; 4 of 4	Supported
H4	3–4/6 bin: lower CI ≥ 2.0	2.66× [2.20, 3.32]	Supported
H5	1–2/6 bin: lower CI ≥ 0.5	0.48× [0.41, 0.57]	Not supported
H6	Pooled ratio ≥ 1.71	1.62×	Not supported

Table 5: All hypotheses in this paper, continued from Table 4: TruthfulQA and NQ-Open registrations.

from Table 3, the four registered benchmarks are alike: 1.89 on NQ-Open, 2.06 on TruthfulQA, 1.82 on MMLU, and 1.77 on ARC-Challenge, the last two under the registered scorer. The outlier is the TriviaQA discovery set at 9.93, whose runs used a different prompt (Appendix E.3).

4.3 What the Threshold Sweep Adds

The first part of the second question is whether clustering at every threshold adds anything to plurality at one. We take it in five steps: where the two can differ at all, how often they do, how the decision rules of Section 3 compare, which single threshold does best, and a direct test of a response's agreement profile against its agreement share at one threshold, with the one stratum in which that test finds a gain. Every analysis in this subsection is post hoc.

Where leaving the threshold open can matter (post hoc). The partition can differ from plurality at a single threshold only where the plurality, the set of responses in a largest cluster, changes with θ. Loosening the threshold only merges clusters (Proposition 6), so this happens when responses that are apart at strict thresholds merge at looser ones, joining the largest cluster or forming a new one at least as large. Whether it happens depends only on the sampled responses and their similarities, so these question-model pairs are found without labels. They are 16.4% of TriviaQA pairs and 34.5% to 50.6% of pairs on the other benchmarks (Table 6); everywhere else the partition equals plurality at any single threshold. We call these pairs the stratum of changing plurality. Within it, admissibility has a higher ratio than plurality at $\theta = 0.75$, a threshold in the middle of the registered range, on TriviaQA and TruthfulQA and a lower one on NQ-Open, ARC-Challenge, and MMLU; on all five, plurality's ratio lies inside the admissibility interval. The comparison is also between different operating points: on TriviaQA admissibility rejects 52 responses in this stratum and plurality rejects 467, and plurality's retained set is the more accurate (57.7% vs. 53.3%).

How much there is to sweep (post hoc). On the cached similarity matrices, every threshold in the registered grid gives the same clustering on 57%, 59%, and 55% of question-model pairs

on TriviaQA, NQ-Open, and TruthfulQA, and only 2.9%, 2.9%, and 3.8% of response pairs have similarity in $[0.55, 0.85)$, well inside the range. On ARC-Challenge and MMLU (Gemma) the clustering is constant on 11% and 1% of pairs, and 10% and 14% of similarities lie in that band. So on most free-form question-model pairs the clustering does not change across the grid, which is the most direct explanation of what follows. Similarities do gather just below the range on NQ-Open and TruthfulQA, where 4.2% and 3.9% lie in $[0.49, 0.50)$. Since the similarity averages the two directions, these may be pairs in which one response entails the other and not conversely, which a lower endpoint would let merge. On these data that matters, post hoc: starting the range at 0.45 or 0.40 instead of 0.50 raises the NQ-Open ratio from 1.62× to 2.42× and 2.78×, at rejection rates of 14.4% and 13.3% against 19.9%. That was found after the fact, among the 35 endpoint pairs (§4.5), and should be registered before it is tested on fresh data.

Benchmark	In stratum	Admissibility	Plurality at $\theta = 0.75$	n_{rej} adm. / plur.
TriviaQA	16.4%	2.13× [1.43, 3.94]	1.59×	52 / 467
ARC-Challenge	41.1%	1.00× [0.93, 1.08]	1.03×	154 / 1530
MMLU (Gemma)	50.6%	1.10× [0.76, 2.72]	1.17×	8 / 518
TruthfulQA†	37.7%	2.14× [1.72, 2.75]	1.75×	1042 / —
NQ-Open	34.5%	1.34× [1.13, 1.62]	1.39×	511 / —

Table 6: Post hoc: question-model pairs whose plurality (the responses in a largest cluster) changes with θ. Off this stratum, admissibility and plurality at $\theta = 0.75$ coincide. A dash marks a count that was not recorded. The two columns compare different operating points, since rejected pools differ in size. †Computed on rebuilt similarity matrices that do not reproduce the registered partition (pooled ratio 1.72× rather than 1.88×).

Decision rules on the threshold credal set (post hoc). Table 7 applies the rules of Proposition 12 to the same similarity matrices. C_{sem}, E-admissibility over all of $\mathcal{M}_\Theta$, and maximality reject the same responses on all five benchmarks, as they must with six equally weighted responses (see the discussion after Proposition 12), and interval dominance differs slightly on NQ-Open and TruthfulQA alone. The rules that can differ do. Γ-maximax, the plain plurality rule at $\theta = 0.50$, has a higher ratio than C_{sem} on all five benchmarks, at a higher rejection rate. Γ-maximin, plurality at $\theta = 0.95$, has a lower ratio than Γ-maximax on four of the five. As a difference of log ratios, with intervals resampling question-model pairs, Γ-maximax leads by 0.45 [0.34, 0.58] on TriviaQA, 0.26 [0.18, 0.34] on NQ-Open, 0.14 [0.05, 0.24] on TruthfulQA (0.09 [0.03, 0.15] under the judge), and 0.16 [0.11, 0.20] on ARC-Challenge; on MMLU (Gemma) the interval is too wide to separate them, −0.01 [−0.48, 0.30]. On TriviaQA, NQ-Open, and TruthfulQA the two rules reject similar shares of responses (Table 7).

C_{sem} rejects fewer responses than Γ-maximax mostly because of ties (post hoc, like the two paragraphs that follow). Under equal weights, an all-singleton clustering ties every response, so the union over thresholds admits them all, including responses that plurality at $\theta_{\min}$ rejects. We call those responses readmitted. Table 8 shows that these responses are less accurate than the ones C_{sem} rejects, on all five benchmarks. Responses readmitted through a cluster of two or more are more accurate than the rejected ones on four of the five.

These two results point to a variant that ignores all-singleton clusterings above $\theta_{\min}$. It gives 3.21× [2.80, 3.77] at 7.2% rejection on TriviaQA and 1.85× [1.68, 2.05] at 26.7% on NQ-Open. Against C_{sem}, its log ratio is higher by 0.07 [−0.02, 0.16] and 0.14 [0.08, 0.19], and only the NQ-Open interval excludes zero. Against Γ-maximax, its ratio is indistinguishable on these two benchmarks (log-ratio differences 0.02 [−0.03, 0.07] and 0.00 [−0.02, 0.02]), and its rejection rate

is lower by construction. So on these two benchmarks the sweep's demonstrated gain is a lower rejection rate at a similar ratio, not a higher ratio. On TruthfulQA, ARC-Challenge, and MMLU the variant has a higher ratio than Γ-maximax at a lower rejection rate (1.94× against 1.81×, 1.29× against 1.24×, 1.39× against 1.37×), on matrices that only partly reproduce the registered partitions (Table 7).

Three further checks bear on the role of ties. Weighting responses by generation log-probabilities, which breaks ties, gives 3.16× [2.87, 3.52] at 9.3% rejection on TriviaQA, also above C_{sem}'s 2.98× but at a higher rejection rate. The accuracy of tied singletons minus rejected accuracy is −0.056 [−0.138, 0.029] on TriviaQA and −0.061 [−0.099, −0.022] on NQ-Open. Under a word-boundary scorer, which refuses bare fragments the substring rule passes, the difference is −0.070 [−0.147, 0.012] and −0.058 [−0.096, −0.020], and the variant's log-ratio gain is 0.100 [−0.001, 0.192] and 0.135 [0.072, 0.193]. All of this was found on these data and needs replication on fresh data.

Benchmark	Γ-maximin ($\theta = 0.95$)	Γ-maximax ($\theta = 0.50$)	C_{sem}	Without tied singletons
TriviaQA	2.00× (8.4%)	3.15× (7.8%)	2.98× (5.2%)	3.21× (7.2%)
NQ-Open	1.44× (25.8%)	1.86× (27.5%)	1.62× (19.9%)	1.85× (26.7%)
TruthfulQA†	1.57× (29.5%)	1.81× (29.2%)	1.72× (22.1%)	1.94× (27.7%)
ARC-Challenge†	1.06× (9.6%)	1.24× (15.8%)	1.12× (5.5%)	1.29× (14.3%)
MMLU (Gemma)†	1.39× (1.9%)	1.37× (16.4%)	1.23× (0.8%)	1.39× (15.2%)

Table 7: Post hoc comparison of decision rules on the threshold credal set: admissibility ratio (rejection rate). The text gives intervals on the main differences. †Stored similarity matrices that do not exactly reproduce the registered partition.

Benchmark	Rejected by C_{sem}	Readmitted: tied singleton	Readmitted: cluster of ≥ 2
TriviaQA	0.262	0.206 ($n = 233$)	0.316 ($n = 76$)
NQ-Open	0.192	0.131 ($n = 814$)	0.181 ($n = 94$)
TruthfulQA†	0.102	0.066 ($n = 1107$)	0.223 ($n = 292$)
ARC-Challenge†	0.777	0.643 ($n = 1059$)	0.937 ($n = 174$)
MMLU (Gemma)†	0.652	0.596 ($n = 433$)	0.722 ($n = 36$)

Table 8: Post hoc: accuracy of responses that plurality at $\theta_{\min}$ rejects and the union over thresholds readmits, against accuracy of the responses the union rejects. †As in Table 7.

At the partition's rejection rate on TriviaQA, two rankings do better (post hoc). Keeping the responses with the highest agreement share at $\theta_{\min}$, their Γ-maximax value, gives a retained accuracy of 0.7826 against the partition's 0.7790 (partition minus ranking −0.0037 [−0.0062, −0.0012]). Ranking by the agreement share averaged over thresholds reaches 3.89× [3.36, 4.60] at the same 5.2% rejection, against the partition's 2.98×. Neither ranking is a decision rule on $\mathcal{M}_\Theta$: each needs a cutoff, and neither is guaranteed to keep every response in a largest cluster at some threshold.

Which threshold (post hoc). The choice of a single threshold matters, mostly at the strict end. We score each response by its agreement share $u_\theta(r)$ at one threshold and measure response-level AUROC over all the responses of a benchmark: the probability that a correct response scores above an incorrect one, with ties counted as one half. Under the original scorers, this AUROC varies by at most .03 from 0.50 to 0.85 on the free-form benchmarks and falls by .05 to .14 from 0.50 to 0.95; on the multiple-choice benchmarks it falls by .08 to .11 at 0.80 and by .23 to .25 at 0.95. At the strict end matching responses fall into separate clusters: at 0.95 every response is a singleton on

27–32% of free-form and 78–97% of multiple-choice question-model pairs (§4.4), so agreement there says little. Under the TruthfulQA judge the strict end does no worse (.558 at 0.95 against .545 at 0.50).

What leaving the threshold open adds (post hoc). Whether leaving the threshold open adds anything to one threshold is the question of §3.6, and a direct test answers it. We predict each response's correctness out of fold, with folds formed by question so that all models' responses to a question fall in one fold, in two ways: by its agreement share at the single threshold that maximizes AUROC on the training folds, and by a logistic regression on its whole agreement profile $(u_\theta(r))_{\theta\in\Theta}$. Table 9 gives the results. The profile's gain in response-level AUROC runs from −.004 to +.007, every interval includes zero, and no upper bound exceeds .023. Gradient-boosted models in place of the logistic regression do no better: across four settings their gains run from −.024 to +.013, with one interval just above zero (ARC-Challenge, .010 [.0003, .020]) and one below (NQ-Open, −.024 [−.042, −.008]). The chosen threshold itself beats a fixed 0.50 only on TruthfulQA, where every fold chose 0.90 and AUROC rises from .670 to .703; on the other four the two differ by at most .006.

Where the sweep helps (post hoc). The exception is the stratum of Table 6, the pairs whose plurality changes with θ, which the sweep identifies without labels. On those pairs the profile fit on all pairs beats the threshold chosen for all pairs by .054 [.020, .087] on TriviaQA (16% of its pairs) and by .026 [.011, .039] on ARC-Challenge, and by about zero on the other three. With both the profile and the threshold fit on the stratum alone, the gap closes on TriviaQA (−.008 [−.030, .012]), so there the sweep marks the questions that one global threshold serves poorly. On ARC-Challenge it does not (.043 [.020, .065]): there, and nowhere else in these data, the profile beats even a threshold chosen on the same pairs. That result comes from a multiple-choice benchmark scored with the lenient registered letter rule, whose AUROC falls sharply above 0.75 and whose matrices only partly reproduce the registered partition, so it needs replication before it is read as information carried by threshold imprecision. Leaving the threshold open adds little to a single threshold chosen on other labeled questions, and outside TruthfulQA that threshold adds little to a fixed 0.50.

Benchmark	One threshold	Profile	Gain [95% CI]
TriviaQA	.779	.786	+.007 [−.010, +.023]
NQ-Open	.710	.706	−.004 [−.012, +.004]
TruthfulQA	.703	.700	−.003 [−.010, +.004]
TruthfulQA (judge)	.558	.558	.000 [−.010, +.011]
ARC-Challenge	.737	.741	+.004 [−.014, +.019]
MMLU (Gemma)	.745	.745	.000 [−.018, +.015]

Table 9: Post hoc direct test of §3.6: out-of-fold AUROC for the correctness of individual responses, with five folds formed by question. One threshold: a response's agreement share at the threshold that maximizes AUROC on the training folds. Profile: a logistic regression on its agreement shares at all ten thresholds, fit on the training folds. Gain: profile minus one threshold, with a 95% interval resampling questions (1,000 resamples). Correctness labels and similarity matrices are those of Table 7, whose ratios they reproduce.

4.4 Does the Entropy Range Flag Errors?

Width of the entropy range. The width of the entropy range is the one output of the threshold construction that is zero at any single threshold, and the MMLU/ARC registration predicted that it would rise with difficulty (H4). It does not: Spearman ρ is -0.54 on MMLU and -0.46 on ARC-Challenge. Its AUROC for majority-vote error, the probability that a question where the majority vote is wrong gets a wider range than one where it is right, is .479 on NQ-Open and .385 on TruthfulQA, and its AUROC for the 1–2/6 bin is .474 on NQ-Open, .514 on TruthfulQA, and .252 on ARC-Challenge (post hoc). So it does not flag errors. Values well below .5 mean narrower ranges where errors are; the next two paragraphs account for them.

Why width falls with difficulty on the multiple-choice benchmarks (post hoc). At $\theta = 0.95$ every response is a singleton on 96.8% of MMLU (Gemma) and 78.4% of ARC-Challenge question-model pairs, because the similarity of two such responses, even matching ones, rarely reaches 0.95: only 0.5% of response pairs on MMLU and 6.2% on ARC-Challenge have similarity that high, against 25% to 52% on the free-form benchmarks. So $\mathrm{SE}(\theta_{\max})$ is nearly the constant $\log 6$, width is nearly $\log 6 - \mathrm{SE}(\theta_{\min})$, and width and $\mathrm{SE}(\theta_{\min})$ have Spearman correlation -0.99 and -0.89. Width then falls wherever $\mathrm{SE}(\theta_{\min})$ rises, as it does with difficulty here. H4 still fails as registered, but on ARC-Challenge it measured a ceiling of the similarities rather than threshold imprecision. On MMLU only Gemma's matrices are stored, so the same reading covers one of the four models behind $\rho = -0.54$. On the free-form benchmarks every response is a singleton at 0.95 on 27–32% of pairs, width is only weakly tied to $\mathrm{SE}(\theta_{\min})$ (ρ from -0.07 to -0.36), and it still does not flag majority-vote error (AUROC .479 on NQ-Open).

The TruthfulQA width value is fragile (post hoc). The .385 above comes from widths stored by the registered runs. Recomputed from the cached matrices, built with the same NLI model, averaging, and 256-token limit, the AUROC is .488. The two sources agree on most pairs at $\theta = 0.50$, where SE differs on 4.1% of question-model pairs, and on fewer at $\theta = 0.95$, where it differs on 13.7%, mostly for Mistral (21.4%) and Gemma (19.3%). Between 7.4% and 10.5% of similarities lie within 0.02 of 0.95, so small numerical differences between runs move responses in and out of singletons at the strictest threshold, and width inherits that instability. Neither value should be read as a signal.

Scalar diagnostic. The scalar $1 - \max_\theta \max_k P_\theta(C_k^\theta)$ is secondary. It is one minus the Γ-maximax value, and across the 35 endpoint pairs (§4.5) its AUROC takes one value per $\theta_{\min}$ on both TriviaQA and NQ-Open, as Proposition 6 requires. As with width, the AUROC here is for majority-vote error, with questions as the unit. For this scalar it is 0.773 on TriviaQA, against 0.774 for SE at $\theta = 0.75$, 0.786 for KLE, and 0.788 for SNNE. On ARC-Challenge, under the registered scorer, the values are 0.737, 0.728, 0.720, and 0.712. The scalar leads only on ARC-Challenge and offers no consistent gain over these baselines (post hoc); Table 27 gives NQ-Open and TruthfulQA.

4.5 Baselines, Range, and Scale

Comparison to baselines. The natural response-level baseline is a self-consistency partition: keep the largest group of identical normalized strings, ties included, and reject the rest (Wang et al., 2023). On TriviaQA it gives 1.92× [1.80, 2.06] at 10.7% rejection, against 2.98× at 5.2% for NLI admissibility (post hoc). The two sit at different operating points, and the comparison does not show that NLI is the better similarity measure. Table 10 gives string-based admissibility,

which clusters by exact, normalized, and substring match, and three heuristic filters: *frequency* rejects responses whose normalized form occurs once among the six, *consistency* does so after also stripping a leading article or a leading phrase such as "it is", and *length* rejects responses of fewer than two words.

Filter	Retained Acc	Rejected Acc	% Rejected	Ratio
Admissibility (string)	79.2%	23.0%	7.1%	3.45×
Consistency	78.6%	28.4%	6.9%	2.77×
Frequency	78.7%	32.8%	7.6%	2.40×
Length (<2 words)	71.1%	80.5%	43.5%	0.88×
Random baseline accuracy: 75.2%; Majority vote accuracy: 75.9%				

Table 10: String-based admissibility vs. heuristic filters on TriviaQA (post hoc; aggregate across 4 models, 12,000 responses). At comparable rejection rates, string-based admissibility has the highest ratio. The length filter is counterproductive: short responses are typically correct direct answers.

All of the following is post hoc. On TriviaQA, string-based admissibility (3.45× at 7.1%) and admissibility on sentence-embedding cosine similarity (all-mpnet-base-v2; 4.17× [3.50, 5.13] at 3.8%) both have higher ratios than NLI admissibility (2.98× at 5.2%), the embedding version at a lower rejection rate and the string version at a higher one. String-based admissibility also exceeds NLI on three of four models on NQ-Open. On MMLU and ARC-Challenge its pooled ratios (1.48× and 1.54×) exceed NLI's at higher rejection rates (12.6% and 8.8%, against 5.1% and 5.3%), and on TruthfulQA the two tie (Appendix E.3). The partition is thus not specific to NLI, and on free-form factual QA with short answers the results do not favor NLI over simpler similarity measures. On TruthfulQA, sentence-embedding admissibility gives 1.36× [1.19, 1.59] under substring and 0.99× [0.91, 1.09] under judge scoring, repeating the attenuation seen with NLI.

Threshold range. The range is a hyperparameter. Across 35 endpoint pairs ($\theta_{\min}$ from 0.40 to 0.70 in steps of 0.05, $\theta_{\max}$ in $\{0.80, 0.85, 0.90, 0.95, 0.99\}$; post hoc), the TriviaQA ratio spans 2.16× to 4.83× and the rejection rate 0.36% to 8.33%; on NQ-Open, 1.52× to 2.93× and 7.40% to 27.47%. Every interval excludes 1, so the direction does not depend on the range, but the size of the ratio and the rejection rate do. The registered range $[0.50, 0.95]$ gives the reported values. A 10-point grid can miss clusterings: Theorem 13 allows up to 16 cells and 6 distinct clusterings for $M = 6$. The registered grid matches the continuum partition on all 2,000 TriviaQA question-model pairs and differs on 2 of 2,000 on NQ-Open (post hoc).

Instruction-tuned and frontier models (exploratory). Table 11 reports runs that were not registered. All are on TruthfulQA under judge scoring. Pooled ratios are 0.95× for Llama-3.1-8B-Instruct, whose prompt carried two begin-of-text tokens, a defect that makes the value unreliable, and 1.08× for Gemma-4-26B-it. Two frontier models, at $n = 200$ questions each, give 1.15× (Claude Sonnet 4.6, $p = 0.038$) and 0.85× (Gemini 2.5 Flash, $p = 0.085$). Each sample of 200 was drawn in two stages, the second 100 added after the first gave $p = 0.067$ and $p = 0.139$ (Appendix E), so these p-values overstate the evidence. The two models' directions disagree, the runs are underpowered, and they reject few responses: 4.3% and 2.7%, against 20.9% for the base models on a shared 100-question subset. We draw no conclusion on the direction of the effect at frontier scale.

Model	Rejection	Adm. Acc	Rej. Acc	Ratio [95% CI]	n_{rej}	Fisher p
Claude Sonnet 4.6	4.3%	86.4%	75.0%	1.15× [1.00, 1.38]	52	0.038
Gemini 2.5 Flash	2.7%	77.1%	90.6%	0.85× [0.77, 0.98]	32	0.085
Base models, subset	20.9%	28.0%	22.9%	1.22× [1.04, 1.48]	502	0.024
Llama-3.1-8B-Instruct	7.3%	52.8%	55.7%	0.95× [0.86, 1.05]	359	0.297
Gemma-4-26B-it	11.3%	72.8%	67.4%	1.08× [1.02, 1.15]	552	0.009

Table 11: Exploratory, not pre-registered: TruthfulQA under LLM-judge scoring. Frontier rows use $n = 200$ questions and their directions disagree. Bootstrap intervals and Fisher tests can disagree at this rejected-pool size; both are reported.

4.6 Falsifiable Commitment

Prediction 3: On free-form factual QA with base models, admissible responses are more accurate than rejected responses, pooled across models. Falsified if a pooled ratio on such a benchmark, under a scorer matched to the benchmark, has an interval that lies entirely at or below 1. The scope is stated because individual cells outside it already fall at or below 1 (§4.2). The TruthfulQA judge result (1.08×), though pooled above 1, comes close to this boundary on an adversarial benchmark.

4.7 What the Null Says

The construction of Section 3 makes one modeling choice explicit: what counts as agreement. Our comparisons separate two questions about it. Which reading one picks can matter (§4.3). Leaving the reading open adds little to a single threshold chosen on other labeled questions (§4.3), for one structural reason and one empirical one. Structurally, with precise weights the partition is a union of pluralities (§3.6), so it can differ from plurality at one threshold only on question-model pairs where the largest cluster moves: 16% to 51% of pairs, depending on the benchmark (Table 6). Empirically, on free-form QA few similarities fall well inside the range (2.9% to 3.8% lie in $[0.55, 0.85)$), so most pairs give one clustering at every threshold (§4.3); as §3.7 notes, a fixed threshold suffices when similarities are bimodal. Where the largest cluster does move, the agreement profile beats one global threshold on TriviaQA and on ARC-Challenge, the latter under the lenient registered letter scorer, and adds about nothing elsewhere. On the multiple-choice benchmarks the width of the entropy range fails for a separate reason: the similarity of two matching answers rarely reaches 0.95, so at the strictest threshold almost every response is a singleton, and width reduces to $\log 6$ minus the entropy at the loosest threshold (§4.4).

Threshold imprecision might matter more where similarities are graded rather than mostly outside the threshold range, as in long-form generation, and where decisions attach to individual claims. Whether it does is open. The null here does not show that useful imprecision lies elsewhere. It does leave the weights on responses as a natural next place to look, since imprecise weights would let a rule depart from plurality (§3.6). Kudukkil Manchingal et al. (2026) put imprecision there (§6.4) and report gains in selective accuracy over single-model baselines. Their result is consistent with that reasoning, which we did not test.

5 The Same Vocabulary for Other Methods

Sections 3 and 4 treated one method, semantic entropy, and found in it a modeling choice that a credal set can leave open. This section asks what the same vocabulary says about the other common ways of quantifying an LLM's uncertainty: token probabilities, verbalized confidence, ensembles,

and conformal prediction. Of each we ask two things. What object of imprecise probability does the method construct, or, for conformal prediction, how does it relate to one? And is what it reports coherent? The answers are brief, because for these methods the reading is mostly definitional. It fixes terms, separates coherence from calibration, and locates one check that needs no labels. The threshold credal set remains the one place where the reading yields a new construction. Table 13 collects the answers, and Appendix C.4 gives worked examples.

5.1 Token Probabilities as Degenerate Credal Sets

Elicitation dependence. A language model reports a precise probability $P(t|\text{context})$ for each next token t, the softmax of its logits. The number is precise, but the question asked does not fix it. The same query yields different probabilities under:

- **Temperature variation:** $P_\tau(t)$ changes continuously with τ
- **Prompt perturbation:** Semantically equivalent prompts yield different distributions
- **Sampling:** A random seed changes which earlier tokens are drawn, and so the contexts on which later token probabilities condition

We read a reported distribution as one member of the set of distributions that temperature, prompt, or sampling can elicit. The interval $[\underline{P}(t), \overline{P}(t)]$ over that set then records how much the reported probability of a token depends on which setting was used. This is a claim about elicitation, not about an internal epistemic state.

Credal set construction. Two constructions make this set explicit. We give them in order of how good a reason there is to treat the variation they collect as uncertainty.

Definition 16 (Prompt Perturbation Credal Set)**.** *Given a base prompt π_0 and a set $\Pi(\pi_0)$ of prompts semantically equivalent to it, write $P(\cdot|\pi)$ for the model's next-token distribution under prompt π. The* prompt perturbation credal set *is*

$$\mathcal{M}_{prompt} = \overline{conv}\{P(\cdot|\pi) : \pi \in \Pi(\pi_0)\},$$

where $\overline{conv}$ denotes the closed convex hull.

The prompt perturbation credal set is the better grounded of the two: semantically equivalent inputs *should* yield consistent outputs, so variation across $\Pi(\pi_0)$ reflects sensitivity to surface form rather than to an arbitrary hyperparameter.

Definition 17 (Temperature Credal Set)**.** *Given logits z and temperature range $[\tau_{\min}, \tau_{\max}]$, the* temperature credal set *is:*

$$\mathcal{M}_{temp} = \overline{conv}\{P_\tau : \tau \in [\tau_{\min}, \tau_{\max}]\}$$

where $P_\tau(t) = \exp(z_t/\tau)/\sum_{t'} \exp(z_{t'}/\tau)$ and $\overline{conv}$ denotes closed convex hull. We use $\tau \in [0.3, 1.5]$, excluding extreme values, which produce a distribution concentrated on one token ($\tau \to 0$) or a near-uniform one ($\tau \to \infty$).[1]

[1]Temperature credal sets represent *elicitable* variation, not spontaneous hedging. A model that spontaneously reported intervals ("I assign probability 0.7 to 0.9") would provide more direct evidence of imprecision. See §C.2.5 for discussion of epistemic status.

The temperature credal set collects the distributions elicitable from the model over the stated range of temperatures. Temperature is imposed from outside; it is not a parameter the model learned to be uncertain about. The construction is cheaper than prompt perturbation, since varying τ only rescales the logits, but less well grounded. We measure it, on first tokens only (§D.5), because it is cheap. The coherence analysis below applies however the credal set is constructed.

Coherence status. The intervals are coherent whichever settings generate them. We state and prove the result for a finite set of prompts; Appendix A.1 checks its consequences and gives the extension to a range of temperatures.

Theorem 18 (Token Interval Coherence)**.** *Let $\Pi(\pi_0)$ be a finite non-empty set of prompt perturbations. The induced probability intervals:*

$$[\underline{P}(t), \overline{P}(t)] = \left[\min_{\pi\in\Pi} P(t|\pi), \max_{\pi\in\Pi} P(t|\pi)\right]$$

define coherent lower and upper probabilities.

Proof. The credal set $\mathcal{M}_{\text{prompt}} = \text{conv}\{P(\cdot|\pi) : \pi \in \Pi\}$ is by construction a closed convex set of probability mass functions. Since $P(t)$ is linear in the distribution, its infimum and supremum over this set are attained at one of the generating distributions $P(\cdot|\pi)$, so they are the minimum and maximum in the statement. As lower and upper envelopes of probability distributions, they are coherent by the Lower Envelope Theorem (Walley, 1991, Theorem 3.3.3). □

A point estimate from a single temperature or prompt is coherent, since a singleton credal set gives precise probabilities.

Verdict. Token-level methods report one member of a set that temperature or prompt can elicit. Whether the width they leave unreported carries information is empirical: on first tokens it rises with the diversity of a model's sampled answers, on whole responses the relation can reverse (§D.5), and neither is the same as flagging errors.

5.2 Verbalized Confidence as Lower Probability

When LLMs state confidence ("I'm 85% sure..."), what do these numbers mean? We read them as lower probabilities, an interpretive choice, and argue that confidences in a claim and its negation summing above one are a coherence violation.

Review. Verbalized confidence methods prompt a model to state its confidence as a number (Xiong et al., 2024):

- In a statement: "Is the following statement true? [statement]. Answer with a probability."
- In its own answer: "How confident are you in your answer? Give a percentage."

The P(True) of Kadavath et al. (2022) asks a similar question and takes the probability the model assigns to the answer "True". That probability is read from the output distribution and is not a stated confidence (Appendix C.1.2).

A model is calibrated when its stated confidence matches its accuracy, and expected calibration error (ECE) is the average gap between the two within bins of confidence. Xiong et al. (2024) report mean ECE above 0.3 for four of the five models they test under a plain prompt (0.18 for GPT-4), and describe verbalized confidence as overconfident.

Imprecise-probability interpretation. We read a stated confidence c as a lower probability: a model that says "I'm 85% confident" commits to $\underline{P}(\text{correct}) = 0.85$, leaves the upper probability open, and does not thereby commit to exactly 15% in the alternative (Walley, 1991). This is an interpretive choice; we have not measured whether models treat such statements as bounds. Yang et al. (2026) take the direct route, prompting models for probability intervals and credal sets, and report that the elicited imprecision tracks predictive error and question ambiguity more closely than precise elicitation does.

Why lower probability? Write $c(A)$ for the confidence a model states in a claim A, and $c(\neg A)$ for the confidence it states in the negation of A. If $c(A)$ were a point probability, coherence would require $c(A) + c(\neg A) = 1$, and a departure in either direction would be a violation. Read as lower probabilities, the two may sum to less than 1, which expresses a wide interval. Only a sum above 1 is a violation, and it occurs at a rate that depends on elicitation (§D.5.2). This reading is more charitable to a model. It does not by itself say more about whether the model is reliable.

Coherence status. The violation can be stated exactly. No distribution gives a claim and its negation probabilities that sum above one, and a model that held such a pair as lower probabilities would accept a combination of bets that loses whatever happens.

Theorem 19 (Complement Sums Above One Admit a Dutch Book)**.** *If an LLM's verbalized confidences for complementary claims satisfy $c(A) + c(\neg A) > 1$, then interpreting c as lower probability yields incoherent assessments exploitable via Dutch book.*

Proof. Let $s = c(A) + c(\neg A) - 1 > 0$ and $0 < \epsilon < s/2$. Read as lower probabilities, $c(A)$ and $c(\neg A)$ are supremum buying prices, so by A2 and A4 the model accepts $\mathbf{1}_A - (c(A) - \epsilon)$ and $\mathbf{1}_{\neg A} - (c(\neg A) - \epsilon)$. Buying both costs $1 + s - 2\epsilon$ and returns exactly 1, since exactly one of A and $\neg A$ obtains. By A4 the constant gamble $-(s - 2\epsilon) < 0$ is then desirable, contradicting A1. The sure loss approaches s as ϵ falls. □

What the lower-probability reading adds. Less than it may seem. That $c(A) + c(\neg A) > 1$ signals an inconsistency, that it can be checked on one instance without labels, and that the excess $c(A) + c(\neg A) - 1$ grades it, all follow from precise coherence, which requires the sum to equal 1. Label-free checks of this kind are already in use (Fluri et al., 2023), and the same property can train a probe on activations (Burns et al., 2023). The lower-probability reading does two narrower things. It names the failure: $\underline{P}(\neg A)$ exceeds $\overline{P}(\neg A) = 1 - \underline{P}(A)$, so no distribution meets both bounds. And it weakens the test: a pair such as $(0.01, 0.01)$ passes as a wide interval, where precise coherence would flag it. Meeting the constraint is not enough either. On Llama's incoherent pairs, post hoc rescaling to meet it removes overconfidence and adds no information (§D.5.2).

What our test of the link showed. We tested whether incoherent pairs, those whose two confidences sum above 1.05, are worse calibrated (Prediction 2, §D.5.2). Under a prompt that stated the statement's truth value before asking for confidence, Llama-3.1-70B's incoherent pairs had 2.21× the ECE of its coherent ones. Most of that came from the prompt. Of its 569 incoherent pairs, 404 put both confidences at 1.0. Since about half the statements are true, those pairs have an ECE near 0.5 by construction, and without them the ratio is 1.46×. On the first 400 statements, where the ratio is 2.76×, a neutral prompt lowers incoherence from 24.3% to 5.0% and the ratio to 1.26× [0.61, 2.22], an interval that includes 1. The rate of violation and its link to calibration are properties of the elicitation as much as of the model, and a coherence check is only as informative as the prompt that produced the pair.

Connection to ECE. When overconfidence leaves $c(A)+c(\neg A) > 1$, the pair violates coherence, and the same confidences may also raise ECE. The two need not move together: ECE is an average over instances and coherence a property of a single pair, so any link between them is empirical, not definitional.

Connection to scoring rules. Can an imprecise forecast be scored for accuracy as a precise one is? Not by a proper scoring rule. Seidenfeld et al. (2012) show that no continuous strictly proper scoring rule exists for imprecise probability forecasts, and Mayo-Wilson and Wheeler (2016) show that the impossibility is robust. It survives weakening the assumptions on the forecaster's decision rule, on the values a score may take, and on the form of the forecasts, and a strictly proper rule would have to be discontinuous on a dense set of forecasts. So a stated confidence read as a lower probability cannot be scored for accuracy the way a point forecast is. This section checks such confidences for coherence, and where it measures calibration, by ECE, it treats each stated confidence as a point value.

Verdict. Verbalized confidence can violate coherence, at a rate that depends on how it is elicited (§D.5.2). Only a pair of confidences summing above one is exploitable. Overstating a single confidence is miscalibration, not incoherence. Whether coherence violations *cause* miscalibration, or both stem from a common source such as RLHF training dynamics, remains open, and our evidence on their association is inconclusive (Prediction 2).

5.3 Ensembles as Empirical Credal Sets

Ensemble methods such as deep ensembles (Lakshminarayanan et al., 2017) and MC dropout (Gal and Ghahramani, 2016) generate K distributions $P_1, \ldots, P_K$ over the same outcomes. Their convex hull is a credal set, and forming it needs no further choice. Self-consistency (Wang et al., 2023) is different: it samples answers from one distribution and takes the plurality, the fixed-threshold case of §3 with exact string groups.

Definition 20 (Ensemble Credal Set)**.** *Given K distributions $P_1, \ldots, P_K$, the* ensemble credal set *is:*

$$\mathcal{M}_{ens} = conv\{P_1, \ldots, P_K\}$$

The lower and upper previsions of a gamble f under $\mathcal{M}_{\text{ens}}$ are $\min_k \mathbb{E}_{P_k}[f]$ and $\max_k \mathbb{E}_{P_k}[f]$, and for an event A they are the lower and upper probabilities $\underline{P}(A) = \min_k P_k(A)$ and $\overline{P}(A) = \max_k P_k(A)$. These are coherent by the Lower Envelope Theorem, and taking the convex hull changes neither, since an expectation is linear in the distribution.

The difficulty is in aggregation. For prediction, current practice averages: $\hat{P} = \frac{1}{K}\sum_k P_k$. That collapses the credal set to a point; disagreement survives only in a separate statistic, such as the spread across members, and the point carries none of it. If three members assign $P_1(A) = 0.9$, $P_2(A) = 0.5$, $P_3(A) = 0.8$, the average $\hat{P}(A) = 0.73$ suggests moderate confidence, while the interval $[0.5, 0.9]$ shows that one member is undecided and another highly confident. Worse, averaging can destroy structural properties (such as judgments of independence) that all ensemble members share (Elkin and Wheeler, 2018). Reporting the full credal set or at minimum the interval $[\underline{P}(A), \overline{P}(A)]$ would preserve imprecision for downstream decision rules (§6.1). Kudukkil Manchingal et al. (2026) keep such a set for language models (§6.4).

Verdict. Ensemble methods construct credal sets and, for prediction, collapse them to a point by averaging. That averaging discards disagreement is well known; the imprecise-probability reading names the alternative, keeping the set, and whether keeping it helps is an empirical question.

5.4 Conformal Prediction: Complementary, Not Equivalent

Conformal prediction (Shafer and Vovk, 2008; Vovk et al., 2005), applied to language models by Mohri and Hashimoto (2024) and Quach et al. (2024), provides distribution-free guarantees of marginal coverage. Given a calibration set and a nonconformity score, it outputs prediction sets Γ_α that satisfy $P(Y \in \Gamma_\alpha) \geq 1 - \alpha$, where Y is the true answer and the probability is over the calibration data and the test instance together. The guarantee needs no distributional assumption beyond exchangeability. Conformal prediction stands to imprecise probability differently from the three methods above. It does not construct one of the theory's objects, and the two answer different questions.

Conformal prediction (coverage, across instances): "What prediction set will contain the true answer with probability at least $1 - \alpha$, over repeated draws of the calibration data and a test instance?" This is a *coverage* guarantee about long-run frequencies.

Imprecise probability (coherence, within instance): "What probability bounds can I hold for this instance without exposing myself to sure loss?" The question is put in the behavioral interpretation we adopt. The guarantee is *coherence*, of which avoiding sure loss is the minimal part, and it says nothing about frequencies.

Neither subsumes the other. A conformal set can have exact coverage while the model's own confidences in its members are incoherent, for instance confidences in a claim and its negation that sum above one. Conversely, the lower and upper probabilities of a credal set are always coherent and need have no coverage guarantee. The distinction parallels that between frequentist confidence intervals and Bayesian credible intervals.

The two are nonetheless connected. Cella and Martin (2022) show that a stronger notion of prediction validity requires imprecise probability, and achieve it by reading conformal output as the contour function of a possibility measure, which determines a credal set. Caprio et al. (2025) show that the classical conformal prediction region is the imprecise highest density region of that credal set and keeps the same coverage guarantee. So conformal prediction can be read as producing a credal set, of a different kind from those above, whose distinctive guarantee is coverage; like any credal set, its lower and upper probabilities are also coherent. We do not pursue that reading here.

	Conformal Prediction	Imprecise Probability
Core question	Coverage across instances	Coherence within instance
Guarantee type	$P(Y \in \Gamma_\alpha) \geq 1 - \alpha$	Coherence
Reference class	Population of instances	Single instance
Requirements	Exchangeability + calibration data	Assessed bounds or a credal set
Failure mode	Miscoverage	Sure loss
Output	Prediction set Γ_α	Lower and upper previsions

Table 12: Conformal prediction vs. imprecise probability: different questions, different guarantees

Verdict. Conformal prediction and imprecise probability answer different questions and give different guarantees (Table 12), though on the reading of Cella and Martin (2022) conformal output determines a credal set.

5.5 Summary Table

Table 13 summarizes the mapping, and Figure 4 in Appendix C gives a visual overview.

Method	Implicit Construct	Coherent?	What It Reveals
Token entropy	Degenerate credal set (singleton)	✓(trivial)	Unreported width
Semantic entropy	Singleton in a threshold credal set; admissible partition	With agreement shares fixed	Implicit threshold commitment
Verbalized confidence	Lower probability $[\underline{P}, 1]$	Can violate	Overconfidence on complements is incoherent
Ensembles	Finitely generated credal set $\text{conv}\{P_k\}$	✓	Aggregation discards imprecision
Conformal	Coverage sets, a complement	Different notion	Coverage $\neq$ coherence

Table 13: LLM uncertainty methods and imprecise-probability constructs

A note on mapping types. The mappings vary in character. Token and ensemble methods map to credal sets by construction: the closed convex hull of the distributions they produce is a credal set. Semantic entropy maps to a singleton in a threshold credal set by our construction, and the admissible partition is a choice function on that set, coherent with agreement shares fixed on the full sample. Verbalized confidence maps to lower probability by interpretation: we read a stated confidence as a lower probability, on which confidences in a claim and its negation that sum above one violate coherence, though one could instead treat verbalized confidence as a miscalibrated point estimate. Conformal prediction is a complement rather than a mapping, reflecting different foundations (frequentist coverage vs. betting coherence). So the mapping is uneven in strength.

The imprecise-probability framework gives these methods one vocabulary. It also gives criteria for locating failures: coherence violations and inappropriate aggregation can be identified from a method's construction. Whether a given violation degrades performance is an empirical question.

6 Discussion

The paper has done two things. It has built a credal set for the threshold of semantic entropy and tested what that set adds, and it has read four other methods in the same vocabulary. This section takes up what follows from both. We first answer an objection to working with sets and intervals at all, that a decision needs a single number (§6.1). We then put the reading in a form a practitioner can use, a short list of checks (§6.2), state what the construction and the evidence do not establish (§6.3), place the work beside related frameworks (§6.4), and say what evidence would change our assessment (§6.5).

6.1 From Credal Sets to Decisions

A common objection is that intervals are not actionable: a decision needs a single number. It does not. The decision rules of §2.3.1, E-admissibility, maximality, interval dominance, Γ-maximin, and Γ-maximax, each choose among options on a credal set without first collapsing it to a point (Troffaes, 2007; Walley, 1991). The admissible partition is such a rule, and its output is an action: keep or reject each response. Where a number is wanted for a gamble f, the lower prevision $\underline{P}(f)$ gives a cautious bound, and the interval $[\underline{P}(f), \overline{P}(f)]$ shows how far the bounds are apart; a midpoint gives one number and hides the width.

Our experiments exercised these rules only in part. On the threshold credal set, E-admissibility and maximality had to coincide with the admissible partition, because each question has six equally weighted responses. The rules that could differ did: Γ-maximax, plurality at the loosest threshold, had a higher admissibility ratio than Γ-maximin, plurality at the strictest, on four of five benchmarks (§4.3, post hoc).

6.2 A Diagnostic Checklist

For practitioners evaluating LLM uncertainty methods, the imprecise-probability framework suggests a diagnostic checklist. It describes what a method does. Whether a method that fails a check performs worse is an empirical question, on which our evidence (Section 4 and Appendix D) is limited.

1. **Points vs. intervals:** Does the method report intervals or single values? A single value leaves its sensitivity to temperature, prompt, and threshold unreported. An interval makes that sensitivity visible and does not thereby flag errors: in our data, the width of the entropy range does not flag majority-vote error (Table 27).

2. **Coherence check:** For verbalized confidence, verify $c(A) + c(\neg A) \leq 1$ (a small tolerance, e.g., ≤ 1.05, accommodates rounding). A pair that violates it incurs sure loss when the confidences are read as lower probabilities. ECE averages over responses and does not detect a violation on a single pair, and in our data rescaling Llama's incoherent pairs to sum to one lowered their ECE and left AUROC near chance (§D.5.2, post hoc).

3. **Agreement between measures:** Compare verbalized confidence with semantic entropy on the same question. High confidence together with high entropy would show that the two have come apart (§D.2); we do not test this.

4. **Threshold sensitivity:** If the method uses a threshold (clustering, conformal), how much do results vary with it? In our data the clustering threshold mattered mostly at the strict end, and leaving it open added little (§4.3).

6.3 Limitations

Describes, does not explain. Imprecise probability describes how elicited uncertainty is structured, not why a model's elicited distributions vary as they do; that needs other tools, such as interpretability and the study of training dynamics. And imprecise probability forecasts have no continuous strictly proper scoring rule (§5.2), so there is no standard accuracy score on which one credal set beats another.

Limits of the admissible partition. The limits are of three kinds. In theory, the admissible partition is a coherent choice function only with agreement shares fixed on the full sample (Theorem 14, Proposition 15). Its one guarantee, that every admitted response is in a largest cluster at some threshold, follows from its definition and is not a finding (Proposition 11). And its threshold range is a hyperparameter, so leaving the threshold open moves the choice to the endpoints. In use, its rejected pool can be thin, its low rejection rate owes much to tied singletons under equal weights, and it fails when the model's most frequent answer is wrong, since it then rejects the correct minority. In evidence, its effect varies with benchmark and scorer, and the registered format test failed on NQ-Open (§4). Our findings about the sweep were made after the fact and await replication on fresh data, among them that it lowers the rejection rate more than it raises the ratio, and the three leads that §6.5 names for registration.

The design cannot separate the rules. With six equally weighted responses per question, E-admissibility and maximality coincide with the admissible partition on every family of nested clusterings (Section 3), so our experiments could not have told these three apart. Testing them needs more responses per question or unequal weights.

Empirical evidence is limited. The registered tests cover one construction, on short-answer and multiple-choice QA, with small open base models. Our two frontier runs are underpowered and point in opposite directions, and we did not measure prompt perturbation credal sets, which we regard as better grounded than temperature ones. Prediction 2 results (Appendix D) apply only to responses where numeric confidence could be extracted; extraction failure rates vary widely (27% for Llama, 39% for Gemma, 84% for Mistral), and the analyzed subset may not be representative of all model responses.

Coherence is diagnostic, not determinative. Our evidence that coherence violations predict miscalibration is inconclusive (Prediction 2), and coherence alone neither guarantees good performance nor is strictly required for it. A coherent model can be confidently wrong; an incoherent model can still rank its right answers above its wrong ones. The criteria diagnose how a method is constructed; they do not certify how it performs. Separately, a strategically deceptive model could exhibit perfect coherence while concealing capabilities; our framework diagnoses structural properties of elicited probabilities, not intentions.

6.4 Related Frameworks

We group related work by what it shares with ours: frameworks that also work with a credal set, or with a distribution over distributions, work on the same coherence check, other sampling-based measures of semantic uncertainty, and interventions on calibration.

Bayesian deep learning. Standard Bayesian approaches maintain distributions over model parameters, yielding a posterior predictive distribution over outputs (Blundell et al., 2015; Wilson and Izmailov, 2020). This is a special case of imprecise probability where the credal set is a singleton (the posterior predictive). However, Bayesian neural networks rarely produce exact posteriors. Approximations like variational inference or MC dropout (Gal and Ghahramani, 2016) induce *sets* of distributions (one per sample or dropout mask). These sets can be read as credal sets, though a Bayesian reads them as samples from one posterior, to be averaged. The distinction is commitment: a Bayesian analysis fixes one prior and so one posterior predictive; an imprecise analysis does not fix one, and may work with a set of priors, a credal set, or lower previsions directly.

Evidential deep learning. Sensoy et al. (2018) place a Dirichlet distribution on the class probabilities, enabling epistemic uncertainty estimation. The connection to imprecise probability is that a region of high Dirichlet density can be read as a credal set (it is convex when every concentration parameter is at least 1), with the concentration parameters controlling its diameter. Our reading is not restricted to parametric families.

Credal deep learning and imprecise probability for LLMs. Caprio et al. (2024) construct credal sets for deep learning. Several recent papers bring imprecise probability to language models. Yang et al. (2026) elicit probability intervals and credal sets from LLMs by prompting. Kudukkil Manchingal et al. (2026) induce a credal set from an ensemble of LoRA adapters and use interval dominance to decide when to commit to an answer, both at the answer token and over meanings of open-ended completions. Garces Arias et al. (2025) compare credal sets of model and human variation in creative generation. In each the imprecision is over distributions or probability values. None treats the clustering threshold of semantic entropy as the source of imprecision, which is the construction of Section 3.

Conformal prediction. As discussed in §5.4, conformal methods provide complementary guarantees with different foundations. The formal bridge described there (Caprio et al., 2025; Cella and Martin, 2022) reads conformal output as determining a credal set.

Consistency checks without labels. The closest prior work on our coherence check lies outside the imprecise-probability literature. Fluri et al. (2023) test language-model forecasts for negation and other logical consistency where ground truth is unavailable, and Burns et al. (2023) find a direction in activations on which a statement and its negation take opposite truth values, with no labels. Both rest on precise coherence. Prediction 2 asks the same kind of question of verbalized confidence, and what the lower-probability reading changes is limited (§5.2).

Sampling-based semantic signals. KLE (Nikitin et al., 2024) generalizes semantic entropy through a similarity kernel, and SNNE (Nguyen et al., 2025) aggregates pairwise similarities without clustering. Both replace the threshold with another parameter (KLE's kernel and its hyperparameters, SNNE's scale factor) and return a score per prompt. With the heat kernel e^{-tL} and the normalization Nikitin et al. give, a direct calculation shows that KLE has a limit that is semantic entropy: as $t \to \infty$, e^{-tL} converges to the projection onto the null space of the graph Laplacian L, which the indicator vectors of the graph's connected components span, and KLE tends to the entropy of those components, each weighted by its share of the responses. That is semantic entropy for equally weighted responses under our clustering rule, with two responses linked when the edge between them has positive weight; where our threshold decides which pairs are linked, t decides how far agreement spreads along chains of linked responses. Semantic Entropy Probes (Kossen et al., 2024) amortize the sampling cost by predicting SE from hidden states with linear probes, and inherit its dependence on the clustering criterion. The admissible partition differs in kind: it labels responses. It does not compete on per-prompt scores. On TriviaQA, KLE and SNNE have higher AUROC than the scalar diagnostic of §4.4 (0.786 and 0.788 against 0.773, post hoc). The approaches can be combined, for instance by training probes to predict admissibility.

Sensitivity to settings. Others report that sampling-based uncertainty scores depend on settings that nothing in the problem fixes. Cecere et al. (2025) find that the best sampling temperature for semantic entropy and related methods shifts across models and datasets, and that drawing the

temperature from a range instead reaches parity with the best fixed temperature chosen on the test set. Abdaljalil et al. (2025) ablate the similarity threshold of their embedding-based clustering and report that it changes detection AUROC. Our results fit both: the choice of threshold matters, and using the whole range of thresholds matches, without beating, a single threshold chosen on other labeled questions (§4.3).

Calibration interventions. Temperature scaling (Guo et al., 2017) fits a single temperature after training by minimizing negative log-likelihood on a validation set. In imprecise-probability terms it selects one point of the temperature credal set: it improves calibration and leaves unreported how much the calibrated value depends on the temperature.

Uncertainty-aware RLHF. Recent work addresses the tension between RLHF and calibration that Appendix D.1 reads in credal terms. Leng et al. (2024) identify systematic reward model bias toward high-confidence responses and propose PPO modifications (PPO-M, PPO-C) that calibrate rewards during training. Zhai et al. (2024) develop uncertainty-penalized RLHF using diverse reward ensembles, explicitly penalizing optimization in regions of high reward uncertainty. Leng et al.'s methods target the direction of the change in stated confidence, which is what matters for calibration on the account of §D.1. Tian et al. (2024) fine-tune toward factuality using the model's own confidence, estimated without references, as the preference signal. All can be read in credal terms, as resisting collapse, and none needs that reading.

6.5 Future Directions

If the imprecise-probability reading is to give more than a shared vocabulary, it needs evidence that a credal set measures something that agreement among sampled responses does not. Four directions follow.

Imprecise weights. With precise weights, every rule on the threshold credal set acts through the agreement profile alone (§3.6), and the null over thresholds leaves the weights as the next place to look. The imprecision can be put there through prompt perturbation credal sets (Definition 16), which we did not measure, or through ensembles. We measured the temperature credal set only on first tokens (§D.5).

Long-form generation and registered follow-ups. On short answers most similarities fall outside the threshold range, so the clustering rarely changes with the threshold (§4.3). The threshold credal set should be tested where similarity is graded and decisions attach to individual claims, as in long-form generation. Three post hoc leads from these data should be registered before they are tested on fresh data: a range that starts below 0.50, the direct comparison of the agreement profile with a single threshold chosen on other labeled questions, and the profile's gain on pairs where the plurality changes with the threshold.

Sequential reasoning. Chain-of-thought can be read as a sequence of credal sets, one per reasoning step. How do they evolve, and when does multi-step reasoning reduce uncertainty rather than propagate errors?

Principled training and inference. Training objectives could preserve credal structure, penalize incoherence, or couple semantic with verbal uncertainty, and inference methods could maintain and report credal sets rather than point estimates. Both require translating the diagnostic reading into optimization targets or decoding rules.

These directions extend the reading toward prescription, which this paper does not support.

7 Conclusion

Semantic entropy commits to a criterion for when two responses mean the same thing. When clusters are formed from pairwise similarities, the criterion is a threshold, and nothing in the problem fixes it. A credal set leaves it open. The joint distributions that pair the response distribution with a distribution over thresholds form a credal set on one outcome space, and fixed-threshold practice is one extreme point of it. The endpoints of the entropy range are coherent lower and upper previsions. The standard decision rules give response-level partitions, among them plurality at either endpoint threshold and the union of pluralities across the range, which is the admissible partition. With agreement shares fixed, the admissible partition is a coherent choice function. Re-run on subsets, it violates Sen's contraction condition, and so does plain plurality. With six equally weighted responses, E-admissibility over the whole credal set and maximality coincide with the admissible partition, so our experiments could not tell the three apart.

The experiments put two questions to this construction. The first was whether admissible responses are more often correct than rejected ones. On free-form factual QA they are, by a registered 1.62× on NQ-Open (short of its 2.0× criterion) and by 2.98× on the TriviaQA discovery set, and 12 of 20 registered hypotheses are supported, three of them only under TruthfulQA's substring scorer. The second was whether leaving the threshold open adds anything to a single threshold, and post hoc comparisons on the same data say that it adds little. The choice of threshold does matter: it can change how well plurality separates correct from incorrect responses. But on held-out questions the full agreement profile changes response-level AUROC by $-.004$ to $+.007$ against a single threshold chosen on other labeled questions, with no interval reaching above .023. The reason is that the range seldom changes anything: on most free-form question-model pairs the clustering is the same at every threshold in the range. And what the union over thresholds adds to plurality at the loosest threshold is mostly tied singletons. The exception is the question-model pairs where the plurality changes with the threshold, which the sweep finds without labels. There the profile beats the threshold chosen for all pairs by .054 in AUROC on TriviaQA (16% of its pairs) and, under the lenient registered letter scorer, by .026 on ARC-Challenge, and by about zero elsewhere. The width of the entropy range does not flag errors on free-form QA. Three further predictions were not registered (Appendix D): those on instruction tuning and on temperature are not supported, and the evidence on coherence and calibration is inconclusive.

For practice, a threshold chosen on other labeled questions and the whole range did about equally well, and outside TruthfulQA neither added much to a fixed 0.50. Under the scorers of the registered runs, separation was lost mainly at the strict end of the range, where the similarities of matching answers often fall short of the threshold. These are findings about these benchmarks and this NLI model rather than rules.

For token probabilities, verbalized confidence, ensembles, and conformal prediction, the same vocabulary is mostly definitional: it says what each method constructs and whether the result is coherent, and it locates one check that needs no labels, which precise coherence already implies. The result for the imprecise-probability reading of LLM uncertainty is a qualified null. Imprecision over what counts as agreement added little here, and the construction shows why it could not

have added much: with precise weights every rule could act only through the agreement profile. Whether imprecision in the weights on responses does better is open. The evidence has limits of its own: most of our runs use small open models, our frontier runs are underpowered, and the post hoc findings need replication on fresh data. Evidence that a credal set measures something that agreement among sampled responses does not would change this assessment. We found hints of it only post hoc, on pairs where the plurality changes with the threshold. On TriviaQA a threshold chosen for those pairs alone matches a profile fit on the same pairs, so the sweep there marks questions that one global threshold serves poorly; on ARC-Challenge, under the lenient registered letter scorer and on matrices that only partly reproduce the registered partition, no such threshold closes the gap (.043 [.020, .065]).

Acknowledgments

The author used Claude (Anthropic) throughout this work, for drafting and editing text, writing analysis code, and checking results against stored data. The author directed the work and takes full responsibility for its content.

References

Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. SINdex: Semantic INconsistency index for hallucination detection in LLMs. *arXiv preprint arXiv:2503.05980*, 2025.

Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to Imprecise Probabilities*. Wiley, 2014.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1985.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1613–1622, 2015.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal Bayesian deep learning. *Transactions on Machine Learning Research*, 2024.

Michele Caprio, Yusuf Sale, and Eyke Hüllermeier. Conformal prediction regions are imprecise highest density regions. In *Proceedings of the Fourteenth International Symposium on Imprecise Probabilities: Theories and Applications*, volume 290 of *Proceedings of Machine Learning Research*, pages 47–59, 2025.

Nicola Cecere, Andrea Bacciu, Ignacio Fernández Tobías, and Amin Mantrach. Monte Carlo temperature: a robust sampling strategy for LLM's uncertainty quantification methods. *arXiv preprint arXiv:2502.18389*, 2025.

Leonardo Cella and Ryan Martin. Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, 2022. doi: 10.1016/j.ijar.2021.07.013.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Jasper de Bock and Gert de Cooman. Interpreting, axiomatising and representing coherent choice functions in terms of desirability. In *Proceedings of the 11th International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, volume 103 of *Proceedings of Machine Learning Research*, pages 125–134, 2019.

Gert de Cooman, Arthur Van Camp, and Jasper de Bock. Desirable sets of things and their logic. In *Proceedings of the 13th International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA)*, 2023.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937. English translation in H.E. Kyburg and H. Smokler (eds.), Studies in Subjective Probability, Wiley, 1964.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lee Elkin and Gregory Wheeler. Resolving peer disagreements through imprecise probabilities. *Noûs*, 52(2):260–278, 2018. Published online 2017.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. arXiv preprint arXiv:2306.09983, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

Esteban Garces Arias, Julian Rodemann, and Christian Heumann. The geometry of creative variability: How credal sets expose calibration gaps in language models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP)*, pages 36–50. Association for Computational Linguistics, 2025.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Shireen Kudukkil Manchingal, Sofiia Nikolenko, and Fabio Cuzzolin. Credal large language models for semantic commitment under uncertainty. *arXiv preprint arXiv:2608.23244*, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:452–466, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. *arXiv preprint arXiv:2410.09724*, 2024.

Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, MA, 1980.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Jack Lindsey. Emergent introspective awareness in large language models. arXiv preprint arXiv:2601.01828, 2026.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Conor Mayo-Wilson and Gregory Wheeler. Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, 93(1):55–78, 2016.

Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2907, 2015.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4530–4540, Vienna, Austria, 2025.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Frank P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Kegan Paul, London, 1931. Written in 1926.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, 2020.

Jay K. Satia and Roy E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Coherent choice functions under uncertainty. *Synthese*, 172(1):157–176, 2010.

Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53(8):1248–1261, 2012.

Amartya K. Sen. Choice functions and revealed preference. *The Review of Economic Studies*, 38 (3):307–317, 1971.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2024.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3), 2025. arXiv:2412.05563.

Cedric A. B. Smith. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society, Series B*, 23:1–37, 1961.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.08401.

Matthias C. M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

Matthias C. M. Troffaes and Gert de Cooman. *Lower Previsions*. Wiley and Sons, Chichester, West Sussex, 2014.

Arthur Van Camp. *Choice Functions as a Tool to Model Uncertainty*. PhD thesis, Ghent University, 2018.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2–3):125–148, 2000.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.

Gregory Wheeler. A gentle approach to imprecise probability. In Thomas Augustin, Fabio Cozman, and Gregory Wheeler, editors, *Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld*, Theory and Decision Library A, pages 37–67. Springer, 2022.

Gregory Wheeler. Choice functions for semantic uncertainty – pre-registration. Open Science Framework, `https://osf.io/ady3k`, 2026a. Registered 2026-04-09.

Gregory Wheeler. Choice function admissibility on Natural Questions: A pre-registered free-form QA replication. Open Science Framework, `https://osf.io/78pb6`, 2026b. Registered 2026-04-18.

Gregory Wheeler. Choice function admissibility on TruthfulQA: A pre-registered test of the Goldilocks-zone hypothesis. Open Science Framework, `https://osf.io/dqbmg`, 2026c. Registered 2026-04-15.

Peter M. Williams. Notes on conditional previsions. Technical report, University of Sussex, 1975. Reprinted in *International Journal of Approximate Reasoning* 44(3):366–383, 2007.

Peter M. Williams. Indeterminate probabilities. In Marian Przełęcki, Klemens Szaniawski, and Ryszard Wójcicki, editors, *Formal Methods in the Methodology of Empirical Sciences*, pages 229–246. Ossolineum, Warsaw, 1976.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:4697–4708, 2020.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Anita Yang, Krikamol Muandet, Michele Caprio, Siu Lun Chau, and Masaki Adachi. Verbalizing LLM's higher-order uncertainty via imprecise probabilities. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 7761–7779, 2026.

Marco Zaffalon and Enrique Miranda. Axiomatising incomplete preferences through sets of desirable gambles. *Journal of Artificial Intelligence Research*, 60:1057–1126, 2017.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward LoRA ensembles. *arXiv preprint arXiv:2401.00243*, 2024.

A Proofs of Main Results

This appendix has two parts. The first returns to Theorem 18, which §5.1 proves in brief. It gives the proof with each step checked, verifies three properties that every coherent lower probability has, and extends the theorem to a range of temperatures. The second proves the results of Section 3, in the order in which that section states them. Theorem 19 is proved where it is stated, and the Lower Envelope and representation theorems are cited. The proofs rely on standard results from imprecise probability, particularly Walley (1991).

A.1 Proof of Theorem 18: Token Interval Coherence

Theorem 18 (Token Interval Coherence, restated)**.** *Let $\Pi(\pi_0)$ be a finite non-empty set of prompt perturbations. The induced probability intervals $[\underline{P}(t), \overline{P}(t)] = [\min_{\pi\in\Pi} P(t|\pi), \max_{\pi\in\Pi} P(t|\pi)]$ define coherent lower and upper probabilities.*

Proof. Define the credal set

$$\mathcal{M}_{\text{prompt}} = \text{conv}\{P(\cdot|\pi) : \pi \in \Pi(\pi_0)\}$$

where conv denotes the convex hull.

We verify that $\mathcal{M}_{\text{prompt}}$ satisfies the requirements for a credal set:

1. **Non-empty:** $\Pi(\pi_0)$ is non-empty by assumption, so $\mathcal{M}_{\text{prompt}}$ contains at least one distribution.

2. **Closed:** For finite Π, $\mathcal{M}_{\text{prompt}}$ is a polytope inside the probability simplex over the vocabulary $\mathcal{V}$, hence closed.

3. **Convex:** By construction as a convex hull.

The Lower Envelope Theorem (Walley, 1991, Theorem 3.3.3) establishes that $\underline{P}(f) = \inf_{P\in\mathcal{M}} \mathbb{E}_P[f]$ defines a coherent lower prevision for any closed, convex credal set $\mathcal{M}$. Specializing to indicator functions $f = \mathbf{1}_A$, we obtain

$$\underline{P}(A) = \inf_{P\in\mathcal{M}} P(A), \quad \overline{P}(A) = \sup_{P\in\mathcal{M}} P(A)$$

which define coherent lower and upper probabilities.

Since $P \mapsto P(A)$ is linear, its infimum and supremum over the convex hull of finitely many distributions are attained at one of them, so:

$$\underline{P}(t) = \min_{\pi\in\Pi} P(t|\pi), \quad \overline{P}(t) = \max_{\pi\in\Pi} P(t|\pi)$$

As a check, they have three properties every coherent lower probability has (necessary for coherence, not sufficient):

1. **Bounds:** $0 \le \underline{P}(t) \le \overline{P}(t) \le 1$ since each $P(\cdot|\pi)$ is a probability distribution.

2. **Conjugacy:** $\underline{P}(A) + \overline{P}(\neg A) = 1$ follows from:

$$\begin{aligned}\underline{P}(A) + \overline{P}(\neg A) &= \inf_P P(A) + \sup_P P(\neg A) \\ &= \inf_P P(A) + \sup_P (1 - P(A)) \\ &= \inf_P P(A) + 1 - \inf_P P(A) = 1\end{aligned}$$

3. **Superadditivity:** For disjoint A, B:

$$\begin{aligned}\underline{P}(A \cup B) &= \inf_P P(A \cup B) = \inf_P [P(A) + P(B)] \\ &\geq \inf_P P(A) + \inf_P P(B) = \underline{P}(A) + \underline{P}(B)\end{aligned}$$

□

Remark. The theorem extends to infinite index sets when the map to distributions is continuous and the index set compact, as for the temperature credal set over $[\tau_{\min}, \tau_{\max}]$; the extremes are then attained by compactness.

A.2 Results on the Threshold Credal Set and the Admissible Partition

Throughout, $\mathcal{R} = \{r_1, \ldots, r_M\}$ and its weights are fixed, with $\sum_r p_r = 1$ and $p_r > 0$, and s is symmetric. Write G_θ for the graph with edge (i, j) when $s(r_i, r_j) \geq \theta$. Cluster probabilities, u_θ, and O^*_θ are as in §3.

Proof of Proposition 6. For $\theta < \theta'$, every edge of $G_{\theta'}$ is an edge of G_θ, so each component of $G_{\theta'}$ lies in a component of G_θ. Hence $C^{\theta'}(r) \subseteq C^\theta(r)$ and $u_{\theta'}(r) \leq u_\theta(r)$, and the largest cluster probability cannot rise with θ. Merging blocks of a partition cannot raise entropy, by the grouping property, so $\mathrm{SE}(\theta) \leq \mathrm{SE}(\theta')$. Two different clusterings in the chain differ in their number of blocks, which lies between 1 and M, so at most M distinct clusterings occur. □

Proof of Proposition 8. $\mathcal{M}_\Theta$ is a closed convex set of probability mass functions, and $\mathbb{E}_{p\otimes\lambda}[g_r] = \sum_\theta \lambda(\theta) u_\theta(r)$ is linear in λ, so its extremes occur at point masses. Proposition 6 places them at $\theta_{\max}$ and $\theta_{\min}$. Coherence follows from the Lower Envelope Theorem, applied to the lower prevision on all gambles and restricted to these. □

Proof of Proposition 9. Since $u_\theta(\omega) \geq p_\omega > 0$, the gamble h is bounded. Clusters partition $\mathcal{R}$, so $\sum_\omega p_\omega(-\log u_\theta(\omega)) = \sum_k P_\theta(C_k)(-\log P_\theta(C_k)) = \mathrm{SE}(\theta)$. Expectation under $p \otimes \lambda$ is linear in λ, so the previsions occur at point masses, and by Proposition 6 at the endpoints. Coherence follows as in Proposition 8. □

Proof of Proposition 11. If $r \in C_{\mathrm{sem}}(\mathcal{R})$, then $r \in O^*_\theta$ for some $\theta \in \Theta$, and $u_\theta(r)$ equals the maximum by the definition of O^*_θ. If $r \notin C_{\mathrm{sem}}(\mathcal{R})$, then $r \notin O^*_\theta$ at each θ, so its cluster falls strictly below the maximum there. Any member of O^*_θ attains that maximum and belongs to $C_{\mathrm{sem}}(\mathcal{R})$. □

The dominating response may differ across thresholds, so this is weaker than dominance by one alternative.

Proof of Proposition 12. By Proposition 8, $\overline{P}(g_r) = u_{\theta_{\min}}(r)$, so Γ-maximax selects $O^*_{\theta_{\min}}$; likewise $\underline{P}(g_r) = u_{\theta_{\max}}(r)$, so Γ-maximin selects $O^*_{\theta_{\max}}$. Both endpoints belong to Θ, so both sets lie in $C_{\mathrm{sem}}(\mathcal{R})$. Each $p \otimes \delta_\theta$ belongs to $\mathcal{M}_\Theta$, which gives $C_{\mathrm{sem}}(\mathcal{R}) \subseteq \mathrm{EA}$. If r maximizes $\sum_\theta \lambda(\theta) u_\theta(\cdot)$, no r' can beat it at every θ, which gives $\mathrm{EA} \subseteq \mathrm{Max}$. If $r \in \mathrm{Max}$ and some r' had $\min_\theta u_\theta(r') > \max_\theta u_\theta(r)$, then r' would beat r at every θ, which gives $\mathrm{Max} \subseteq \mathrm{ID}$. □

Remark. The set Max is the weak Pareto frontier of $\{u_\theta\}$, by definition. C_{sem} and EA relate to the Pareto frontier in neither direction, while ID contains the Pareto frontier, since it contains Max. The two middle inclusions can be strict. Take $\Theta = \{0.5, 0.9\}$ and $s = 0.1$ except as stated. (i) Weights $(.2, .2, .25, .22, .13)$ on (a, x, b, r, w) with $s(a, x) = s(r, w) = 0.6$ give $C_{\mathrm{sem}} = \{a, x, b\}$;

r, with scores $(.35, .22)$, is beaten at $\theta = 0.5$ by a and at $\theta = 0.9$ by b, by no single response at both, and it is optimal for $\lambda(0.5) \in [3/13, 2/7]$. So $C_{\text{sem}} \subsetneq \text{EA}$. (ii) Weights $(.2, .2, .25, .22, .08, .05)$ on (a, x, b, r, w, z) with $s(a, x) = s(r, w) = 0.6$ give $C_{\text{sem}} = \text{EA} = \{a, x, b\}$, while r, with scores $(.30, .22)$, belongs to Max and to the Pareto frontier. So $\text{EA} \subsetneq \text{Max}$. (iii) In the worked example of §3, r_C is in C_{sem} and is Pareto-dominated by r_A.

Proof of Theorem 13. Edges are present exactly when $s \geq \theta$. Take $\theta < \theta'$ in $(b_k, b_{k+1}]$. Every similarity in the range is some b_j, so no similarity lies in $[\theta, \theta')$, and $G_\theta = G_{\theta'}$. If no similarity equals $\theta_{\min}$, the same holds for $\theta = \theta_{\min}$ and the first cell. Otherwise $\{\theta_{\min}\}$ is its own cell, since the edge at $\theta_{\min}$ is present there and absent above. Equal graphs give equal components, cluster probabilities, O^*, and SE, which proves constancy.

Let $\hat{O}_c$ be the constant value of O^* on cell c. Since O^* is determined by the clustering, a grid meeting every clustering gives $C_{\text{sem}} = \bigcup_c \hat{O}_c$, which does not depend on the grid and equals the union over the whole range. Let $d \leq \binom{M}{2}$ be the number of distinct similarities in the range. If none equals $\theta_{\min}$, there are at most $d + 2$ breakpoints and at most $d + 1$ cells. If one does, there are at most $d + 1$ breakpoints, hence at most d half-open cells, plus $\{\theta_{\min}\}$. Proposition 6 gives at most M distinct clusterings, so one threshold per clustering recovers the union. A grid missing a clustering omits the corresponding set from the union, so it admits a subset. □

Proof of Theorem 14. Each $\mathcal{D}_\theta$ is coherent. If $f \leq 0$, then $\mathbb{E}_{p \otimes \delta_\theta}[f] \leq 0$, and f is not a non-zero non-negative gamble, so $f \notin \mathcal{D}_\theta$ (A1). A2 holds by construction. A3 and A4 hold because expectation is linear, the sum of a gamble with positive expectation and a non-negative gamble has positive expectation, and the sum of two non-zero non-negative gambles is non-zero and non-negative.

$\mathcal{K}_\Theta$ is coherent. For a coherent $\mathcal{D}$, $\mathcal{K}_\mathcal{D} = \{A : A \cap \mathcal{D} \neq \emptyset\}$ satisfies K0–K4: $0 \notin \mathcal{D}$ gives K0 and K1, A2 gives K2, A3 and A4 give K3, and K4 is immediate. Each of K0–K4 is preserved under intersection.

The induced choice function. Write $g_A = \{g_v : v \in A\}$. Option g_r is rejected from g_A when $g_A - g_r \in \mathcal{K}_\Theta$, that is, when at each θ some g_v with $v \in A$ has $g_v - g_r \in \mathcal{D}_\theta$. Suppose $g_v - g_r \geq 0$. Since $g_r(r, \theta) = 1$, also $g_v(r, \theta) = 1$, so $r \in C^\theta(v)$ and the two clusters coincide. This holds at every θ, so $g_v = g_r$. Hence no response gamble is a non-zero non-negative shift of another, and $g_v - g_r \in \mathcal{D}_\theta$ exactly when $u_\theta(v) - u_\theta(r) = \mathbb{E}_{p \otimes \delta_\theta}[g_v - g_r] > 0$. So r is rejected from A exactly when it is strictly below $\max_{v \in A} u_\theta(v)$ at every θ, which gives the stated C_Θ. If $r \in C_\Theta(A)$ and $r \in B \subseteq A$, then r maximizes u_θ over A at some θ, hence over B, which is Sen's α. □

Remark. Responses in the same cluster at every $\theta \in \Theta$ define the same gamble and receive the same label. The argument also covers E-admissibility over all of $\mathcal{M}_\Theta$: intersect over all λ in place of the point masses.

Proof of Proposition 15. Take five responses with equal weights. Let $s = 0.9$ within $\{r_1, r_2, r_3\}$ and within $\{r_4, r_5\}$, and $s = 0.1$ between these groups. Let $\Theta = \{0.5\}$. On $\mathcal{R}$ the clusters are the two groups, with probabilities $3/5$ and $2/5$, so $r_1 \in C_{\text{sem}}(\mathcal{R})$. Let $A = \{r_1, r_4, r_5\}$ with weights renormalized over A. The clusters are now $\{r_1\}$ and $\{r_4, r_5\}$, with probabilities $1/3$ and $2/3$, so $r_1 \notin C_{\text{sem}}(A)$. □

Remark. Sen's α is recovered for the re-run map in one case. Let $A \subseteq \mathcal{R}$ be a union of $\theta_{\min}$-clusters of $\mathcal{R}$, hence by Proposition 6 a union of θ-clusters for every $\theta \in \Theta$. Then the clusters of A at each θ are clusters of $\mathcal{R}$, with probabilities rescaled by a common factor, so a cluster that is maximal in $\mathcal{R}$ and lies in A remains maximal in A. Hence $r \in C_{\text{sem}}(\mathcal{R})$ and $r \in A$ imply $r \in C_{\text{sem}}(A)$, and the same argument applies to nested pairs of such subsets.

B Choice Functions: Technical Details

Section 2.3.2 introduced sets of desirable option sets without their axioms. This appendix gives the definitions and axioms in full, states the representation theorem of de Bock and de Cooman (2019), defines the choice function that a set of desirable option sets induces, and relates these to the admissible partition of Section 3.

B.1 From Credal Sets to Desirable Gambles

We begin with sets of desirable gambles, which §2.2.2 introduced, and their relation to credal sets.

Definition 21 (Gamble)**.** *A* gamble *on outcome space Ω is a bounded real-valued function $f : \Omega \to \mathbb{R}$. Gambles represent uncertain rewards: if outcome ω obtains, the agent receives $f(\omega)$.*

Definition 22 (Set of Desirable Gambles)**.** *Write $\mathcal{L}(\Omega)$ for the set of all gambles on Ω. A* set of desirable gambles $\mathcal{D} \subseteq \mathcal{L}(\Omega)$ *collects the gambles an agent would accept in exchange for the status quo. The status quo is the zero gamble, and coherence excludes it from $\mathcal{D}$, so $f \in \mathcal{D}$ says that the agent strictly prefers f to zero.*

Definition 23 (Coherence for Desirable Gambles)**.** *A set $\mathcal{D}$ is* coherent *if it satisfies A1–A4 of §2.2.2:*

A1. If $f \leq 0$, then $f \notin \mathcal{D}$ (avoid partial loss)

A2. If $f \geq 0$ and $f \neq 0$, then $f \in \mathcal{D}$ (accept partial gain)

A3. If $f \in \mathcal{D}$ and $\lambda > 0$, then $\lambda f \in \mathcal{D}$ (positive scaling)

A4. If $f, g \in \mathcal{D}$, then $f + g \in \mathcal{D}$ (combination)

A coherent set $\mathcal{D}$ is *strictly desirable* if each $f \in \mathcal{D}$ that is not a non-zero non-negative gamble satisfies $f - \delta \in \mathcal{D}$ for some $\delta > 0$. Strict desirability is the condition under which a set of desirable gambles and a credal set carry the same information.

Theorem 24 (Equivalence)**.** *For finite Ω, sets of strictly desirable gambles correspond bijectively to non-empty closed convex credal sets (Walley, 1991; Zaffalon and Miranda, 2017):*

- *Given $\mathcal{M}$: $\mathcal{D}_{\mathcal{M}} = \{f : \inf_{P \in \mathcal{M}} \mathbb{E}_P[f] > 0\} \cup \{f : f \geq 0, f \neq 0\}$*
- *Given $\mathcal{D}$: $\mathcal{M}_{\mathcal{D}} = \{P : \mathbb{E}_P[f] \geq 0$ for all $f \in \mathcal{D}\}$*

B.2 Why Sets of Gambles

A set of desirable gambles says of each gamble separately whether it is desirable. It cannot say that at least one of two gambles is desirable while leaving open which one (§2.3.2). Option sets add that disjunctive form, and the admissible partition uses it. Under $\mathcal{K}_\Theta = \bigcap_\theta \mathcal{K}_{\mathcal{D}_\theta}$ (Theorem 14), an option set is desirable when, at every threshold, one of its options is desirable, and which one may vary with the threshold.

B.3 Sets of Desirable Option Sets

Definition 25 (Option Set)**.** *An* option set *is a finite subset $A \subseteq \mathcal{L}(\Omega)$ of gambles.*

Definition 26 (Desirable Option Set)**.** *An option set A is* desirable *if at least one of its elements is desirable, leaving open which. A set of desirable option sets $\mathcal{K}$ collects the option sets judged desirable, and we write $A \in \mathcal{K}$.*

Definition 27 (Coherence for $\mathcal{K}$)**.** *A set of desirable option sets $\mathcal{K}$ is* coherent *if it satisfies the following, which are equivalent to the axioms of de Bock and de Cooman (2019):*

K0. If $A \in \mathcal{K}$ then $A \setminus \{0\} \in \mathcal{K}$

K1. $\{0\} \notin \mathcal{K}$

K2. $\{u\} \in \mathcal{K}$ for all $u \geq 0$, $u \neq 0$ (partial gains)

K3. Combination: If $A, B \in \mathcal{K}$ and, for each $f \in A$ and $g \in B$, $(\lambda_{f,g}, \mu_{f,g}) \geq 0$ is non-zero, then $\{\lambda_{f,g} f + \mu_{f,g} g : f \in A, g \in B\} \in \mathcal{K}$

K4. If $A \subseteq B$ and $A \in \mathcal{K}$, then $B \in \mathcal{K}$ (supersets)

B.4 The Representation Theorem

Theorem 28 (de Bock and de Cooman 2019, Theorem 9)**.** *A set of desirable option sets $\mathcal{K}$ is coherent if and only if there exists a non-empty collection $\{\mathcal{D}_i\}$ of coherent sets of desirable gambles such that:*

$$\mathcal{K} = \bigcap_i \mathcal{K}_{\mathcal{D}_i}$$

where $\mathcal{K}_{\mathcal{D}} = \{A : A \cap \mathcal{D} \neq \emptyset\}$ is the binary set of desirable option sets induced by $\mathcal{D}$.

So an option set A is desirable exactly when it contains a gamble from every $\mathcal{D}_i$ in the collection, and the gamble may differ from one $\mathcal{D}_i$ to the next. The collection is not unique; the largest is the set of all coherent $\mathcal{D}$ with $\mathcal{K} \subseteq \mathcal{K}_{\mathcal{D}}$.

B.5 Choice Functions and Rejection

Definition 29 (Choice Function)**.** *For a finite non-empty set of options A, the choice function induced by $\mathcal{K}$ is:*

$$C(A) = A \setminus R_{\mathcal{K}}(A)$$

where $R_{\mathcal{K}}(A)$ is the rejected set:

$$R_{\mathcal{K}}(A) = \{u \in A : A - u \in \mathcal{K}\}$$

and $A - u = \{v - u : v \in A\}$.

So an option u is rejected from A when the option set $A - u$ is desirable: under each $\mathcal{D}_i$ of a representing collection, some option in A is strictly preferred to u, and that option may vary with $\mathcal{D}_i$.

B.6 Relation to the Admissible Partition

On the fixed outcome space $\mathcal{R} \times \Theta$ of §3, each response r is represented by its gamble g_r, and the theory applies directly. Holding agreement shares fixed on $\mathcal{R}$, Theorem 14 gives a coherent $\mathcal{K}_\Theta$ of the form in Theorem 28, whose choice function agrees with the admissible set on $\mathcal{R}$. Re-running the construction on a subset is a different map. It changes the agreement shares, since each P_θ is computed from the responses at hand, and by Proposition 15 it violates Sen's α, which coherent choice functions satisfy. As for representation, $\mathcal{K}_\Theta$ itself is neither mixing nor Archimedean when $|\Theta| \geq 2$, because each $\mathcal{D}_\theta$ accepts every partial gain; accepting only gambles of positive expectation gives a model that is both and induces the same C_Θ, E-admissibility over $\{p \otimes \delta_\theta\}$ (de Bock and de Cooman, 2019, Theorem 29 and Proposition 24).

B.7 Computational Complexity

For a grid Θ and M responses, the admissible partition takes $O(|\Theta| \cdot M^2)$ time after the similarity matrix: at each threshold, connected components take $O(M^2)$ and finding the largest cluster takes $O(M)$. The similarity matrix itself requires $M(M-1)$ directed NLI calls.

C Further Background

This appendix collects background that the main text uses only in passing: a survey of LLM uncertainty methods (§C.1), a fuller introduction to imprecise probability for an ML audience (§C.2), including the epistemic status of temperature credal sets, a figure summarizing the reading of Section 5 (§C.3), and worked examples for that section (§C.4).

C.1 LLM Uncertainty Methods

This subsection surveys existing approaches to LLM uncertainty quantification. It is not a comprehensive review: it sets out the methods that Section 5 reads in one vocabulary. Shorinwa et al. (2025) give a fuller treatment.

C.1.1 Token-Level Methods

The most direct approach to uncertainty quantification derives from the model's own output distribution. Given a context c, a language model defines a probability distribution $P(T|c)$ over the next token T, which ranges over the vocabulary $\mathcal{V}$. This distribution carries information about how confident the model is in different continuations.

Predictive entropy measures the spread of this distribution (Malinin and Gales, 2018):

$$H(T|c) = -\sum_{t \in \mathcal{V}} P(t|c) \log P(t|c)$$

When entropy is high, probability mass spreads across many tokens and the model is uncertain which to produce. When entropy is low, mass concentrates on a few tokens and the model is confident. The related measure of perplexity, $\exp(H)$, is the effective number of tokens the model is choosing among.

Token-level methods extend to sequences by the chain rule of probability. The entropy of a full response decomposes into the sum of conditional entropies at each position.

The limitation of token-level methods is that they conflate linguistic variation with semantic uncertainty. Asked "What is the capital of France?", a model may answer "Paris", "It is Paris", or

"The capital is Paris". Token entropy is high because probability spreads across these phrasings, yet all of them give the same answer. Variation in wording is not uncertainty about the answer, and token-level methods cannot tell the two apart (Kuhn et al., 2023).

C.1.2 Verbalized Confidence

Rather than deriving uncertainty from output distributions, verbalized methods elicit the model's own assessment of its reliability. The simplest version prompts the model directly: "How confident are you in your answer? Give a percentage." A related method, the P(True) of Kadavath et al. (2022), asks the model whether a proposed answer is true and takes the probability it assigns to the answer "True"; the confidence is then read from the model's output distribution, not stated.

Verbalized confidence has two attractions. If a model knows something about its own reliability, asking for it is more direct than inferring it from output distributions. And a user can interpret a stated confidence of 85% more readily than an entropy of 1.3 bits.

The challenge is calibration. A model is well-calibrated if its stated confidence matches empirical accuracy: when the model says 80% confident, it should be correct roughly 80% of the time across many such statements (Guo et al., 2017). Miscalibration occurs when stated confidence systematically diverges from accuracy, most commonly in the form of overconfidence. Expected Calibration Error (ECE) quantifies this gap by binning predictions by confidence and measuring the average difference between confidence and accuracy within each bin (Naeini et al., 2015). ECE has known limitations (sensitivity to binning, small-sample bias), but remains the standard metric for verbalized confidence calibration; we use it in §D.5, while the coherence check of §5.2, which is already in use (Fluri et al., 2023), is complementary and does not depend on ECE.

Xiong et al. (2024) report mean expected calibration errors above 0.3 for verbalized confidence on four of five models and describe verbalized confidence as overconfident.

Reinforcement learning from human feedback (RLHF) is reported to make the problem worse. Xie et al. (2024) state, citing earlier studies, that pre-trained models have well-calibrated conditional probabilities and that calibration degrades significantly after RLHF; their subject is token-level probability, not stated confidence. Leng et al. (2024) show that reward models exhibit systematic bias toward high-confidence responses, and fine-tuning against such reward models trains the model to express confidence regardless of whether it is warranted. The result is a trained-in disconnect between expressed and actual uncertainty.

C.1.3 Ensemble and Sampling Methods

Ensemble methods generate multiple predictions and quantify uncertainty through their disagreement. The intuition is that if different predictions agree, the model is confident in a robust answer; if they disagree, uncertainty is higher.

Deep ensembles (Lakshminarayanan et al., 2017) achieve this by training multiple models independently and comparing their outputs. Each model may converge to slightly different solutions due to random initialization and training order, and the variance across ensemble predictions measures epistemic uncertainty about which solution is correct.

Monte Carlo dropout (Gal and Ghahramani, 2016) approximates this approach more cheaply by exploiting dropout regularization. Rather than training multiple models, one keeps dropout active at test time and performs multiple forward passes, each with a different random dropout mask. The variation across these passes stands in for ensemble disagreement without the cost of training multiple models.

Self-consistency (Wang et al., 2023) applies similar reasoning to chain-of-thought prompting (Wei et al., 2022). Generate multiple reasoning chains for the same question and examine whether they converge on the same answer. If different reasoning paths lead to the same conclusion, confidence in that answer increases; if they lead to different conclusions, uncertainty is higher.

These methods face two challenges. The first is computational cost: generating multiple predictions requires multiple forward passes, and training true ensembles multiplies the training budget. The second is aggregation: given diverse predictions, how should we combine them? For prediction, standard practice averages, and the average discards the magnitude and structure of disagreement. If three ensemble members predict Paris with confidence 0.9 and two predict Lyon with confidence 0.8, the average obscures this bimodal structure.

C.1.4 Conformal Prediction

Conformal prediction gives distribution-free guarantees on the coverage of prediction sets (Shafer and Vovk, 2008; Vovk et al., 2005). Given a user-specified error rate α and a calibration dataset, conformal methods output prediction sets Γ_α satisfying

$$P(Y \in \Gamma_\alpha) \geq 1 - \alpha$$

where Y is the true outcome. This coverage guarantee is marginal, over the calibration data and the test instance together, and holds with no distributional assumptions beyond exchangeability between calibration and test data.

The method rests on a nonconformity score, which measures how unusual a candidate prediction is. Given calibration examples, we compute nonconformity scores for the true labels and use their distribution to threshold predictions on new examples. Any candidate with a nonconformity score below the calibrated threshold enters the prediction set.

Quach et al. (2024) apply conformal prediction to language modeling, producing sets of plausible continuations with guaranteed coverage. Mohri and Hashimoto (2024) develop conformal methods for factuality, ensuring that generated claims are true with high probability. We cite representative work from a growing literature. These methods give formal guarantees that the other approaches lack.

The limitation is that coverage guarantees do not imply useful predictions. A conformal set achieving 90% coverage might do so by including many implausible answers along with the correct one. The sets can become uninformatively large on difficult questions. And conformal methods give sets, not probability distributions: they indicate which predictions are plausible but not how probability distributes among them.

C.1.5 Open Challenges

Three patterns recur in this work.

Overconfidence is widely reported, and RLHF can make it worse (§C.1.2): pre-trained models are often reasonably calibrated, and fine-tuned ones overstate their confidence (Leng et al., 2024; Xie et al., 2024; Xiong et al., 2024).

Semantic uncertainty and verbalized confidence measure different things and need not agree. Semantic entropy detects confabulations better than P(True) does (Farquhar et al., 2024), and verbalized confidence is often overconfident (Xiong et al., 2024). A model can express high verbalized confidence while producing semantically diverse samples. One reading is that verbalized confidence and sampling variation come from largely separate processes, so that verbal reports do not give reliable access to the model's own uncertainty.

Methods disagree. Token entropy, semantic entropy, verbalized confidence, ensemble disagreement, and conformal sets all purport to measure uncertainty, but they give different answers and do not combine into one account. What each measures, why they differ, and how they relate remain open questions.

Section 2 introduces imprecise probability, and Section 5 states what each of these methods constructs in its terms.

C.2 Imprecise Probability for an ML Audience

We present imprecise probability through its behavioral interpretation, on which it is a generalization of the Bayesian theory of fair prices and not an exotic departure from it. On that interpretation, the path from standard Bayesian probability to imprecise probability requires one conceptual step. Section 2.2 gives the definitions the main text uses; this subsection gives the history and the interpretation.

C.2.1 From Bayesian Foundations

We begin with the canonical, "Bayesian" theory of subjective probability. On this view, probability statements represent an agent's degrees of belief about uncertain quantities. Degrees of belief must satisfy the axioms of probability. But why? The textbook answer comes from the betting arguments of Frank Ramsey and Bruno de Finetti (de Finetti, 1937; Ramsey, 1931).

The theory identifies your degrees of belief with your fair prices on a finite set of gambles (bets with uncertain payoffs). If your probability for an event A is 0.7, you regard a certain reward of \$0.70 as equivalent to the uncertain reward of \$1 if A and \$0 otherwise, thus committing you to buy or sell such a contract at that price. De Finetti (1937) showed that incoherent prices admit a *Dutch book*: an arbitrage that guarantees you a sure loss. The theorem: you avoid sure loss if and only if your prices satisfy the axioms of (finitely additive) probability.

This much is familiar. De Finetti's framework is built on *linear previsions*, the fair prices that commit you equally to buy or sell a gamble. The two-sided commitment forces your price to equal your expectation, and this linearity is what lets de Finetti treat previsions as primitive and probabilities as derived.[2]

But must we require such two-sided commitments?

C.2.2 A Simple Extension

C.A.B. Smith (1961) proposed a heretical idea within Bayesian orthodoxy: allow upper and lower odds on the same event to differ. Peter Williams (1975; 1976) took the key next step of recasting this as *lower previsions* (supremum buying prices) and *upper previsions* (infimum selling prices), generalizations of de Finetti's linear previsions. Where de Finetti required buying and selling prices to coincide, Williams allowed them to differ, capturing states where the evidence warrants withholding commitment in either direction.

Williams also developed *sets of desirable gambles* as an axiomatic foundation that sits under both linear and lower previsions. Desirability distills the core of the betting approach (sure-loss avoidance, dominance reasoning) and supplies minimal logical machinery for constructing an object (a convex cone of acceptable gambles) that is either coherent or not. Peter Walley (1991) later systematized these ideas into a comprehensive theory. Matthias Troffaes and Gert de Cooman

[2] De Finetti's dictum "Probability does not exist" expresses the companion view that probability is not an objective property of events.

Property	Bayesian	Imprecise Probability
Representation	Single prior P	Desirable gambles, or a credal set $\mathcal{M}$
Output	Expectation	Lower and upper previsions
Ambiguity	Must resolve	Can represent
Coherence	Automatic	By construction, or to be checked
Unresolved clustering threshold	One distribution over thresholds	Vacuous over thresholds

Table 14: Comparison of Bayesian and imprecise-probability frameworks

(2014) and colleagues have extended it, building the theory from sets of desirable gambles rather than from lower previsions. The progression extends de Finetti's move. He took previsions as primitive and derived probabilities from them; lower previsions generalize his linear previsions; and desirability sits beneath both, with credal sets as a derived representation.

Consider a gamble on an event A. Your maximum buying price b is the most you would pay for a contract paying \$1 if A; your minimum selling price s is the least you would accept to sell it. De Finetti requires $b = s$: a precise fair price. But you might reasonably set $b = 0.60$ and $s = 0.80$. The interval $(0.60, 0.80)$ represents prices between which you decline to transact in either direction, not from irrationality, but because the evidence does not warrant sharper commitment.

What constraints apply to one-sided prices? The betting argument generalizes in two steps. Your prices *avoid sure loss* if no finite combination of transactions you accept guarantees a loss. They are *coherent* if, in addition, no such combination implies a higher buying price for some gamble than the one you state (Walley, 1991). Both are weaker than de Finetti's requirement. A coherent agent need not accept every transaction; the gap between b and s is a region of permissible non-commitment that fair, two-sided prices do not allow.

Bounds on the probabilities of events do not suffice. Different credal sets can give every event the same probability bounds yet give some gamble different expectation bounds (Levi, 1980; Walley, 1991); Figure 3 shows a case. This is why the theory takes gambles, rather than events, as primitive. And when the question is which options to retain under an unresolved modeling choice (as with semantic clustering in LLMs), choice functions apply, operating directly on options. Imprecise probability thus encompasses a hierarchy of increasingly expressive representations:

$$\text{precise probability} \subset \text{credal sets} \subset \text{desirable gambles} \subset \text{choice functions}$$

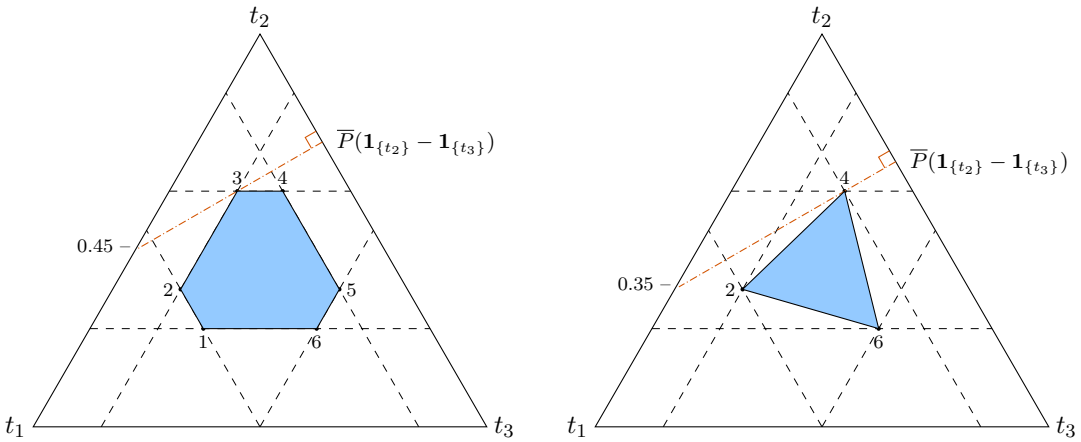
where each "$\subset$" denotes representation-theoretic containment (each level embeds into the next), not literal set inclusion.[3] The LLM uncertainty methods in §5 draw on different levels of this hierarchy, and Section 2.2 introduces each.

Table 14 summarizes the key differences between Bayesian and imprecise-probability approaches.

C.2.3 A Worked Credal Set

Example 1 (Temperature Credal Set)**.** *Consider Qwen2.5-7B completing "The capital of France is ___." Measured from its next-token logits, at temperature $\tau = 0.3$ it assigns $P(Paris) = 0.994$. At $\tau = 1.0$, this drops to 0.445. At $\tau = 1.5$, it falls further to 0.108. The three distributions span*

[3]These are different mathematical objects. The notation indicates that any uncertainty representation expressible at a given level can also be expressed at the next level, with the right-hand construct strictly more expressive.



(a) $\mathcal{M}_1$ (b) $\mathcal{M}_2$

Figure 3: Two uncertainty-aware language models with identical marginal token probability intervals can still disagree on comparative token rankings. The figure illustrates credal sets $\mathcal{M}_1 = \text{conv}\{p_1, \ldots, p_6\}$ and $\mathcal{M}_2 = \text{conv}\{p_2, p_4, p_6\}$ within the 2-simplex of probability mass functions over tokens $\{t_1, t_2, t_3\}$. The marginal intervals $[\underline{P}(\{t_i\}), \overline{P}(\{t_i\})]$, for $i = 1, 2, 3$ are identical in $\mathcal{M}_1$ and $\mathcal{M}_2$, yet the upper prevision for the gamble $f = \mathbf{1}_{\{t_2\}} - \mathbf{1}_{\{t_3\}}$ is 0.45 for $\mathcal{M}_1$ versus 0.35 for $\mathcal{M}_2$. Each credal set constrains the token probabilities jointly in its own way, which determines how far the model commits to $t_2 \succ t_3$. Marginal intervals do not determine the credal set.

a temperature credal set:

$$\mathcal{M}_{temp} = conv\{P_{0.3}, P_{1.0}, P_{1.5}\}$$

This yields $\underline{P}(Paris) = 0.108$ *and* $\overline{P}(Paris) = 0.994$*, with imprecision 0.886. Because Paris is the top token, whose probability falls monotonically as* τ *rises, additional temperatures between 0.3 and 1.5 would enlarge the credal set and leave these bounds unchanged. A token that is not on top can peak inside the range.*

The imprecision in Example 1 quantifies how much the model's confidence depends on a hyperparameter that nothing fixes. Current practice ignores this by reporting a single temperature's output.

Example 1 applies Definition 2 to a single event (the token "Paris"). But credal sets contain more information than their marginal probability bounds. As noted in §C.2.2, distinct credal sets can induce identical probability bounds on all events while disagreeing on expectation bounds for some gambles (Figure 3). The figure illustrates two credal sets, $\mathcal{M}_1$ and $\mathcal{M}_2$, with identical marginal intervals $[\underline{P}(\{t_i\}), \overline{P}(\{t_i\})]$ for each token, yet different upper previsions for the gamble $f = \mathbf{1}_{\{t_2\}} - \mathbf{1}_{\{t_3\}}$ that compares two tokens. Probability bounds on events constrain but do not determine the credal set; to recover $\mathcal{M}$ fully, we need expectation bounds on all gambles. This is the content of the Lower Envelope Theorem (Theorem 3).

C.2.4 Interpreting Coherence

Section 2.2.2 defined coherence. This subsection says how we read it: as a way of building models, as a diagnostic, and in terms of bets. It then answers the objection that a language model does not bet.

Constructive character. The desirability axiom system, A1–A4, is a normative standard for judging beliefs and also a *constructive* procedure for building coherent models from a finite set of initial assessments. You provide some gambles you are disposed to accept, add the gambles that A2 requires, close the result under A3 and A4 to extend your initial assessments to a full model of those commitments, then verify with axiom A1 that the model avoids partial loss.

Diagnostic framing. Coherence is binary: a set of assessments is coherent or it is not. Our claim is that *coherence violations are diagnostic*: identifying which axiom is violated reveals structural problems with the uncertainty representation. For example, when an LLM's verbalized confidence breaks the requirement $\underline{P}(A) + \underline{P}(\neg A) \leq 1$ (that is, $c(A) + c(\neg A) > 1$), whether this violation goes with miscalibration is an empirical question, on which our evidence is inconclusive (§D.5.2).

Behavioral interpretation. The behavioral interpretation of coherence, in terms of acceptable bets, is the one we adopt. Following de Finetti, we take betting behavior as primitive, bracketing questions about whether such behavior reflects "genuine beliefs" or some other internal state. This operationalist stance is particularly apt for LLMs, where the question "What does the model truly believe?" may be ill-posed or beside the point.

For imprecise probability, the operationalization involves one-sided commitments. A model's lower probability $\underline{P}(A)$ represents the supremum price at which it would hypothetically buy a gamble paying \$1 if A; its upper probability $\overline{P}(A)$ represents the infimum price at which it would hypothetically sell. When $\underline{P}(A) < \overline{P}(A)$, the gap represents uncertainty the model declines to price, a region of permissible non-commitment. Coherence constrains these one-sided commitments, but it does not require the two-sided commitment of a fair price.

Some coherence violations, such as $c(A)+c(\neg A) > 1$, make the model's commitments exploitable for sure loss; the others show that its stated prices fall short of what its commitments imply. Whether this reflects "internal beliefs," patterns in training data, or artifacts of elicitation is a question we set aside.

Why coherence matters for non-betting agents. One might object: LLMs do not bet, so why care if their outputs violate betting-derived coherence axioms? We offer three answers.

First, *downstream decision-makers do bet.* When a medical system uses LLM confidence to triage patients, or a financial system uses it to assess risk, the stated probabilities inform decisions with stakes. An agent who acts on confidences that incur sure loss, such as confidences in a claim and its negation that sum above one, can be exploited, even if the LLM itself faces no stakes.

Second, *coherence violations can be checked without access to ground truth*, which makes them a candidate signal of unreliable uncertainty estimates. Our evidence on that link is inconclusive: Prediction 2 (§D.5.2) finds it in one model under a leading prompt, and under a neutral prompt its interval includes 1. We do not claim coherence is necessary or sufficient for reliability: a coherent system can be confidently wrong, and an incoherent one might still be useful if its outputs track correctness.

Third, *coherence is a structural property, not a psychological one.* De Finetti's coherence is a condition on a set of prices, not on the agent who states them. A set of one-sided prices either is the lower envelope of a set of probability measures or it is not, which is a mathematical fact. When we say LLM outputs are incoherent, we are making a structural claim about the elicited numbers, not a psychological claim about the model's internal states.[4]

[4]Indeed, some schools of Bayesianism (Jaynes, 2003) treat these constraints as logical rather than behavioral. For our practical purposes here, these denominational differences do not matter.

C.2.5 What Temperature Credal Sets Represent

One might object: "Temperature variation isn't epistemic uncertainty; it's a hyperparameter imposed externally. The model doesn't 'believe' different things at different temperatures; we're just rescaling its outputs." The objection is right about τ but mistakes the question. The coherence question for temperature credal sets is not "Does the model have genuine uncertainty about τ?" but rather: "Given that different temperatures elicit different betting prices from the model, what is the coherent representation of these elicitable commitments?"

The operationalist response. In de Finetti's framework, probabilities are defined by betting behavior, not by introspection into mental states. Temperature variation reveals that the model can be induced to offer different prices for the same gamble depending on τ. If we commit to any single P_τ, we implicitly claim precision that our elicitation procedure cannot support, since a different τ would yield different commitments. The credal set $\mathcal{M}_{\text{temp}}$ represents the range of distributions consistent with what we can coherently elicit, remaining agnostic about which point is "correct."

Hierarchy of epistemic strength. Prompt perturbation credal sets have the strongest grounding: semantically equivalent queries *should* yield consistent answers, so variation across $\Pi(\pi_0)$ represents failure of a normative constraint the model ought to satisfy. Temperature credal sets are weaker: variation across τ reveals sensitivity to a hyperparameter that lacks independent justification. We measure temperature credal sets, on first tokens only (§D.5), because they are cheap (varying τ requires only logit rescaling), while noting that the stronger construction is available when the stakes warrant additional inference cost. The coherence analysis applies to both: the lower envelope of a credal set is coherent however the set was constructed.

C.3 The Mapping at a Glance

Figure 4 summarizes the reading of Section 5.

C.4 Worked Examples for Section 5

Token probabilities. Example 1 gave the temperature credal set for one prompt at three temperatures. Table 15 gives the same prompt on Qwen2.5-7B at five, with the probabilities of two other tokens, measured from the model's next-token logits:

Temperature τ	P(Paris)	P(Lyon)	P(Marseille)
0.3	0.994	<0.001	<0.001
0.7	0.748	<0.001	<0.001
1.0	0.445	<0.001	<0.001
1.3	0.206	<0.001	<0.001
1.5	0.108	0.001	<0.001

Table 15: Next-token probabilities for "The capital of France is" on Qwen2.5-7B, measured from its logits at each temperature.

The temperature credal set yields $\underline{P}(\text{Paris}) = 0.108$ and $\overline{P}(\text{Paris}) = 0.994$, a width of 0.886, and Lyon and Marseille stay below 0.002 at every temperature. Reporting only $P_{1.0}(\text{Paris}) = 0.445$

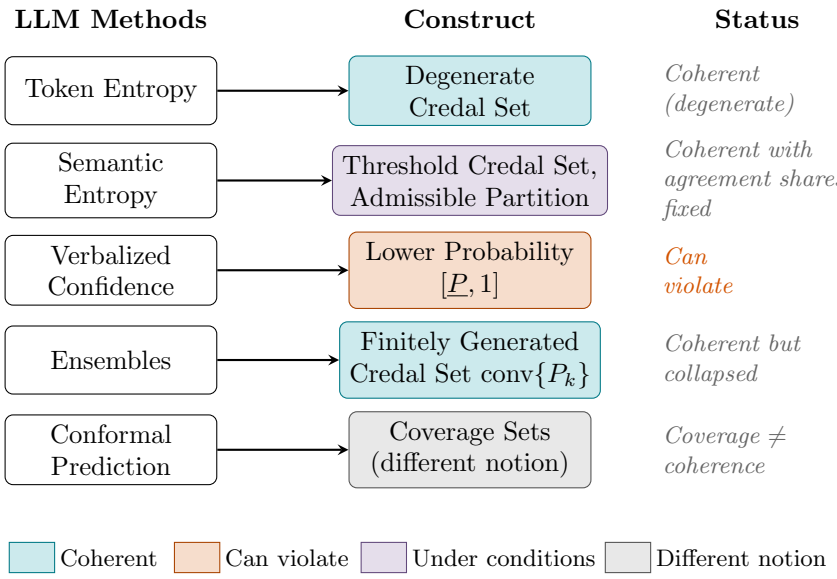


Figure 4: LLM uncertainty methods read as imprecise-probability constructs. Color indicates coherence status: teal for coherent constructions, orange for constructions that can violate coherence, purple for constructions coherent only under stated conditions, and gray for a different notion of validity. On this reading, token methods report one member of an elicitable set, fixed-threshold semantic methods use one extreme point of a threshold credal set, verbalized confidence can violate coherence, ensembles are coherent but poorly aggregated, and conformal prediction operates under a different notion of validity.

hides this dependence on an arbitrary choice. The example is one model on one prompt: on Llama-3.1-8B the most likely next token is "a" and $P_{1.0}(\text{Paris}) = 0.14$.

Hedging language. Phrases like "fairly confident" or "I think so" can be glossed as probability intervals. The values below are our illustration, not measurements:

- "I'm certain": $[0.95, 1.0]$
- "I'm fairly confident": $[0.70, 0.85]$
- "I'm not sure": $[0.30, 0.60]$

A method that maps such phrases to point values discards that width.

Verbalized confidence. For illustration, suppose a model is asked to evaluate two claims: "The Eiffel Tower is in Paris" (claim A) and "The Eiffel Tower is not in Paris" (claim $\neg A$).

The model is slightly underconfident about the true statement (0.92 stated vs. 0.95 frequency) but overconfident about the false negation (0.25 stated vs. 0.05 frequency). The overconfidence on the negation drives the coherence violation. Interpreting these as lower probabilities:

- $\underline{P}(A) = 0.92$: "I'm at least 92% confident A is true"
- $\underline{P}(\neg A) = 0.25$: "I'm at least 25% confident A is false"

Question	Stated confidence c	Frequency true p	Gap
"True" (A)	0.92	0.95	-0.03
"False" ($\neg A$)	0.25	0.05	$+0.20$
Sum	1.17	1.00	$+0.17$

Table 16: Verbalized confidence coherence violation, with illustrative values; p is the frequency with which claims like this one are true. The sum $c(A) + c(\neg A) = 1.17 > 1$ violates the coherence constraint $\underline{P}(A) + \underline{P}(\neg A) \leq 1$.

The sum exceeds 1, creating a Dutch book opportunity. An adversary offering both bets just below these prices collects nearly \$1.17 and pays out exactly \$1.00. The \$0.17 is the supremum of the sure loss.

D Other Predictions Tested

The paper states four predictions. Prediction 3, the admissible partition, is the subject of Section 4. The other three were not registered, and they bear on the reading less directly: Prediction 1 uses response diversity as a proxy for credal width, Prediction 2 tests a property that needs only precise coherence (§5.2), and Prediction 4 uses no imprecise-probability construct. Predictions 1 and 4 are not supported, and the evidence for Prediction 2 is inconclusive. We report them in full here, with the readings that motivated them and the harness defects that first made their effects look real. Table 4 lists their outcomes alongside the registered hypotheses.

D.1 A Credal Reading of RLHF-Induced Miscalibration

Reinforcement learning from human feedback (RLHF) (Bai et al., 2022; Christiano et al., 2017; Ouyang et al., 2022) improves response quality but can damage calibration (Leng et al., 2024; Xie et al., 2024). The imprecise-probability framework offers a reading of what changes.

D.1.1 The Phenomenon

Xie et al. (2024) state, citing earlier studies, that pre-training yields language models with well-calibrated conditional probabilities and that calibration degrades significantly after RLHF fine-tuning. For stated confidence, Leng et al. (2024) report that RLHF-trained models express confidence predominantly at the higher levels.

D.1.2 Imprecise-Probability Reading

The imprecise-probability framework reads this as credal set collapse. The reading is a model. Our runs do not measure it directly (§D.5.1).

Pre-training. We represent a pre-trained model's state as a credal set $\mathcal{M}_{\text{pre}}$ with more than one member and a substantial diameter, the greatest total-variation distance between two of its members, reflecting limited, conflicting training data.

RLHF training. Reward models trained on human preferences favor high-confidence responses (Leng et al., 2024). If raters likewise prefer "The answer is X" to "The answer might be X, but I'm not certain," the RLHF objective penalizes hedging. The same bias may feed sycophancy (Sharma et al., 2024), where models give agreeable responses in place of appropriate uncertainty.

Credal collapse. On this reading RLHF changes the model's *expressed* uncertainty and leaves its underlying uncertainty in place: training narrows the expressed credal set to $\mathcal{M}_{\text{RLHF}} \subseteq \mathcal{M}_{\text{pre}}$.

Model 1 (RLHF as Credal Diameter Reduction)**.** *We model the* effect *of RLHF on expressed uncertainty as a map taking each credal set $\mathcal{M}$ to a narrower set $\mathcal{M}' \subseteq \mathcal{M}$. This is a phenomenological characterization of observable behavior, not a mechanistic model of training dynamics.*

Assumptions:

1. *RLHF narrows expressed uncertainty: $\text{diam}(\mathcal{M}') < \text{diam}(\mathcal{M})$*
2. *Actual accuracy is unchanged (RLHF affects expression, not knowledge)*[5]

Consequence: *These assumptions do not by themselves imply worse calibration.*

What narrowing implies. Model 1 fixes only that the set shrinks. What changes calibration is how the reported value moves relative to accuracy. Suppose a reported confidence c was calibrated before training, so that accuracy given c equals c, and accuracy does not change. If each question's set shrinks about its reported value, calibration is unchanged. If reports are compressed toward a common value, $c' = \bar{c} + \lambda(c - \bar{c})$ with $0 < \lambda < 1$ and $\bar{c} = \mathbb{E}[c]$, then with bins fine enough to separate values of c', ECE rises to $(1 - \lambda)\,\mathbb{E}|c - \bar{c}|$, with zero mean signed gap. At $\lambda = 0$ the report is constant, so all of it falls in one bin: ECE is zero, and the report carries no information. This compression acts across questions, so it is narrowing in the model's sense only if each set contains $\bar{c}$ and contracts about it. And if a set shrinks toward its confident end, it both narrows and raises the report, and overconfidence grows with that shift. Published evidence fits the last case: RLHF models state confidence predominantly at the higher levels, where models trained only with supervised fine-tuning spread more widely, and reward models favor confident responses (Leng et al., 2024). So the credal reading describes the change and does not explain the damage. The reward bias supplies the direction. Our runs do not test this.

Recent work on uncertainty-aware RLHF proposes modifications that mitigate this effect. Leng et al. (2024) develop PPO variants that calibrate reward scores during training. Zhai et al. (2024) penalize optimization in regions of high reward uncertainty using ensemble disagreement. These approaches target the reward bias directly. They do not test the credal reading.

D.1.3 Testable Implication

Model 1 assumes accuracy is unchanged. Prediction 1 departs from that: questions where the pre-trained model exhibits wider credal sets should lose more accuracy, since selecting a point from a wide credal set leaves more room to sacrifice accuracy. Section D.4 states the prediction and §D.5.1 reports that it is not supported.

[5]This assumption is plausible for factual recall, where RLHF mainly shapes response style; §D.5.1 finds overall accuracy changes compatible with it. For reasoning tasks, RLHF may improve accuracy through chain-of-thought elicitation, requiring a refined model that accounts for genuine capability gains.

D.2 Semantic-Verbal Disconnect

Semantic uncertainty, measured by disagreement among sampled responses, and verbalized confidence can come apart. A model can express high verbal confidence while producing semantically diverse responses.

D.2.1 The Phenomenon

For illustration, when asked "Who wrote *Hamlet*?", a model might:

- State: "I'm 95% confident the answer is Shakespeare"
- Yet produce diverse samples: "Shakespeare" (4×), "William Shakespeare" (3×), "Christopher Marlowe" (2×), "Francis Bacon" (1×)

The high verbal confidence (95%) masks the disagreement among samples, of which 30% name a different author.

D.2.2 Imprecise-Probability Reading

The imprecise-probability framework reads verbal and semantic uncertainty as *different constructs over different spaces*.

Semantic uncertainty. When we sample responses and cluster by meaning, we obtain a distribution over semantic clusters. Each threshold gives one such distribution, over its own clusters; §3 places them on one outcome space, responses × thresholds, as a credal set. The spread of responses measures disagreement about which answer is correct.

Verbal uncertainty. When the model states confidence, it produces a numerical estimate. Repeated elicitation, with varied prompts or sampling, yields a spread of such estimates, and that spread measures uncertainty in self-assessment.

The disconnect. These two constructs need not align. Models have some self-knowledge (Kadavath et al., 2022), and some functional introspective awareness, which Lindsey (2026) finds highly unreliable and context-dependent. We do not know how far it reaches semantic uncertainty. A stated confidence may have been shaped by RLHF independently of the underlying semantic state. Verbal confidence can be narrow (confident self-assessment) while semantic uncertainty remains wide (diverse sampled responses).

D.2.3 Implications for Alignment

On this reading, RLHF shapes verbal confidence without reference to semantic uncertainty. Aligning the two would then differ from calibration in the standard sense, because it concerns two representations rather than one. We do not test this here.

D.3 Overconfidence as Coherence Violation

As established in §5.2, overconfidence on complementary claims is a coherence violation: when $c(A) + c(\neg A) > 1$, the model's stated confidences are inconsistent and a Dutch book can be constructed (Theorem 19). Here we highlight the connection to calibration metrics.

ECE (Expected Calibration Error) measures the average gap between stated confidence $c(A)$ and empirical accuracy $p(A)$:

$$\text{ECE} = \sum_b \frac{n_b}{n} \left| \bar{c}_b - \text{acc}_b \right|,$$

where bin b holds n_b of the n claims, $\bar{c}_b$ is their mean stated confidence, and acc_b is the share of them that are true; we use 10 equal-width bins (Appendix E.2).

When the stated confidences in a claim and its negation are inflated enough to sum above one, the model becomes vulnerable to a Dutch book. Overstatement that leaves the sum at or below one is miscalibration without incoherence. Still, one might expect high ECE to go with coherence violations. But calibration and coherence are distinct. Calibration is a frequentist property (long-run alignment); coherence is a single-instance property (no sure loss on this bet). An agent can be calibrated on average but incoherent on individual instances, or coherent everywhere but miscalibrated overall. The relationship is empirical rather than definitional, and our evidence on it is inconclusive (§D.5.2).

D.4 Predictions

We state four testable predictions. Predictions 1 and 3 come from the reading, Prediction 1 by departing from Model 1's assumption that accuracy is unchanged. Prediction 2 tests a condition that precise coherence already implies. Prediction 4 is a companion hypothesis that uses no imprecise-probability construct.

D.4.1 Prediction 1: Instruction Tuning Costs Accuracy Where Credal Sets Are Wide

Hypothesis: Questions on which the base model has a wider implicit credal set lose more accuracy under instruction tuning (RLHF, supervised fine-tuning, and other steps). We first stated the prediction this way. After seeing the first results we restated it: the relation is non-linear in base-model diversity, with damage concentrating at moderate diversity under concise tuning and at low diversity under verbose tuning. Section D.5.1 reports that neither form holds.

Test: For a dataset of questions, measure:

1. Pre-trained credal width proxy (e.g., response diversity of sampled responses)
2. Accuracy change: $\Delta\text{Acc} = \text{Acc}_{\text{instruct}} - \text{Acc}_{\text{base}}$

D.4.2 Prediction 2: Coherence Violations Cluster with Miscalibration

Hypothesis: Questions where models exhibit coherence violations (sure loss on complementary claims) should have worse calibration than questions where models are coherent.

Rationale: Exactly one of a claim and its negation is true, so a pair of confidences that sums above one overstates at least one of them, and overstatement is miscalibration.

Test: For binary questions, elicit $c(A)$ and $c(\neg A)$. Read as lower probabilities (§5.2), the two confidences avoid sure loss only if $c(A) + c(\neg A) \leq 1$; in practice, we allow a 5% tolerance for rounding (classifying as coherent if the sum ≤ 1.05) and compare ECE between coherent and coherence-violating subsets.

Prediction: $\text{ECE}_{\text{violating}} > \text{ECE}_{\text{coherent}}$.

Two further predictions:

Prediction 3: Choice function admissibility predicts answer quality. Responses in the admissible set C_{sem} (Definition 10) should have higher accuracy than rejected responses, since a rejected response falls below the largest cluster at every threshold in the range. Section 4 reports its tests.

Prediction 4: Temperature sensitivity varies with base-model diversity. The effect of temperature on accuracy should depend on how concentrated the base model already is. For questions where it is concentrated (low diversity), increasing temperature adds noise and degrades accuracy. For questions where it is already diffuse (high diversity), temperature increases have less to destroy. This concerns accuracy. Definition 17 concerns next-token probabilities and says nothing about how width varies with diversity. When we measure it, that width rises with diversity on first tokens and, in one noisy estimate, falls with it on whole responses (§D.5). The prediction: questions with low base-model diversity should show greater sensitivity to temperature than questions with high diversity. We first stated this prediction as optimal temperature varying with question difficulty, and restated it after seeing the results (§D.5.3).

D.5 Results

This subsection reports Predictions 1, 2, and 4. Section 4 reports Prediction 3, and Appendix E gives experimental details.

Does diversity track credal width? Prediction 1 uses response diversity as a proxy for credal width, which we had not checked, and Prediction 4 stratifies by it. We measured the temperature credal set of Definition 17 directly from next-token logits on the 500 TriviaQA questions of Prediction 1 (post hoc). The width of the top token's probability over $\tau \in [0.3, 1.5]$ rises with diversity on all four base models (Spearman ρ from 0.40 to 0.55), and the total-variation diameter of the set is more strongly associated with it (ρ from 0.67 to 0.79). From the lowest quartile of diversity (Q1) to the highest (Q4), mean top-token width goes as follows: 0.41 to 0.69 on Qwen, 0.42 to 0.63 on Mistral, 0.51 to 0.70 on Llama, 0.45 to 0.59 on Gemma. This covers the first token only, which need not be the answer (it is "The" on up to 61 of 500 questions). For whole responses the relation can reverse: on the Prediction 4 run, with one instruct model and three samples per temperature, so noisily, the width of the most frequent answer's frequency falls from 0.85 in Q1 to 0.62 in Q4 (Q1 minus Q4 [0.17, 0.29]). So Prediction 1 tests diversity, which is related to credal width and not the same thing.

D.5.1 Prediction 1: Instruction Tuning and Credal Width

Evidence status: Not registered, and all results below are exploratory. We first stated the prediction as linear, with damage growing with base-model diversity. The first run contradicted that, and we restated it after seeing the results, so the restated form is post hoc. The pairs compare base and instruction-tuned checkpoints and do not isolate RLHF from other tuning steps.

Protocol: Four base and instruction-tuned pairs (Qwen2.5-7B (Qwen Team, 2024), Mistral-7B (Jiang et al., 2023), Llama-3.1-8B (Dubey et al., 2024), and Gemma-2-9B (Gemma Team, 2024)) on the same $n = 500$ TriviaQA questions (Joshi et al., 2017). Response diversity, our proxy for credal width, is the proportion of unique normalized responses among 15 from the base model (3 samples $\times$ 5 temperatures: 0.3, 0.5, 0.7, 1.0, 1.3). Quartiles are computed per model.

Results as run: The stored runs compare the mean accuracy of those 15 base responses with one instruct response at $\tau = 1.0$. Table 17 gives the differences. They suggest two patterns: Qwen

loses at moderate diversity (Q3), and Mistral and Llama lose at low diversity (Q1). Across the 16 quartile tests, only Mistral Q1 and Llama Q1 pass Holm correction at 0.05; Qwen Q3 ($p = 0.015$) does not. Neither survives the checks below.

Model	Run	Q1 (low)	Q2	Q3	Q4 (high)	Overall
Qwen	Stored, no template	−1.7 [−4, 0]	−0.1 [−9, +8]	−9.0 [−16, −2]	0.0 [−4, +4]	−2.9
	Matched, no template	−0.8 [−3, +1]	+5.6 [−1, +12]	+1.1 [−4, +7]	+3.6 [−1, +8]	+1.5 [−1, +3]
	Matched, chat	−7.9 [−12, −4]	+1.7 [−7, +10]	+1.4 [−6, +8]	+6.3 [+2, +11]	−1.1 [−4, +2]
Mistral	Stored	−13.4 [−18, −9]	−5.3 [−13, +2]	−4.9 [−15, +5]	+1.2 [−6, +9]	−7.1
	Matched	−6.9 [−10, −4]	−6.1 [−14, +1]	−0.8 [−10, +8]	+9.1 [+2, +17]	−1.9 [−5, +1]
Llama	Stored, two tokens	−8.9 [−13, −5]	−2.4 [−10, +5]	+2.6 [−8, +13]	+5.0 [−2, +13]	−2.4
	Matched, two tokens	−4.6 [−8, −2]	−12.3 [−19, −6]	+1.0 [−7, +9]	+4.8 [−2, +11]	−2.5 [−5, 0]
	Matched, one token	−1.4 [−4, +1]	−4.8 [−11, +1]	+7.5 [0, +15]	+7.7 [+1, +15]	+1.9 [−1, +5]
Gemma	Stored	−0.5 [−2, +1]	+3.8 [−3, +11]	+3.9 [−2, +10]	+6.0 [−3, +15]	+2.4
	Matched	+0.2 [−2, +2]	−3.1 [−11, +5]	+3.7 [−3, +10]	+4.5 [−4, +13]	+1.1 [−1, +4]

Table 17: Instruct minus base accuracy (percentage points), by base-model diversity quartile, with 95% bootstrap intervals over questions. Exploratory. Stored rows use the stored responses and scorer. Matched rows use regenerated responses, three draws per arm at $\tau = 1.0$, and the repaired scorer. Quartiles come from the stored runs throughout.

Defects in the stored runs. There are five. The arms are unmatched: base accuracy averages over five temperatures and instruct accuracy rests on one draw, and the quartiles are formed from the same base responses whose accuracy is measured. Qwen-Instruct ran with no chat template. Llama-Instruct's prompt carried two begin-of-text tokens. Mistral's instruct checkpoint (v0.2) was tuned from a later base model than the v0.1 base used here, so that pair also differs in pretraining. The scorer cut each response at its first period, which truncates at abbreviations and initials, and capped normalized text at 100 characters, which mislabels long responses. On regenerated Mistral responses, the cap and the cut alone move instruct accuracy from 64.3% to 70.8%.

Matched rerun. We regenerated both arms locally: three responses per question at $\tau = 1.0$ from each base and instruct model, on all 500 questions, scored by one rule with the cut repaired and no cap. Quartiles stay those of the stored runs, so the new draws played no part in forming them. One difference remains: base models take a two-shot prompt and instruct models a direct one.

Qwen's Q3 loss disappears with matching alone: −9.0 as stored, +1.1 [−4, +7] matched on the same prompt, and +1.4 [−6, +8] with the chat template. Llama's Q1 loss falls from −8.9 to −4.6 [−8, −2] when matched and to −1.4 [−4, +1] with one begin-of-text token. Mistral's Q1 loss is −14.8 on matched responses under the stored scorer and −6.9 [−10, −4] under the repaired one. Qwen shows a Q1 loss under its chat template and none without it (−7.9 [−12, −4] against −0.8 [−3, +1]). Gemma shows none (+0.2 [−2, +2]). With prompts corrected, no model's Q3 interval lies below zero, and accuracy rises at Q4 in all four models, with intervals above zero in three. Overall, accuracy moves by −1.1 [−4, +2], −1.9 [−5, +1], +1.9 [−1, +5], +1.1 [−1, +4] points for Qwen, Mistral, Llama, and Gemma, with every interval covering zero, which is compatible with the assumption in §D.1 that accuracy is unchanged.

Stratification check. A question on which a model has low diversity is usually one it answers correctly. A second model that knows partly different things would tend to do worse there and

better at high diversity, with no tuning involved. We tested this two ways on the same draws. First, we replaced the instruct model with another base model under the same prompt. Llama base minus Mistral base, on Mistral's quartiles, gives −5.3 [−8, −3] at Q1 and +9.5 [+3, +16] at Q4, close to Mistral-Instruct's −7.0 and +8.6 on the same 498 questions. Among the six ordered pairs of Mistral, Llama, and Gemma base models, three have a Q1 interval below zero and two a Q4 interval above zero. Adding Qwen's base model, whose accuracy is 9.7 to 12.7 points below each of the others overall, gives six and five of twelve. Second, we kept instruct minus base and stratified on the mean diversity of the other three models, which does not favor either arm of the pair. Mistral's Q1 loss falls to −1.1 [−4, +2], and Llama and Gemma stay near zero (−0.2 and +0.2). Qwen's persists (−7.4 [−12, −3]). Mistral's intervals at Q2 and Q3 on these strata reach zero (−5.2 [−11, 0] and −6.9 [−14, 0]), so a moderate-diversity loss in Mistral is not excluded. Table 23 gives the full results.

Spread. The stored runs keep one instruct response per question, so they do not measure narrowing. The rerun allows a crude measure: the share of questions whose three draws are not all scored alike. On the other models' strata, instruct draws disagree less than base draws in 15 of 16 cells, with intervals below zero in five: Qwen Q3 (−11.0 [−21, −2]), Mistral Q3 (−12.3 [−22, −3]), Llama Q2 (−10.7 [−18, −4]), Gemma Q2 (−10.6 [−19, −2]), Gemma Q3 (−12.6 [−21, −4]). On a model's own strata the drops look larger, but another base model reproduces much of them (Mistral base minus Qwen base at Q2 is −18.6 [−34, −3], against Qwen-Instruct's −22.0 [−36, −8]), so we do not use those figures. The result on the other models' strata points toward narrowing without aggregate accuracy loss, as §D.1 assumes. It does not measure the credal set: it compares correctness labels across three draws, and the prompt differs between arms.

Status. The first-stated prediction is not supported: with prompts corrected, loss does not grow with diversity, and no model's Q4 interval lies below zero on either stratification. The restated form was post hoc, and its two patterns do not survive. Qwen's moderate-diversity loss came from unmatched arms, Llama's low-diversity loss largely from the token defect, and Mistral's is reproduced by a different base model and disappears on other models' strata. The split between concise and verbose tuning fails as well: under its chat template Qwen answers in a median of 3 words and loses at Q1, where the restated form assigned concise tuning to moderate diversity. One effect remains, a single exploratory result: Qwen-Instruct under its chat template loses accuracy on questions the other models find easy.

D.5.2 Prediction 2: Coherence and Calibration

Evidence status: Not registered.

Protocol: For each statement A from MMLU (Hendrycks et al., 2021), elicit $c(A)$, the stated confidence that A is true, and $c(\neg A)$, the stated confidence that it is false, separately via verbalized confidence prompting. We test three models: Llama-3.1-70B-Instruct (Dubey et al., 2024) (n = 3000 statements across 20 subjects), Gemma-2-9B-it (Gemma Team, 2024), and Mistral-7B-Instruct (Jiang et al., 2023) (both $n = 1000$ across 10 subjects).

Classification: Responses where both $c(A)$ and $c(\neg A)$ can be extracted as numeric values are classified as:

- Coherent: $c(A) + c(\neg A) \leq 1.05$ (5% tolerance)
- Incoherent: $c(A) + c(\neg A) > 1.05$

Pairs are excluded when either query lacks a parseable numeric confidence. Extraction failure rates vary substantially across models (27% for Llama, 39% for Gemma, 84% for Mistral), reflecting differences in how reliably each model follows verbalized confidence prompts.

Results: Table 18 presents the results. The models exhibit markedly different coherence profiles: Llama-3.1-70B produces coherent responses 74% of the time, while Gemma-2-9B and Mistral-7B are predominantly incoherent (9% and 4% coherent, respectively).

Model	n_{coh}	n_{incoh}	% Coherent	ECE (coh)	ECE (incoh)	Ratio
Llama-3.1-70B	1611	569	74%	22.7%	50.2%	2.2×
Gemma-2-9B	56	551	9%	9.1%	38.1%	4.2×
Mistral-7B	6	158	4%	26.7%†	34.6%	1.3×†

Table 18: Expected Calibration Error by coherence status, under the original prompt. †Mistral's coherent sample ($n = 6$) is too small for reliable ECE estimation.

Llama-3.1-70B gives 2.21× [1.97, 2.52] and Gemma-2-9B gives 4.18× [2.24, 13.75]. These ratios do not establish that incoherence degrades calibration. The checks below are post hoc and exploratory.

Prompt and parser. The prompt states "This statement is TRUE." before asking for confidence that it is true, and "This statement is FALSE." before asking the reverse, and the parser takes the first number it finds. Of Llama's 569 incoherent pairs, 404 have both confidences at 1.0. About half the statements are true, so those pairs sit near an ECE of 0.5 by construction. Without them the ratio is 1.46× [1.19, 1.89].

Gemma and Mistral. Gemma's 56 coherent pairs are nearly all statements it calls false, with mean stated confidence 0.016 and 8.9% true, so its low ECE reflects selection. Discrimination does not differ between its groups (AUROC .588 and .605). Mistral has 6 coherent pairs.

What remains. Llama's coherent responses separate true from false statements and its incoherent responses do not (AUROC .787 and .493). Each statement falls in one group, so this compares different statements. It is not a ceteris paribus comparison.

Neutral prompt. On the first 400 stored statements we reran Llama-3.1-70B with a prompt that asserts nothing, asks for a probability from 0 to 100, and accepts only a bare number. Extraction rises from 76.2% to 99.8%, and incoherence falls from 24.3% to 5.0%. The ratio falls from 2.76× [2.03, 3.80] to 1.26× [0.61, 2.22], on 20 incoherent pairs, and its interval includes 1. The discrimination gap persists at this small size (AUROC .714 and .438). So both the rate of incoherence and the size of the link depend on how confidence is elicited. On this evidence the link is at most weak, in one model.

Renormalization. Rescaling incoherent pairs to sum to 1 lowers the ECE of Llama's incoherent pairs from 0.502 to 0.076. A constant 0.5 gives 0.047 on the same responses, and AUROC stays near chance (.493 before rescaling, .535 after). Rescaling removes overconfidence and adds no information.

Tolerance and scale. The ratio does not depend on the 5% tolerance: across tolerances from 0% to 20% it stays between 2.21× and 2.31× for Llama and between 3.55× and 4.18× for Gemma. An initial Llama run on $n = 1000$ statements across 10 subjects gave 2.9×; the run reported here covers $n = 3000$ across 20.

D.5.3 Prediction 4: Temperature Sensitivity and Base-Model Diversity

Evidence status: Not registered. The run's script states the prediction as optimal temperature varying with question difficulty, with intermediate temperatures best at moderate diversity and higher temperatures possibly helping at high diversity. All quartiles peak at low temperature, so that prediction did not hold. We restated it as temperature sensitivity after seeing these data, so the restated form is post hoc.

Protocol: Using the $n = 500$ TriviaQA questions from Prediction 1, stratified into quartiles of the Llama base model's response diversity (quartile boundaries computed from the diversity distribution, yielding unequal bin sizes due to ties at low diversity values: Q1=209, Q2=76, Q3=98, Q4=117), we run Llama-3.1-8B-Instruct at 8 temperatures ($\tau \in \{0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 2.0\}$) with 3 samples per temperature. We measure accuracy at each temperature within each quartile.

Results: Table 19 presents accuracy by quartile and temperature. Optimal temperature is similar across quartiles ($\tau = 0.1$–0.5).

	τ=0.1	τ=0.3	τ=0.5	τ=0.7	τ=1.0	τ=1.3	τ=1.5	τ=2.0	n
Q1	90.7	**91.7**	91.5	89.8	81.0	44.7	24.4	4.5	209
Q2	80.7	**81.1**	80.7	78.1	71.9	35.1	13.2	3.1	76
Q3	67.3	66.7	**68.4**	64.6	48.0	25.9	11.2	3.4	98
Q4	**43.9**	**43.9**	41.0	37.3	27.9	10.8	6.0	4.0	117

Table 19: Accuracy (%) by base-model diversity quartile and temperature (Llama-3.1-8B-Instruct, $n = 500$ TriviaQA questions, 3 samples per temperature), scored against the full TriviaQA alias lists. Bold indicates best temperature per quartile. All quartiles peak at low temperatures and differ in sensitivity to increases.

The standard deviation of accuracy across temperatures falls with diversity (.329, .310, .254, .162 for Q1 to Q4; Q1 minus Q4 [0.133, 0.200]), but that measure is tied to peak accuracy: a quartile that starts higher has further to fall. Divided by each quartile's own peak, it is flat (.359, .381, .371, .370; Q1 minus Q4 [−0.035, 0.024]). Retention at $\tau = 1.0$ runs opposite to the restated prediction: 88%, 89%, 70%, 64% of peak accuracy (Q1 minus Q4 [0.143, 0.370]).

The run gave an instruction-tuned model a plain completion prompt with no chat template. We reran 250 of the questions with the template (exploratory). The pattern holds: peaks at $\tau = 0.1$–0.3, standard deviation .239, .244, .210, .121, normalized .259, .286, .283, .280 (Q1 minus Q4 [−0.064, 0.023]), retention 94%, 86%, 84%, 83% (Q1 minus Q4 [−0.029, 0.255]). Accuracy at high temperature is much better with the template: at $\tau = 1.3$ on Q1 it is 73.6% against 46.2% on the same questions, so part of the decline in the table reflects the prompt format. Prediction 4 is not supported, in either form, in this one model.

The evidence on Prediction 2 is inconclusive: a link appears in one model, at a size that depends on the prompt. Predictions 1 and 4 are not supported. The framework does not claim that coherence is necessary or sufficient for good performance. It provides criteria for evaluating methods, and these results bear on how much violating them costs.

D.5.4 Falsifiable Commitments

Prediction 1: As first stated, falsified if accuracy loss does not grow with base-model diversity. The matched reruns falsify it (§D.5.1). The restated form excluded only damage concentrated at high diversity or spread evenly across quartiles. It was written after the data, and we do not count it as a commitment.

Prediction 2: Coherence violations predict miscalibration. Falsified if incoherent responses show *equal or better* calibration than coherent responses across diverse models and domains. Our evidence is inconclusive: Llama's 2.21× falls to 1.46× without its both-1.0 pairs, its interval includes 1 under a neutral prompt (1.26× [0.61, 2.22]), and Gemma's value reflects selection.

Prediction 4: The temperature sensitivity of accuracy varies with base-model diversity. Falsified if questions with low base-model diversity show *equal or less* sensitivity to temperature increases than questions with high diversity. Our evidence does not support it: the absolute measure agrees but tracks peak accuracy, the peak-normalized measure is flat (Q1 minus Q4 [−0.035, 0.024]), and retention runs the other way. The restated form was written after the data, and we do not count it as a commitment.

E Experimental Details

This appendix gives the methodology and detailed results for the experiments of Section 4 and Appendix D. Code and stored runs are at `https://github.com/welr/ip-llm-uq`, commit `38ffd43`.

Infrastructure. Prediction 1 and the initial Prediction 2 runs were conducted on Modal.com serverless GPU infrastructure (NVIDIA A100 80GB for Prediction 1, CPU for Prediction 2). The scaled-up Prediction 2 evaluation (Llama-3.1-70B, $n = 3000$) and Prediction 4 (Llama-3.1-8B-Instruct, temperature sweeps) were run locally on Apple Silicon (M5, 128GB) using 4-bit quantized models via MLX. Prediction 3 generation used Modal.com GPU infrastructure, with NLI post-processing on a local workstation. The matched Prediction 1 rerun, the neutral-prompt Prediction 2 test, and the chat-template Prediction 4 rerun were run locally on the same machine. Frontier-model runs used the vendors' APIs.

E.1 Prediction 1: Instruction Tuning and Credal Width

Models and data. We compare base/instruct pairs from four model families on $n = 500$ TriviaQA questions (Joshi et al., 2017):

Family	Base Model	Instruct Model
Qwen	Qwen/Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct
Mistral	mistralai/Mistral-7B-v0.1	mistralai/Mistral-7B-Instruct-v0.2
Llama	meta-llama/Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Gemma	google/gemma-2-9b	google/gemma-2-9b-it

Table 20: Models for Prediction 1: Qwen2.5 (Qwen Team, 2024), Mistral (Jiang et al., 2023), Llama (Dubey et al., 2024), and Gemma (Gemma Team, 2024) (all via HuggingFace). Mistral's instruct checkpoint (v0.2) was tuned from a later base model than the v0.1 base used here, so that pair also differs in pretraining.

Protocol. For each question: (1) generate 15 base-model responses (3 samples × 5 temperatures); (2) compute response diversity as proportion of unique normalized responses; (3) measure base accuracy (average over 15 samples) and instruct accuracy (one response sampled at $\tau = 1.0$, top-p 0.95); (4) stratify by diversity quartile. All results are exploratory. Section D.5.1 gives the defects in these runs and the corrected comparisons.

	Qwen				Mistral			
Quartile	n	Base	Inst	Δ	n	Base	Inst	Δ
Q1 (low)	198	96.1	94.4	−1.7	207	97.4	84.1	−13.4
Q2	59	71.3	71.2	−0.1	88	82.6	77.3	−5.3
Q3	121	56.1	47.1	−9.0	88	67.4	62.5	−4.9
Q4 (high)	122	17.2	17.2	0.0	117	32.1	33.3	+1.2
Overall	500	64.3	61.4	−2.9	500	74.3	67.2	−7.1
	Llama				Gemma			
Quartile	n	Base	Inst	Δ	n	Base	Inst	Δ
Q1 (low)	209	94.5	85.6	−8.9	217	95.9	95.4	−0.5
Q2	76	81.3	78.9	−2.4	87	77.8	81.6	+3.8
Q3	98	65.8	68.4	+2.6	100	68.1	72.0	+3.9
Q4 (high)	117	30.1	35.0	+5.0	96	32.5	38.5	+6.0
Overall	500	71.8	69.4	−2.4	500	75.0	77.4	+2.4

Table 21: Accuracy (%) by base-model diversity quartile in the stored runs, before the corrections in §D.5.1. Base averages 15 responses over five temperatures; instruct is one response at $\tau = 1.0$.

Matched rerun. Both arms were regenerated locally on an Apple GPU at 16-bit precision (Gemma at bfloat16), with raw text kept: three responses per question at $\tau = 1.0$, top-p 0.95, up to 50 new tokens. Instruct models took the stored direct prompt (`Answer the following trivia question with a brief, direct answer.`) in each model's chat format. Base models took the stored two-shot prompt. Qwen-Instruct was run with and without its chat template, and Llama-Instruct with one and two begin-of-text tokens. Table 22 gives the scorer sensitivity and Table 23 the stratification checks.

E.2 Prediction 2: Coherence and Calibration

Models and data. Three models on MMLU statements (Hendrycks et al., 2021): Llama-3.1-70B-Instruct (Dubey et al., 2024) ($n = 3000$ across 20 subjects), Gemma-2-9B-it (Gemma Team, 2024) and Mistral-7B-Instruct (Jiang et al., 2023) (both $n = 1000$ across 10 subjects).

Protocol. For each statement A, separately elicit $c(A)$ and $c(\neg A)$, the stated confidences that A is true and that it is false, via verbalized confidence prompting. The prompt states the statement's truth or falsity before asking for confidence, and the parser takes the first number in the response. Pairs are excluded when either query lacks a parseable numeric confidence; extraction failure rates vary by model (Llama: 27%, Gemma: 39%, Mistral: 84%), reflecting differences in prompt-following reliability. Valid responses are classified as coherent if $c(A) + c(\neg A) \leq 1.05$ (5% tolerance for rounding), incoherent otherwise. ECE is computed using 10 equal-width bins.

Model	Scorer	Q1 (low)	Q2	Q3	Q4 (high)	Overall
Qwen	Cut, two-way	−7.9	+1.7	+1.4	+6.3	−1.1
Qwen	Cut, word-bounded	−8.1	+2.3	+1.1	+6.8	−1.0
Qwen	Full text, two-way	−6.7	+2.3	+1.7	+6.0	−0.5
Mistral	Cut, two-way	−6.9	−6.1	−0.8	+9.1	−1.9
Mistral	Cut, word-bounded	−7.1	−6.4	+1.1	+8.3	−1.9
Mistral	Full text, two-way	−5.0	−4.5	+3.0	+11.7	+0.4
Llama	Cut, two-way	−1.4	−4.8	+7.5	+7.7	+1.9
Llama	Cut, word-bounded	−1.4	−4.8	+8.2	+7.4	+2.0
Llama	Full text, two-way	−1.3	−4.4	+8.5	+8.5	+2.5
Gemma	Cut, two-way	+0.2	−3.1	+3.7	+4.5	+1.1
Gemma	Cut, word-bounded	+0.5	−4.2	+4.0	+3.1	+0.9
Gemma	Full text, two-way	+0.2	−3.1	+3.7	+4.5	+1.1

Table 22: Matched instruct minus base accuracy (percentage points) under three scorers, with prompts corrected (exploratory). Cut is the repaired first-sentence cut. The two-way rule is the substring rule of the stored runs, the word-bounded rule accepts a gold answer only at word boundaries, and the full-text scorer drops the cut and stops at an invented next question.

Results. Table 18 gives counts and ECE by coherence status. §D.5.2 gives the post hoc checks, including the neutral-prompt rerun, and the limits on the ratios: both-1.0 pairs inflate Llama's, Gemma's reflects selection, and Mistral has 6 coherent responses. Mistral's 96% incoherence comes from the same prompt and parser and was not retested under the neutral prompt.

Robustness to coherence threshold. Table 24 shows that the ECE ratio changes little with the coherence tolerance: by at most 0.10 for Llama, and for Gemma only above 10%.

E.3 Prediction 3: Choice Function Admissibility

Registrations. The registrations were filed when these experiments were planned as a separate choice-function paper, so their titles and text refer to that study.

Models and data. Four base models (Qwen2.5-7B, Mistral-7B-v0.1, Llama-3.1-8B, Gemma-2-9B) on TriviaQA ($n = 500$), MMLU ($n = 500$), ARC-Challenge ($n = 500$), TruthfulQA (full validation split, $n = 817$), and NQ-Open ($n = 500$). For each question and model, we generate 15 responses (3 samples at each of 5 temperatures, in increasing order of temperature) and retain the first 6 in generation order, which are the three samples at each of $\tau = 0.3$ and $\tau = 0.5$. That gives 12,000 responses per benchmark except TruthfulQA (19,608). The TruthfulQA and NQ-Open registrations state this rule. The MMLU/ARC registration states one sample per temperature plus one, so its runs deviate on this point. On TriviaQA, keeping the first k, the last k, or a random k of the six, $k \in \{3, 4, 5, 6\}$, gives ratios from 2.33× to 3.13× (post hoc). The retained six cannot be compared with all 15, because the 15-response similarity matrices were not stored.

Protocol. Similarity is the mean of the two directed entailment probabilities from DeBERTa-large-MNLI. For each $\theta \in \{0.50, 0.55, \ldots, 0.95\}$, we take connected components of the thresholded graph, weight responses equally, and mark the largest clusters, ties included, as the TruthfulQA and NQ-Open registrations state. A response is admissible if marked at some θ. A response counts as correct if a normalized gold answer is a substring of it or vice versa, ignoring gold answers that normalize to the empty string. p-values are two-sided Fisher exact tests on the pooled or

Model	Strata	Q1 (low)	Q2	Q3	Q4 (high)
Qwen	Own quartiles	−8.0 [−12, −4]	+1.7 [−7, +10]	+1.4 [−6, +8]	+6.3 [+2, +11]
	Other models' quartiles	−7.4 [−12, −3]	−3.6 [−10, +2]	−2.8 [−9, +3]	+10.6 [+5, +16]
Mistral	Own quartiles	−7.0 [−11, −4]	−6.1 [−14, +1]	−0.8 [−9, +8]	+8.6 [+1, +16]
	Other models' quartiles	−1.1 [−4, +2]	−5.2 [−11, 0]	−6.9 [−14, 0]	+4.8 [−3, +13]
Llama	Own quartiles	−1.4 [−4, +1]	−4.8 [−11, +1]	+7.5 [0, +15]	+7.7 [+1, +15]
	Other models' quartiles	−0.2 [−2, +2]	+4.2 [−2, +10]	+0.5 [−6, +7]	+4.3 [−2, +11]
Gemma	Own quartiles	+0.2 [−2, +2]	−3.1 [−11, +5]	+3.7 [−3, +11]	+4.5 [−3, +13]
	Other models' quartiles	+0.2 [−1, +2]	0.0 [−5, +5]	+4.2 [−1, +10]	−0.3 [−8, +7]

Base models	Q1 (low)	Q2	Q3	Q4 (high)
Llama minus Mistral	−5.3 [−8, −3]	−3.8 [−10, +3]	−3.0 [−12, +5]	+9.5 [+3, +16]
Gemma minus Mistral	−0.6 [−3, +1]	−2.7 [−9, +4]	+2.7 [−5, +10]	+8.6 [+2, +16]
Mistral minus Llama	−0.6 [−4, +2]	−5.3 [−11, 0]	+8.8 [+1, +17]	+2.3 [−5, +9]
Gemma minus Llama	+0.6 [−1, +3]	−0.4 [−7, +5]	+9.9 [+3, +17]	+3.4 [−3, +10]
Mistral minus Gemma	−3.1 [−6, −1]	−7.7 [−14, −2]	−1.3 [−8, +5]	+6.2 [−1, +14]
Llama minus Gemma	−3.7 [−6, −1]	−11.5 [−18, −6]	0.0 [−7, +7]	+3.5 [−3, +10]

Table 23: Stratification checks, in percentage points with 95% bootstrap intervals (exploratory). Top: instruct minus base with prompts corrected, on the model's own quartiles and on quartiles of the other three models' mean diversity. Bottom: one base model minus another, with no tuning involved, on the second model's quartiles. Questions are the 498 distinct ones among the 500, so own-quartile values can differ slightly from Table 17.

	Llama-3.1-70B			Gemma-2-9B		
Tolerance	n_{coh}	n_{inc}	Ratio	n_{coh}	n_{inc}	Ratio
0%	1523	657	2.22×	56	551	4.18×
2%	1565	615	2.24×	56	551	4.18×
5%	1611	569	2.21×	56	551	4.18×
10%	1636	544	2.23×	58	549	4.18×
15%	1647	533	2.25×	65	542	3.73×
20%	1660	520	2.31×	69	538	3.55×

Table 24: ECE ratio (binned, 10 bins) as a function of coherence tolerance. The ratio ranges from 3.55× to 4.18× for Gemma and 2.21× to 2.31× for Llama, so it does not depend on the 5% tolerance. This table does not address the limits in §D.5.2.

per-model contingency table. Intervals are bootstrap percentile intervals from 10,000 resamples, drawn independently within the admissible and rejected groups.

Scorers and deviations. On MMLU and ARC-Challenge the registered scorer also accepts the gold answer letter by substring, and the prompts included the answer options although the registration specified prompts without them. Both matter. A letter inside any word earns credit under this rule. We compare it with a stricter rule (post hoc): the gold letter counts only as a capital standing alone, as a leading label, or after "answer", "option", or "choice", and the gold text counts only with word-bounded containment. Over all 15 stored samples, accuracy under the registered rule and the stricter rule is 0.923 and 0.815 for Qwen, 0.745 and 0.452 for Mistral, 0.793 and 0.472 for Llama, and 0.853 and 0.695 for Gemma on ARC-Challenge, and 0.762 and 0.542 for Gemma on MMLU. We report the registered results as registered. Rebuilding the retained six from the stored samples reproduces all 12,000 stored labels on ARC-Challenge and all 3,000 for Gemma

on MMLU, and rescoring those six over the stored partitions under the stricter rule gives the post hoc ratios in §4: admissible and rejected accuracy are 0.745 and 0.434 on ARC-Challenge, and 0.647 and 0.520 for Gemma on MMLU. The other three MMLU models cannot be rescored because their samples were not stored. The TruthfulQA judge is Claude Haiku 4.5, given the question and the benchmark's correct and incorrect answer lists; the rescoring is post hoc and the partition is unchanged.

TriviaQA numbers. All TriviaQA figures for Prediction 3 score the stored responses against each question's full alias list. Prediction 1 scored against the full lists during generation but stored only the first three aliases, and the first admissibility analyses reused those stored lists as gold. The canonical answer lies outside the first three on 267 of the 500 questions, and 1,785 correct responses were marked wrong: total correct is 7,238 against three aliases and 9,023 against the full lists. Partitions are unchanged, and the ratio against three aliases is 3.16×. Prediction 1's accuracies, scored during generation, are unaffected. An earlier run, also scored against three aliases, gave 11,503 admissible and 497 rejected responses (2.94×), and the registrations cite that value. Its NLI matrices were not stored. The stored matrices, computed with a different model snapshot, give 11,373 and 627. Total correct against three aliases is 7,238 in both runs, overall and identical per model, so that difference comes entirely from the partition. No registered hypothesis threshold depends on these values. The difference is also an indication of sensitivity to the similarity computation.

Protocol details missing from the registrations. All four registered runs gave base models the same prompt: `Answer the following question concisely.`, then `Q: [question]` and `A:` on separate lines, with the answer options inside the question on MMLU and ARC-Challenge. The documents attached to the registrations say "No system prompt; question presented directly". That is accurate on the system prompt and not on the question. The TriviaQA discovery runs used a different prompt, two worked question-answer pairs followed by the question, although the MMLU/ARC registration calls its protocol identical and the NQ-Open registration calls its design the same protocol. Each response was cut at its first newline and then its first period before similarity and scoring, a step no registration states. The cut truncates at abbreviations and can leave an empty string. Normalized text was capped at 100 characters in the runs before April 17, 2026, including TruthfulQA, and at 500 in the NQ-Open run; a 500 cap would change 7 of 19,608 TruthfulQA labels.

Post hoc, these choices change the pooled ratios as follows. Removing empty responses from the table takes NQ-Open (263 empty) from 1.62× to 1.60×, TruthfulQA (101) from 1.88× to 1.87×, and ARC-Challenge (178) from 1.10× to 1.09×. Removing responses that end in an abbreviation or initial gives 3.06× on TriviaQA, 1.61× on NQ-Open, and 1.86× on TruthfulQA. The similarity of two identical strings can fall below 1: among identical pairs, 36.8% on TriviaQA, 24.6% on NQ-Open, and 24.7% on TruthfulQA have similarity below 0.95. Setting identical pairs to 1 rejects more responses. It gives 3.14× on TriviaQA at 906 rejected (from 627) and 1.73× on NQ-Open at 2,777 (from 2,389), which would clear H6's 1.71×. On the cached TruthfulQA matrices, which reproduce 2,964 of 3,268 partitions, it moves the ratio from 1.72× at 4,332 rejected to 1.54× at 4,678, which would miss H6. These are post hoc, and every registered outcome stands as registered.

Empty outputs and bootstrap unit. The forms differ on empty outputs. The MMLU/ARC form excludes a question-model pair with an empty output and reports the count. The TruthfulQA and NQ-Open forms say in one answer that empty responses are kept and noted, and in another

that an empty generation is re-sampled once and the pair then dropped. The runs kept every pair and never re-sampled. Stored text is cut before storage, so empties from the model cannot be separated from empties the cut produced, and the re-sampling rule cannot be applied now. Dropping every pair with an empty response among its six (post hoc) removes 109 of 2,000 pairs on ARC-Challenge (1.10× to 1.09×), 68 of 3,268 on TruthfulQA (1.88× unchanged), and 160 of 2,000 on NQ-Open (1.62× to 1.64×), and changes no primary outcome on those three. On MMLU only Gemma's samples are stored: 3 of 500 pairs, 1.17× to 1.18×. The TruthfulQA and NQ-Open forms also state the bootstrap unit two ways, responses in two answers and questions in a third; the MMLU/ARC form states responses, and we resample responses. Resampling whole questions across the four models (post hoc) gives H4 lower bounds of 2.23 on TruthfulQA and 2.07 on NQ-Open and H5 lower bounds of 0.43 and 0.40, so those outcomes stand. The MMLU/ARC form sets H1b and H2b on the lower confidence bound and its attached document on the ratio; both fail on either reading.

Tie rule on MMLU and ARC-Challenge. The MMLU/ARC registration breaks ties for the largest cluster by lowest cluster index. The runs keep all tied clusters, which is the rule the TruthfulQA and NQ-Open registrations state. We recompute with the registered rule (post hoc), taking the cluster with the smallest response index, since the registration does not define the index further. On the same cached matrices, ARC-Challenge goes from 1.12× at 5.5% rejection to 1.29× [1.24, 1.34] at 16.0%, with 4 of 4 models above 1 rather than 3. H2a, H2b, and H2c keep their outcomes. The stored 1.10× differs from 1.12× because 23 of 2,000 partitions do not reproduce from the cached matrices. On MMLU only Gemma's matrices are stored; its ratio goes from 1.23× to 1.37× and its rejected responses from 23 to 491. The stored 1.17× differs from 1.23× because 1 of 500 partitions does not reproduce. H1a, H1b, and H1c cannot be recomputed under this rule, and neither can H3, because three MMLU models stored no samples.

Filing and code. Registration records date the MMLU/ARC filing to April 9, 2026, TruthfulQA to April 15, and NQ-Open to April 18; their first runs followed on April 10, April 16, and, for NQ-Open, later the same day. The NQ-Open run used a later revision of its frozen script; the changes are checkpoint frequency, command-line checks, and comments. The MMLU/ARC code entered version control on April 11, after those runs began, so its state at registration cannot be verified from the repository. The NQ-Open registration refers to the TruthfulQA registration by its project page, osf.io/vze74; the registration is osf.io/dqbmg.

Analyses named in the registrations. The registrations name the analyses in Table 27, as analyses to report or as measured variables. We computed them after the fact from stored data. Three results bear on the method. Rejected TruthfulQA responses match a listed incorrect answer less often than admissible ones (14.2% and 28.5%), so the partition does not single out the benchmark's listed falsehoods. The width of the entropy range does not detect the 1–2/6 bin on NQ-Open (AUROC .474). And at comparable rejection, string admissibility gives 1.94× on NQ-Open and 1.93× on TruthfulQA, against 1.62× and 1.88× for NLI. With intervals that resample whole questions, string minus NLI is 0.32 [0.07, 0.61] on NQ-Open and 0.05 [−0.22, 0.35] on TruthfulQA, so only the NQ-Open gap excludes 0.

Dependence. Responses to one question are dependent, so Fisher tests and bootstrap intervals over responses overstate precision. We keep them for the registered analyses and report intervals

over question-model pairs alongside them. Hypothesis outcomes in Table 4 use the registered method.

Results. Table 25 gives per-model results, and Table 26 gives stratified results. For TriviaQA (post hoc), under stratifiers that do not condition on correctness (rank bins of SNNE, of SE at $\theta = 0.75$, and of mean pairwise similarity), the ratio on bins with a non-empty rejected set lies between 1.46× and 2.62×, with no inversion. TruthfulQA stratification was registered, and its row in Table 26 uses the registered partition. Its row in Table 6 does not: the cached similarity matrices behind it reject 4,332 responses rather than the registered 3,847, so that row is indicative only. Using generation log-probabilities in place of equal weights (teacher-forced mean per-token log-probability, softmax over the six) gives 3.16× [2.87, 3.52] at 9.3% rejection on TriviaQA (post hoc). Replacing the mean of the two entailment probabilities by their minimum, maximum, or product gives 2.60×, 4.92×, and 2.53×, also post hoc.

Benchmark	Model	Adm. Acc	Rej. Acc	Ratio [95% CI]	n_{rej}	Fisher p
TriviaQA	Qwen	71.6%	22.9%	3.13× [2.50, 4.10]	210	$< 10^{-10}$
	Mistral	80.4%	26.9%	2.99× [2.38, 3.99]	160	$< 10^{-10}$
	Llama	78.6%	22.9%	3.43× [2.60, 4.90]	131	$< 10^{-10}$
	Gemma	80.8%	34.1%	2.37× [1.91, 3.09]	126	$< 10^{-10}$
MMLU	Qwen	81.9%	84.6%	0.97× [0.89, 1.08]	78	0.654
	Mistral	68.8%	51.3%	1.34× [1.18, 1.55]	195	1.1×10^{-6}
	Llama	81.0%	70.0%	1.16× [1.08, 1.25]	320	8.5×10^{-6}
	Gemma	79.9%	68.0%	1.17× [0.94, 1.65]	25	0.139
ARC-Challenge	Qwen	94.5%	90.9%	1.04× [0.96, 1.16]	44	0.308
	Mistral	77.2%	67.3%	1.15× [1.01, 1.33]	101	0.030
	Llama	84.2%	80.0%	1.05× [1.01, 1.11]	484	0.023
	Gemma	91.0%	100.0%	0.91×	7	1.000
TruthfulQA	Qwen	17.4%	9.9%	1.77× [1.46, 2.20]	902	5.0×10^{-9}
	Mistral	20.1%	8.3%	2.43× [1.95, 3.12]	831	$< 10^{-10}$
	Llama	13.5%	7.4%	1.83× [1.50, 2.32]	1153	5.9×10^{-9}
	Gemma	19.0%	12.2%	1.56× [1.32, 1.89]	961	2.6×10^{-7}
TruthfulQA (judge)	Qwen	34.5%	30.8%	1.12× [1.01, 1.25]	902	0.035
	Mistral	24.8%	21.1%	1.18× [1.03, 1.37]	831	0.021
	Llama	24.8%	24.5%	1.01× [0.90, 1.13]	1153	0.907
	Gemma	25.7%	25.0%	1.03× [0.92, 1.17]	961	0.650
NQ-Open	Qwen	30.6%	18.9%	1.62× [1.38, 1.96]	594	5.8×10^{-9}
	Mistral	32.0%	17.7%	1.81× [1.51, 2.21]	519	$< 10^{-10}$
	Llama	26.0%	21.2%	1.23× [1.06, 1.44]	750	0.008
	Gemma	35.0%	18.3%	1.92× [1.61, 2.34]	526	$< 10^{-10}$

Table 25: Per-model admissibility results. Gemma on ARC-Challenge rejected 7 responses, all correct, so its ratio carries no interval.

Exploratory runs. The frontier runs used $n = 200$ TruthfulQA questions per model, drawn as two disjoint seeded samples of 100, with the second added after the first gave marginal p-values ($p = 0.067$ Sonnet, $p = 0.139$ Gemini). Temperatures were $\{0.3, 0.5, 0.7, 0.9, 1.0\}$, since one API caps at 1. The first sample gave 1.22× for Claude Sonnet 4.6 and 0.82× for Gemini 2.5 Flash; the second gave 1.10× and 0.89×. A registered study at larger n would be needed to settle the direction.

Benchmark	1–2/6	3–4/6	5/6
TriviaQA (post hoc)	0.24× [0.17, 0.31]	11.36× [6.43, 28.23]	6.91× [4.02, 18.33]
TruthfulQA, substring (registered)	0.51× [0.45, 0.59]	2.97× [2.38, 3.89]	2.39× [1.83, 3.35]
TruthfulQA, judge (post hoc)	0.43× [0.39, 0.47]	1.63× [1.46, 1.84]	1.86× [1.64, 2.16]
NQ-Open (registered)	0.48× [0.41, 0.57]	2.66× [2.20, 3.32]	1.90× [1.63, 2.28]
MMLU (registered, with deviations)	0.54× [0.44, 0.69]	1.47× [1.21, 1.88]	1.52× [1.30, 1.84]
ARC-Challenge (registered, with deviations)	0.37× [0.29, 0.49]	0.99× [0.85, 1.18]	1.48× [1.30, 1.72]

Table 26: Admissibility ratio by the number of correct responses among six. Bins 0/6 and 6/6 are trivial. Binning conditions on the outcome, so the peak and the inversion should not be read as effect sizes.

E.4 Prediction 4: Temperature Sensitivity and Base-Model Diversity

Model and data. Llama-3.1-8B-Instruct (4-bit quantized via MLX) on the $n = 500$ TriviaQA questions from Prediction 1, stratified into quartiles of the Llama base model's response diversity.

Protocol. For each question, generate 3 responses at each of 8 temperatures ($\tau \in \{0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 2.0\}$). The prompt was a plain completion with no chat template (`Answer the following question concisely.`, then `Question: [question]` and `Answer:`), and sampling used temperature alone, with no top-p. A response counts as correct if a normalized gold alias is a substring of it or vice versa. Compare accuracy profiles across quartiles.

Scoring. Responses are scored against the full TriviaQA alias lists, ignoring gold answers that normalize to the empty string. The original run scored against the three aliases stored with Prediction 1, which understated accuracy; for example, Q1 at $\tau = 0.5$ scored about 52% instead of 91.5%, while the stored Prediction 1 run gives the same model 85.6% on Q1 with one response at $\tau = 1.0$. The rescoring uses the stored responses, with no new generation.

Results. All quartiles peak at low temperatures ($\tau = 0.1$–0.5). Section D.5 gives the sensitivity measures and the rerun with the chat template.

E.5 Summary

Tables 4 and 5 in §4 collect every hypothesis and its outcome.

Benchmark	Analysis	Result
NQ-Open	AUROC for majority-vote error	KLE .729, averaged KLE .725, averaged SE .720, $1 - \text{maxCP}$ .710, SE(0.75) .708, SNNE .588
	Majority-vote accuracy, all and admissible	30.4%, 30.6%
	SE width as detector of the 1–2/6 bin	AUROC .474
	Distinct largest-cluster sets across the grid	mean 1.42; one set on 65.5%
	String admissibility; frequency, consistency, length filters	1.94× at 19.7%; 2.22× at 51.7%; 2.24× at 51.2%; 0.65× at 22.4%
	SE width, AUROC for majority-vote error	.479
	Incorrect under the registered scorer, rejected and admissible	80.8%, 69.0%
TruthfulQA	AUROC for majority-vote error	KLE .692, SE(0.75) .690, $1 - \text{maxCP}$ .649, SNNE .503
	Majority-vote accuracy, all and admissible	17.7%, 17.6%
	Adversarial; non-adversarial	1.83× [1.53, 2.21]; 1.95× [1.60, 2.43]
	By category	33 of 38 have at least 30 rejected; 31 of those have a defined ratio: 0.00× to 8.13×, 9 below 1
	SE width as detector of the 1–2/6 bin	AUROC .514
	Matches a listed incorrect answer, rejected and admissible	14.2%, 28.5%
	String admissibility; frequency, consistency, length filters	1.93× at 19.2%; 4.22× at 51.8%; 4.24× at 51.4%; 0.26× at 18.7%
	SE width, AUROC for majority-vote error	.385
MMLU	STEM; social sciences; humanities	1.10× [0.99, 1.22]; 1.29× [1.15, 1.47]; 1.16× [0.99, 1.39]
	Jaccard with frequency, consistency, length filters	.216, .217, .033
	Threshold range, 35 endpoint pairs (Gemma)	1.17× to 1.44×, rejection 0.6% to 16.6%
	Majority vote, all and admissible (Gemma)	81.2%, 81.2%
	Best of six, read as at least one correct (four models, registered scorer): all, admissible	92.5%, 90.4%
	Admissibility rate against difficulty	Spearman $\rho = -0.21$ $[-0.29, -0.13]$
ARC-Challenge	Jaccard with frequency, consistency, length filters	.276, .278, .026
	SE width as detector of the 1–2/6 bin	AUROC .252
	Threshold range, 35 endpoint pairs	1.05× to 1.35×, rejection 3.4% to 15.6%
	Majority vote, all and admissible	87.1%, 87.1%
	Best of six, read as at least one correct: all, admissible	96.2%, 94.7%
	Admissibility rate against difficulty	Spearman $\rho = -0.18$ $[-0.26, -0.09]$

Table 27: Secondary and exploratory analyses named in the registrations, computed after the fact from stored data. Intervals resample question-model pairs, except those on Spearman correlations, which resample questions. Best of six is our reading of an analysis the MMLU/ARC form names without defining. TruthfulQA AUROCs use cached matrices that reproduce 2,964 of 3,268 partitions and are indicative, except SE width, which uses stored values. Filter rows give the ratio and rejection rate on stored labels. The MMLU domain assignment is ours. Jaccard values pool over question-model pairs. The threshold-range grid contains the MMLU/ARC form's $[0.70, 0.90]$ and $[0.40, 0.95]$ and extends past them.