CREDAL LARGE LANGUAGE MODELS FOR SEMANTIC COMMITMENT UNDER UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) often produce fluent but incorrect answers with unwarranted confidence. A central limitation is that standard LLMs represent uncertainty through a single predictive distribution, conflating epistemic ignorance with genuine ambiguity. We introduce Credal Large Language Models (CLLMs): an ensemble of LoRA adapters induces a credal set whose lower and upper probabilities expose the spread of plausible predictive distributions rather than collapsing to a single softmax output. From this representation, we derive a single commitment rule: the model commits to an answer only when its lower probability exceeds the upper probability of every alternative, and otherwise returns the set of answers that no plausible predictor rules out. We apply this commitment rule at two depths: Credal Token Commitment (CTC) applies it to answer tokens from one ensemble forward pass, which decides constrained answers without any generation; for open-ended answers, credal decoding extends a partial answer only when no completed answer dominates it, so that the completions produced are those the plausible predictors license, and Credal Semantic Commitment (CSC) applies the rule to their meaning clusters. We evaluate CLLMs with Gemma-2-9B, Llama-3.1-8B and Qwen2.5-7B on OpenBookQA, CoQA, TriviaQA and ARC-Challenge. On multiple choice, CTC commits on 73–91% of questions at 89–98% accuracy, returns sets of 1.1–1.5 options containing the gold one on 89–98%, and its intervals contain the observed accuracy in 24 of 30 confidence bins without calibration; corrupted context lowers commitment from 87–92% to 65–71%, and on Gemma the credal bound detects corruption better than every baseline. On open-ended QA, CLLM outperforms semantic entropy and Laplace-LoRA at a fixed coverage by up to 19% and 9.5% absolute accuracy on CoQA and TriviaQA with context, for every backbone.

1 INTRODUCTION

Large Language Models (LLMs) achieve strong performance on question answering, reasoning, code generation and open-ended dialogue (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Jiang et al., 2023), yet they continue to produce fluent-but-incorrect answers with unwarranted confidence, particularly when context is incomplete, conflicting or adversarial (Maynez et al., 2020; Ji et al., 2023; Huang et al., 2025; Farquhar et al., 2024). In safety-critical settings the model often fails not by being uncertain, but by appearing *insufficiently uncertain*. A deployed system needs to make a decision for each question: whether the model may answer and, if not, which answers remain plausible. A standard LLM cannot supply one, because it represents its uncertainty through a single next-token distribution obtained by softmax normalisation (Fig. 1). One such distribution can be sharp for two very different reasons: because the evidence supports one answer, or because fine-tuning happened to settle on one of several answers that the data does not adjudicate between.

Existing methods build a family of models and then summarise it. Bayesian and Laplace approximations place distributions over adapter parameters (Yang et al., 2024; Wang et al., 2024b), and LoRA (Hu et al., 2021) ensembles, the most widely used remedy at LLM scale, fine-tune several adapters on the same data and average their next-token distributions (Lakshminarayanan et al., 2017; Balabanov & Linander, 2025; Wang et al., 2023a). Averaging is where the information is lost. Five adapters that each prefer a different answer and five that agree on a hedged answer can give the same mean, so the pooled model reports how spread out its distribution is, not whether its members agree (Hüllermeier & Waegeman, 2021), and the disagreement survives only as a scalar, such as mutual information, that must again be thresholded. Generation compounds the loss: an open-ended

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

answer has no fixed label set, the adapters often disagree about how an answer begins (*Special* or *special*) while agreeing on what it says. Sampling from the mean, or from one model as semantic entropy does (Kuhn et al., 2023; Farquhar et al., 2024), yields candidates without any adapter's probability of them. Credal readings of ensembles in classification keep the members as a set of distributions and improve epistemic uncertainty over averaging (Shaker & Hüllermeier, 2021; Wang et al., 2025a; 2024a), as do set-valued classifiers such as random-set neural networks (Kudukkil Manchingal et al., 2025); however, they assume a fixed label set and return another score. In every case the family is collapsed to one distribution or one scalar before a decision is taken, and none of these methods tells a deployed system when it is in a position to answer and when it must present alternatives.

We address this from the perspective of *imprecise probability* (Walley, 1991; Levi, 1980), in which epistemic uncertainty is represented by a *set* of plausible distributions rather than a single point. We propose **Credal Large Language Models** (CLLMs): an ensemble of LoRA adapters on a frozen backbone induces a credal set, a closed convex set of distributions that bounds each candidate answer's probability from below and above, and the model *commits* to a candidate only when no other candidate could be preferred by any distribution in that set. *This is the interval-dominance rule of imprecise-probability decision theory (Troffaes, 2007), and it changes the question from "how uncertain is the prediction?" to* ***"which answers does the evidence still license?"***. The rule contains no threshold, temperature or sharpness parameter, reduces to argmax prediction when the adapters agree, and returns the set of licensed answers; the model commits when exactly one remains.

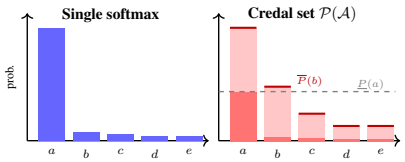


Figure 1: **Left:** one softmax over five candidate answers a-e is sharp whether the model knows the answer or merely settled on one. **Right:** the credal set $\mathcal{P}(\mathcal{A})$ (Eq. (1)) of five LoRA adapters gives each answer a lower (solid) and an upper (band) probability. Here the lower bound of a (dashed) is below the upper bound of b, so some adapter prefers b: the model should not commit to a alone.

Our main **contributions** are: **(i)** ***CLLMs***, which read a fine-tuned LoRA ensemble on a frozen backbone as a credal set, bound each candidate answer's probability by its minimum and maximum over the adapters, and decide by interval dominance, committing when one answer survives and otherwise returning the answers the evidence still licenses (§3.1–§3.2); **(ii)** ***Credal Token Commitment (CTC)***, the interval-dominance rule of (i) applied to the whole vocabulary at the first answer position after one ensemble forward pass, committing when a single token survives, which decides multiple-choice questions with no generation (§3.4); **(iii)** ***credal decoding***, which extends a partial answer only while its largest probability under any adapter is at least the best completed answer's smallest probability over adapters, records every adapter's probability of each completion, and adds to the upper bounds the unexplored mass, the probability each adapter gives to completions never written out, so it needs no temperature or sample count and unwritten completions can only make the model more cautious (§3.3); **(iv)** ***Credal Semantic Commitment (CSC)***, which groups completed answers into *meanings* that every adapter judges equivalent, bounds each meaning by the adapters' summed probabilities, adding the unexplored mass to the upper bound, and applies the same interval dominance rule, so the model commits to what an answer says rather than to how it is spelled (§3.4); and **(v)** an evaluation of selective accuracy, calibration, set-valued prediction, hallucination and corrupted-context detection, and abstention on five datasets, three backbones and six baseline families, showing that the rule commits on the questions the model knows and withdraws when the evidence is corrupted, with no threshold.

Paper outline. §2 reviews related work; §3 defines the credal set and its bounds (§3.1), the commitment rule (§3.2), credal decoding (§3.3) and CTC and CSC (§3.4); §4 gives the setup (§4.1), results (§4.2) and limitations (§4.3); and §5 concludes. The appendix gives proofs (§A), extended related work (§B), the algorithm with an example (§C), implementation, baselines and setup (§D), additional results on selective accuracy and calibration (§E.1), detection (§E.2), abstention (§E.3), set-valued prediction (§E.4), the four-option setting (§E.5), the two depths (§E.6), ablations (§E.7), cost-sensitive commitment (§E.8) and computational cost (§E.9), and broader impact (§F).

2 Related Work

Uncertainty in LLMs is most often quantified by a scalar read off the predictive distribution. The aleatoric/epistemic split is formalised at the level of that distribution (Hüllermeier & Waegeman, 2021), although linguistic variation complicates it for generated text (Baan et al., 2023; Bickford Smith et al., 2025). LLMs are systematically miscalibrated on question answering and few-shot tasks (Jiang

et al., 2021; Zhao et al., 2021); sequence-level scores can be obtained by ensembling and decomposed into token-level contributions (Malinin & Gales, 2021), verbalised confidence is often a better signal than the probabilities of RLHF-tuned models (Tian et al., 2023; Lin et al., 2022; Kadavath et al., 2022), and semantic calibration can be improved after the fact (Lamb et al., 2026; Dixon, 2026). Conformal Language Modeling calibrates sampling and rejection to return answer sets with coverage guarantees (Quach et al., 2024), and later conformal methods drop the need for logits, trade specificity for factuality or learn when to abstain (Su et al., 2024; Mohri & Hashimoto, 2024; Jiang et al., 2025; Wang et al., 2025c; Tayebati et al., 2025). In every case the model answers when a calibrated threshold on a scalar is met, whereas CLLMs answer only when the evidence licenses a single answer, with no fitted cutoff. Approximate posteriors retain more than a single distribution. Full Bayesian inference is intractable at LLM scale, so recent work places the posterior over the parameters of LoRA adapters (Hu et al., 2021): Laplace-LoRA fits a KFAC Gaussian posterior around the fine-tuned weights (Yang et al., 2024), BLoB trains a variational posterior through an ELBO (Wang et al., 2024b), and LoRA ensembles replace the explicit posterior with several independently trained adapters (Balabanov & Linander, 2025; Wang et al., 2023a). Related methods combine expert adapters, learn calibrated mixtures of LoRA experts or perturb low-rank weights during decoding (Li et al., 2025b; Niu et al., 2024; Zhang et al., 2026). Each constructs a family of predictors and then summarises it, as a mean distribution, a mixture or a disagreement score, before any decision is taken; CLLMs keep the family as a set, so whether its members agree becomes the decision itself.

Hallucination detection and abstention are instances of selective prediction with a reject option (Huang et al., 2025; Ji et al., 2023). Because many surface forms express one answer, token-level dispersion is a poor hallucination signal; semantic entropy instead clusters sampled generations by bidirectional entailment and scores the diversity of the clusters (Kuhn et al., 2023; Farquhar et al., 2024). Later work keeps graded similarities or semantic distances instead of hard classes (Nikitin et al., 2024; Qiu & Miikkulainen, 2024; Lin et al., 2024), measures consistency across sampled answers (Wang et al., 2023b; Manakul et al., 2023), approximates it from hidden states or corrects it for incomplete sampling (Kossen et al., 2024; McCabe et al., 2026), or varies the input or the model (Hou et al., 2024; Li et al., 2025a; 2026; Hamidieh et al., 2026). Most of these reason about the frequencies of samples from one model. Credal decoding, instead, writes out the candidates a set of predictors licenses, obtains each predictor's probability of them and accounts for the probability each predictor gives to completions it never wrote out, so the candidates, their probabilities and their grouping into meanings all come from the same predictors, rather than from one model's samples and an external entailment model, as in semantic entropy. Imprecise probability and robust Bayesian analysis represent epistemic uncertainty by a set of plausible distributions rather than a selected mixture (Walley, 1991; Levi, 1980; Berger et al., 1994), with belief functions and credal uncertainty measures as related tools (Cuzzolin, 2014; 2025; Gonzalez-Garcia et al., 2026). Credal-set and random-set neural networks bring these representations to classification (Wang et al., 2025b;a; Kudukkil Manchingal et al., 2025), and recent work carries credal ideas to language, through credal sets of creative continuations (Garces Arias et al., 2025) and attention parameterised by Dirichlet evidence (Ji et al., 2025). These works assume a fixed label set, measure diversity rather than choose an answer. To our knowledge, CLLMs are the first to use interval-dominance decisions to decide whether an LLM should answer, in both next-token prediction and open-ended generation; an extended discussion is in §B.

3 METHODOLOGY

Given a prompt and M LoRA adapters on a frozen backbone, CLLMs bound the probability (upper and lower) of every candidate answer (§3.1), decide from these bounds using interval dominance whether to commit to one answer or to return a set (§3.2), generate open-ended candidates using the same interval-dominance rule applied to partial answers (§3.3), and apply it at two depths, to answer tokens and to meanings (§3.4). Fig. 2 illustrates the full methodology.

3.1 CREDAL SET FORMULATION

Let x be a prompt, $y_{<t}$ a partial answer, and $\{f_1, \ldots, f_M\}$ LoRA adapters (Hu et al., 2021) sharing a frozen backbone. Adapter m gives a next-token distribution $p_m(\cdot \mid x, y_{<t})$ over the vocabulary $\mathcal{V}$. The adapters are deliberately different fits of the same training data (different target modules, ranks, learning rates and bootstrap resamples; §D), so the family represents a sensitivity analysis over fine-tuning decisions the data does not adjudicate between (Breiman, 2001; Marx et al., 2020). A *candidate* is an answer the model could give: at depth one a token of $\mathcal{V}$ at the first answer position, and in the semantic space a completed answer written out by credal decoding (§3.3); $\mathcal{A}$ denotes a

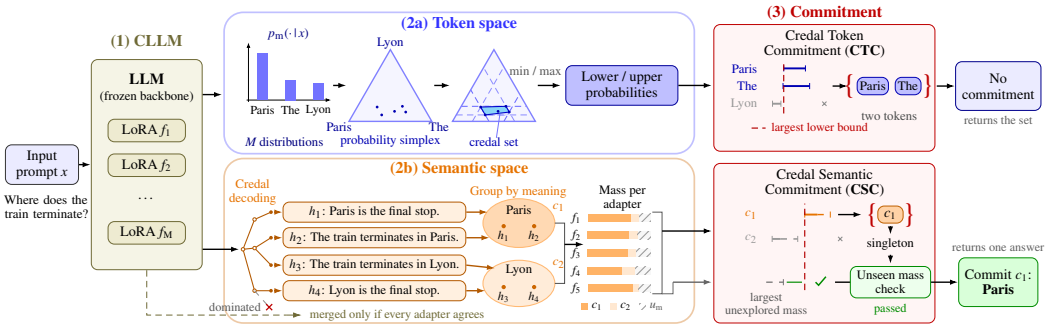


Figure 2: **Credal Large Language Models. (1)** M LoRA adapters on a frozen backbone give M next-token distributions. **(2a)** Plotted in the probability simplex, reduced for display to the three tokens that carry mass here, *Paris*, *The* and *Lyon*, they generate the credal set of Eq. (1), their convex hull (shaded). The dashed region is the larger set of distributions consistent with the bounds alone; the two agree on the lower and upper probabilities of Eq. (2), which is all the rule reads. One interval-dominance test, Eq. (3), is then applied in both spaces. In the **token space** *Paris* and *The* both reach the largest lower bound, so CTC does not commit and returns the surviving set $\mathcal{D}_1$. **(2b)** Credal decoding writes out the completions no adapter dominates (Algorithm 1), the adapters themselves group them by meaning, and each meaning's upper bound is widened by that adapter's unexplored mass u_m (Eq. (4)) (dashed extension). **(3)** In the **semantic space** only c_1 survives and its lower bound clears the largest unexplored mass, so CSC commits to the class $c_1 = \{h_1, h_2\}$, that is to the answer *Paris* rather than to either sentence: the bounds are computed from the pooled mass of both completions. The two spaces are read independently on the same prompt. Here the token space does not close, because the adapters disagree about how the answer begins rather than about what it is, and the semantic space decides.

finite set of candidates of one of these two kinds, with elements $a \in \mathcal{A}$. On any finite candidate set $\mathcal{A}$ to which each adapter assigns probabilities, the M distributions generate a *credal set*, their convex hull

$$\mathcal{P}(\mathcal{A}) = \operatorname{co}\{p_1, \ldots, p_M\} = \Big\{ \textstyle\sum_m \lambda_m p_m \;:\; \lambda_m \geq 0,\ \sum_m \lambda_m = 1 \Big\}, \tag{1}$$

a closed convex set of distributions whose extreme points are the adapters themselves; every weighting of the adapters the evidence might have justified is a point of it. Writing $p_m(a)$ for adapter m's probability of a and suppressing the conditioning on x, the *lower* and *upper probability* of a candidate a are the envelopes of $\mathcal{P}(\mathcal{A})$ at a, the smallest and largest probability $q(a)$ that any distribution $q \in \mathcal{P}(\mathcal{A})$ gives it,

$$\underline{P}(a) = \min_{q \in \mathcal{P}(\mathcal{A})} q(a) = \min_m p_m(a), \qquad \overline{P}(a) = \max_{q \in \mathcal{P}(\mathcal{A})} q(a) = \max_m p_m(a), \tag{2}$$

since $q(a) = \sum_m \lambda_m p_m(a)$ is linear in λ and attains its extremes at vertices (Proposition A.1, §A); Eq. (2) therefore gives the credal set's exact bounds. Mean-pooling the adapters would replace $\mathcal{P}(\mathcal{A})$ by its barycentre and discard the gap; keeping the set is the robust-Bayesian position that declines to choose weights the evidence does not supply (Berger et al., 1994; Walley, 1991).

3.2 Interval-Dominance Commitment

Candidate b *dominates* candidate a when $\underline{P}(b) > \overline{P}(a)$, in which case every distribution in the credal set ranks b above a, since $q(b) \geq \underline{P}(b) > \overline{P}(a) \geq q(a)$ for every $q \in \mathcal{P}(\mathcal{A})$. The *credal commitment set* is the set of undominated candidates,

$$\mathcal{D}(\mathcal{A}) = \big\{ a \in \mathcal{A} \;:\; \overline{P}(a) \;\geq\; \max_{b \in \mathcal{A}} \underline{P}(b) \big\}. \tag{3}$$

The model **commits** when $|\mathcal{D}(\mathcal{A})| = 1$, to the unique element a^*, with confidence $\underline{P}(a^*)$, the support it retains under every distribution in the credal set. Otherwise it **presents** $\mathcal{D}(\mathcal{A})$, the candidates the credal set does not rule out.

Three properties follow directly. *(i) No hyperparameters.* Eq. (3) contains no threshold, temperature, sharpness or support size; whether the model commits is decided per input by the width of its own intervals. *(ii) Singleton limit.* If the adapters agree, $\underline{P} = \overline{P}$ and $\mathcal{D}(\mathcal{A})$ is the set of maximisers of the common distribution (Proposition A.3, §A), so credal commitment departs from argmax prediction only where the credal set is non-degenerate. *(iii) Monotonicity in M.* The bounds are order statistics (the smallest and the largest of the M adapter probabilities), so adding predictors can only widen intervals and enlarge $\mathcal{D}$ (Proposition A.4); set sizes are comparable at fixed M only. Interval dominance is the most conservative of the standard imprecise-probability decision rules (Walley, 1991; Troffaes, 2007) and, unlike maximality and E-admissibility, needs only the bounds on single

candidates; maximality keeps every answer that no other answer beats under all distributions in the set, and E-admissibility every answer that is best under some distribution in the set, so both need the full set.

3.3 CREDAL DECODING

Credal decoding keeps extending a partial answer, that is *follows* it, as long as no completed answer is more probable under every adapter than everything that continues it, that is as long as the ensemble cannot *rule it out* (Algorithm 1). It does so up to a width B (B is the number of partial sequences kept at each step, as in a beam), by applying the rule of Eq. (3) to sequences rather than to a single next token. A *partial sequence* $y_{<t}$ carries, for every adapter m, the log-probability $\ell_m(y_{<t}) = \sum_{s<t} \log p_m(y_s \mid x, y_{<s})$ of its tokens; one that reaches the end-of-answer token becomes a *completed sequence*.

1. Start with the empty partial sequence, $\ell_m = 0$ for every m.
2. *Expansion.* For each partial sequence $y_{<t}$, run the M adapters one step. A continuation y yields $y_{<t} \circ y$ (the partial sequence extended by y) with $\ell_m \leftarrow \ell_m + \log p_m(y \mid x, y_{<t})$ for every m, and is followed unless its upper mass $\overline{q}(y_{<t} \circ y) = \max_m \exp \ell_m(y_{<t} \circ y)$ is below the largest lower mass $\underline{q}^* = \max_h \min_m p_m(h \mid x)$ of any completed sequence h so far, where $p_m(h \mid x) = \exp \ell_m(h)$.
3. *Width.* If more than B undominated partial sequences result, keep the B with the largest upper mass and record a *truncation*.
4. Stop at maximum length L. The *completion set* $\mathcal{H}(x)$ holds the sequences that reached the end-of-answer token and any partial sequence still live at step L, each with its M log-probabilities $\ell_m(h)$.

A continuation pruned in step 2 is *dominated as a sequence*: its upper mass is the most any adapter can give to everything that continues it, so every completion extending it has upper probability below $\underline{q}^*$ and is dominated by the completion that attains $\underline{q}^*$, while the mass it carries is retained in u_m below. Before any sequence has completed, $\underline{q}^* = 0$ and the width alone limits exploration; once a strong completion exists, weaker variants of it are pruned. Adapter m's probability of a completed sequence, $p_m(h \mid x) = \exp \ell_m(h)$, is produced during decoding, so no second scoring pass is needed. The candidates do not exhaust the adapter's distribution: the mass it places on sequences never written out, because a token of theirs was dominated or the width was reached, is the *unexplored mass*

$$u_m(x) = 1 - \sum_{h \in \mathcal{H}(x)} p_m(h \mid x), \tag{4}$$

which §3.4 carries as imprecision rather than discarding. The width B and the maximum length L are the only operational quantities (B is ablated in §E.7.2, and few completions reach L; §D.4). Unlike a sampling temperature, neither decides which completions exist: B only limits how many of the completions the rule licenses are followed, and whatever it leaves out remains visible through u_m.

3.4 CREDAL TOKEN AND SEMANTIC COMMITMENT

Credal Token Commitment (CTC). The rule at depth one: $\mathcal{A} = \mathcal{V}$, the whole vocabulary at the first answer position. Write $\mathcal{D}_1$ for the commitment set of Eq. (3) on that space. The model is *token-committed* when $|\mathcal{D}_1| = 1$. CTC costs one forward pass per adapter and no generation.

Candidates and meanings. In the semantic space the candidates are completed answers: completed sequences that are identical after lower-casing and removing punctuation and articles form one *candidate* a, with $p_m(a \mid x) = \sum_{h \in a} p_m(h \mid x)$. Meaning is decided by the ensemble itself: for two candidates a, b and the question x, each adapter is asked whether a and b give the same answer to the question, and its probability of "yes" against "no" at the first response token is $e_m(a, b)$; adapter m judges the two equivalent when $e_m(a, b) \geq \frac{1}{2}$. Using a language model to judge whether two answers are equivalent follows the entailment check of semantic entropy (Farquhar et al., 2024); the threshold $\frac{1}{2}$ is simply the argmax of "yes" against "no", the same decision applied to entailment classes in Kuhn et al. (2023). Our choice is to require every adapter to agree, which makes the grouping conservative; its quality is measured in §E.7.3. Two candidates are *surely* one meaning when every adapter judges them equivalent and *possibly* one meaning when at least one does. The candidates are partitioned into meanings $\mathcal{C}(x) = \{c_1, \ldots, c_n\}$ greedily, in decreasing order of mass: each candidate joins the meaning whose representative it is surely equivalent to, or starts a new meaning if there is none (§D). The transitive closure of the possibly-equivalent relation gives a coarser partition, reported as a diagnostic of how imprecise meaning itself is for the ensemble (Tab. 18).

Credal Semantic Commitment (CSC). Adapter m's mass on a meaning is $P_m(c \mid x) = \sum_{a \in c} p_m(a \mid x)$. We do not know what the unwritten completions say, so each adapter's unexplored mass $u_m(x)$ is added to the upper bound of every meaning, as if all of it belonged to that meaning: the bounds on a meaning are

$$\underline{P}(c) = \min_m P_m(c \mid x), \qquad \overline{P}(c) = \max_m \big[P_m(c \mid x) + u_m(x)\big], \tag{5}$$

and the interval dominance rule of Eq. (3) is applied to $\mathcal{C}(x)$ with these bounds. The model is *semantically committed* when $|\mathcal{D}(\mathcal{C}(x))| = 1$, to the meaning c^*, *and* $\underline{P}(c^*) \geq \max_m u_m(x)$, so that the unexplored mass, the probability of completions never written out, cannot exceed the lower bound of the committed meaning; its confidence is $\underline{P}(c^*)$. Otherwise it presents the meanings the credal set does not rule out, together with an unseen element when the second condition fails (Algorithm 1, §D). CSC costs the decoding of Sec. 3.3 plus one short forward pass per adapter per judged candidate pair.

Why the two depths are not chained. CTC and CSC are the same rule on two candidate sets. Expanding only the tokens that survive at depth one would be unsound, because a token dominated at depth one can still begin the completion with the largest lower mass; expansion therefore prunes on sequence upper mass rather than on the commitment set $\mathcal{D}_t$ over the next token after $y_{<t}$ (Sec. 4.2(vi)).

Cost-sensitive commitment. Where a deployment must trade coverage against error, the same bounds give a rule with a declared cost rather than a tuned threshold: with utility 1 for a correct answer, $-\kappa$ for a wrong one and 0 for abstaining, the model commits to a^* iff $\underline{P}(a^*) \geq \kappa/(1+\kappa)$. This is the lower-expected-utility, or Γ-maximin, decision (Troffaes, 2007), derived in §E.8.

4 Experiments

4.1 Experimental Setup

Backbones and adapters. Gemma-2-9B-Instruct, Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, each with $M{=}5$ LoRA adapters ($M{=}5$ is compute-bounded; the adapter sweep in §E.7.1 saturates at four) trained on the training split of each dataset (ARC-Challenge uses the OpenBookQA-trained adapters and AdvBench the CoQA-trained ones); the adapters differ in target modules, rank (4, 8, 16, 4, 8), learning rate (1-5 $\times$ 10^{-4}), bootstrap resample and seed (§D). **Datasets. OpenBookQA** (Mihaylov et al., 2018) (4-way multiple choice), **ARC-Challenge** (Clark et al., 2018) (4-way reasoning, a transfer test), **CoQA** (Reddy et al., 2019) (free-form conversational QA), **TriviaQA** (Joshi et al., 2017) (long-tail entities, with the passage and closed-book) and **AdvBench** (Zou et al., 2023) (harmful prompts, against benign Dolly instructions). Prompt counts, the corruption protocol, the development splits on which the width B and length L were fixed, and the correctness rules are in §D.4.

Baselines. Six families, under the same names throughout: (i) *Standard LLM*, the base model without adapters, scored by its predictive entropy or max probability at the first answer token, with its greedy answer; (ii) *LoRA Ensemble* (Balabanov & Linander, 2025; Lakshminarayanan et al., 2017), the same five adapters mean-pooled, scored by predictive entropy, mutual information or variance, which is the control that isolates the credal reading from the ensemble it is read from; (iii) *Bayesian-LoRA (KFAC)* and (iv) *Laplace-LoRA (diagonal)*, the Laplace-LoRA posterior of Yang et al. (2024) over the LoRA weights of one adapter with a Kronecker-factored or a diagonal Fisher (Daxberger et al., 2021), with $S{=}20$ posterior samples; (v) *Semantic entropy* (Kuhn et al., 2023; Farquhar et al., 2024) with its original NLI clustering and with a cosine-similarity variant at the threshold that favours it, from $K{=}16$ samples of the base model; and (vi) the ablated credal scores of Tab. 8. Each baseline uses its standard budget ($K{=}16$, $S{=}20$); and the compute of every method is in §E.9.

Metrics. *Selective accuracy* is the accuracy at 80% coverage with questions ranked by a method's score; ensemble methods rank the same answer, the CSC representative, that is the candidate with the largest lower probability, so only the score differs, and single-model methods rank the base model's greedy answer. *Calibration* is the expected calibration error (ECE) (Naeini et al., 2015; Guo et al., 2017) and negative log-likelihood (NLL) of each method's own per-answer confidence, for CLLM the midpoint of its interval. *Gold-in-set* is the percentage of questions whose set contains the gold answer; as our rule chooses the set size per question, rankings are compared at that size and at a fixed number k of top-ranked answers. *Hallucination* and *corrupted-context detection* are AUROCs of a score against the model's own errors (Kuhn et al., 2023; Farquhar et al., 2024) and against corrupted inputs, our proxy for the conditions under which a model hallucinates. For the rule itself we report commit rate, accuracy when committed, accuracy of the representative, set size and unexplored mass.

4.2 RESULTS

Tabs. 1 and 2 report Gemma-2-9B on all five datasets, with Llama-3.1-8B and Qwen2.5-7B in Tabs. 5 and 6 and Fig. 5. Further results are in §E, and §C.2 follows one prompt through both depths.

Table 1: **Hallucination and corrupted-context detection** (AUROC, Gemma-2-9B; corrupted columns for Llama-3.1-8B and Qwen2.5-7B in Fig. 5; 0.5 is chance and no score is inverted). *Clean*: does the score separate the questions the model's own answer got wrong from the ones it got right (Kuhn et al., 2023; Farquhar et al., 2024)? *Corrupted*: does it separate corrupted inputs from intact ones? Where the answer rests on supplied evidence, the evidence is replaced (the passage in CoQA and TriviaQA, the supporting fact in OpenBookQA); otherwise the question is replaced and the answer side kept (closed-book TriviaQA and ARC-Challenge).

		CoQA		**TriviaQA (passage)**		**TriviaQA (closed book)**		**OpenBookQA**		**ARC-Challenge**	
Score	Source	Clean ↑	Corrupted ↑	Clean ↑	Corrupted ↑	Clean ↑	Corrupted ↑	Clean ↑	Corrupted ↑	Clean ↑	Corrupted ↑
Standard LLM (predictive entropy)	single model	0.786	0.739	**0.850**	0.568	0.729	0.495	0.866	0.538	0.762	0.472
Standard LLM (max prob)	single model	0.774	0.754	0.829	0.594	0.702	0.502	0.863	0.537	0.760	0.488
Semantic entropy (cosine)	single model, K=16	0.646	0.739	0.736	0.697	0.722	0.543	0.668	0.638	0.633	0.490
Semantic entropy (NLI)	single model, K=16	0.709	0.698	0.811	0.675	0.723	0.556	0.535	0.629	0.530	0.502
Bayesian-LoRA (KFAC)	posterior, S=20	0.812	0.827	0.845	0.694	0.792	0.597	0.931	0.642	0.819	0.512
Laplace-LoRA (diagonal)	posterior, S=20	0.822	0.828	**0.850**	0.688	0.789	0.601	0.924	0.634	0.805	0.528
LoRA Ensemble (predictive entropy)	5 adapters, mean-pooled	0.806	0.830	0.847	0.683	0.777	0.693	**0.944**	0.766	**0.834**	0.742
LoRA Ensemble (mutual information)	5 adapters, mean-pooled	0.723	0.854	0.787	0.723	0.706	0.699	0.941	0.767	0.815	0.702
CLLM	ensemble adapters, credal set	**0.844**	**0.873**	**0.850**	**0.882**	**0.833**	**0.701**	0.942	**0.781**	0.830	**0.783**

Table 2: **Selective accuracy and calibration** (Gemma-2-9B; Llama-3.1-8B and Qwen2.5-7B in Tabs. 5 and 6). Acc.@80%: accuracy at 80% coverage when questions are ranked by the score. ECE and NLL: calibration of each method's own per-answer confidence; for CLLM this is the midpoint of the interval, since the lower bound is a guaranteed floor and under-confident by construction. Methods built on the same adapters select the same answer and differ only in the score. ARC-Challenge uses the OpenBookQA-trained adapters.

	CoQA			TriviaQA (passage)			TriviaQA (closed book)			OpenBookQA			ARC-Challenge		
Score	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓
Standard LLM (predictive entropy)	0.795	0.156	0.694	0.710	0.241	0.784	0.735	0.160	0.768	0.973	0.067	0.302	0.943	0.095	0.438
Standard LLM (max prob)	0.790	0.156	0.694	0.700	0.241	0.784	0.715	0.160	0.768	0.973	0.067	0.302	0.940	0.095	0.438
Semantic entropy (cosine)	0.780	0.198	3.141	0.720	0.287	3.610	0.755	0.180	2.480	0.948	0.070	1.237	0.915	0.107	2.043
Semantic entropy (NLI)	0.775	0.129	1.775	0.725	0.240	2.353	0.750	0.144	2.125	0.917	0.138	0.835	0.875	0.130	1.412
Bayesian-LoRA (KFAC)	0.835	0.271	0.645	0.900	0.435	0.854	0.755	0.110	0.537	0.993	**0.009**	0.130	0.940	0.068	0.322
Laplace-LoRA (diagonal)	0.815	0.281	0.663	0.900	0.470	0.934	0.750	0.126	0.544	0.990	0.041	0.146	0.945	**0.043**	**0.290**
LoRA Ensemble (predictive entropy)	0.845	0.080	0.591	**0.920**	0.045	0.704	0.835	**0.032**	0.416	0.995	0.017	0.137	0.950	0.068	0.352
LoRA Ensemble (mutual information)	0.825	0.080	0.591	0.880	0.045	0.704	0.780	**0.032**	0.416	0.993	0.017	0.137	0.950	0.068	0.352
CLLM	**0.850**	**0.045**	**0.579**	0.915	**0.042**	**0.614**	**0.845**	0.038	**0.403**	**0.998**	0.014	**0.123**	**0.953**	0.045	0.291

(i) Detection: the credal bound matches the best detectors of the model's own errors and leads on corrupted inputs. Read as a score, $1 - \underline{P}(c^*)$ is the best detector of the model's own errors on Gemma's three open-ended datasets, tied on TriviaQA with the passage, and within a few thousandths of the mean-pooled ensemble on the two four-option datasets (Tab. 1). The single-model baselines are weaker and less stable: semantic entropy with NLI clustering is close to chance on the four-option datasets. On corrupted inputs the credal bound is the strongest score on every Gemma dataset, by more than 0.15 on TriviaQA with the passage, and averaged over the two open-ended datasets it is also the strongest on Llama (Fig. 5). Two cells go the other way: on Qwen the mean-pooled mutual information of the same adapters is ahead, and on Llama TriviaQA semantic entropy with NLI clustering edges the credal bound, while the other single-model scores on Qwen sit barely above chance (Fig. 5). The ensemble carries the signal about the model's own errors and the bounds keep it; where corruption makes the adapters disagree, the bounds are the sharper reading.

(ii) Selective and set-valued prediction: CLLM beats every single-model baseline and matches or exceeds the ensemble it is read from. At 80% coverage CLLM is 5 to 7 points more accurate than the best single-model baseline on CoQA for every backbone (Tabs. 2, 5 and 6). The gap is expected: one softmax cannot tell a distribution that is sharp because the evidence is strong from one that is sharp because fine-tuning settled, whereas independently fine-tuned adapters disagree exactly where the evidence is thin. More notably, CLLM is also more accurate than the same adapters mean-pooled on every dataset except TriviaQA with the passage for Gemma and Llama, with the largest gains, about 8 points, on Llama ARC-Challenge and Qwen CoQA, and its interval midpoint is better calibrated than the mean-pooled confidence on most datasets. The ensemble keeps the edge on TriviaQA with the passage, and for Qwen on closed-book TriviaQA and ARC-Challenge. The credal sets contain the gold meaning on 76-91% of clean questions (Tab. 12); given our per-question size the ensemble's top-k coincides with them, and against any k fixed in advance the credal set is ahead or level (Tab. 9), so the set content is the ensemble's and the size is the rule's contribution. In Fig. 3 (left) the rule's own operating point, the coverage at which it commits and its accuracy there, chosen with no threshold, lies on the ensemble's risk-coverage curve, and the single-model curves lie below it (Fig. 6). CLLM gives up no ranking quality for its decision, and on most datasets gains some.

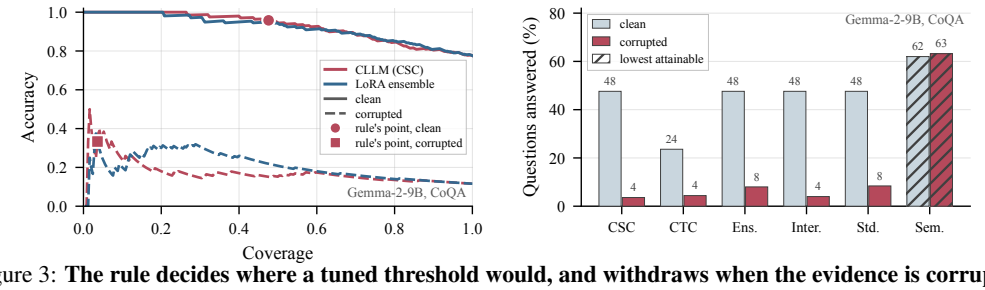


Figure 3: **The rule decides where a tuned threshold would, and withdraws when the evidence is corrupted** (Gemma-2-9B on CoQA; Llama-3.1-8B and Qwen2.5-7B, and TriviaQA, in Fig. 6 and Fig. 7). **Left**, risk-coverage: solid is clean prompts, dashed the same prompts with the passage replaced; the circle and square mark the rule's own operating point, chosen without a threshold. Corrupting the passage moves it from 47.6% coverage to 3.6%, while a ranker's curve only shifts down. **Right**, each method thresholded on clean prompts to the coverage the rule chose: bars are what it still answers. Semantic entropy is zero on most prompts, so no threshold reaches that coverage and its lowest attainable one is shown (hatched).

(iii) Abstention under corrupted context without a threshold. Corrupting the passage lowers CSC commitment on CoQA from 31-48% to 4-8% of prompts and to at most 1% on TriviaQA with the passage (Tab. 12), and token-space commitment falls with it (Tab. 13). The mechanism is the unexplored mass: with the wrong passage the adapters spread their mass over completions credal decoding does not write out, the upper bounds widen and dominance fails. Thresholded on clean prompts to the coverage CSC chose, the LoRA Ensemble's entropy still answers up to 8% of corrupted prompts, more than twice as many as CSC on Gemma, the Standard LLM's entropy still answers up to 16% of corrupted prompts, and semantic entropy, which is zero on most prompts, cannot be thresholded to the target and keeps answering most of them (Fig. 3, right, and Fig. 7). This comparison favours the baselines, whose thresholds are set to the coverage the rule chose for itself: even with that information, the LoRA Ensemble answers more corrupted prompts than CSC on Llama and Gemma. The rule as stated also agrees with the cost-sensitive rule at $\kappa{=}1$ on almost every prompt (Tab. 19), so it behaves as a declared-cost decision in which a wrong answer costs as much as a right one is worth. The rule withdraws where it should, with no threshold to tune or to transfer.

(iv) The four-option setting: CTC is at least as accurate as a tuned ranker at its own coverage, and the intervals are calibrated as intervals. On OpenBookQA and ARC-Challenge CTC decides from one ensemble forward pass. It is never told that an answer is one of four letters, yet almost every token it keeps begins an option; it commits on most questions with high accuracy, returns small sets that usually contain the gold letter, and at its own coverage is as accurate as the LoRA Ensemble's entropy ranking cut at the same coverage (Fig. 4, Tab. 10). At 80% coverage it is the most accurate ranker on both datasets for Llama and Gemma and on OpenBookQA for Qwen (Tab. 7). Corrupting the passage lowers commitment by about twenty points while the committed answers stay accurate, and token-space lower bound is the best detector of the corrupted passage on Gemma (Tab. 1). The intervals contain the observed accuracy in 24 of 30 bins with no calibration step (Fig. 8, Tab. 11); the misses fall mostly in the most confident bins, where 5 adapters slightly overstate their agreement. On a closed answer space, one forward pass of ensemble is enough for a calibrated decision that limits itself.

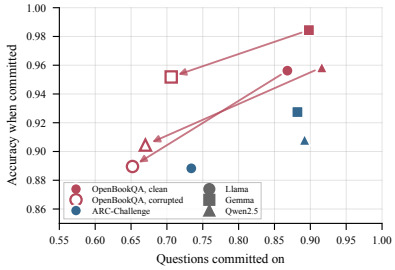


Figure 4: **CTC withdraws under corrupted evidence rather than degrading**. Each marker is one backbone at its commit rate and its accuracy when committed. Filled: OpenBookQA, clean; hollow: the same questions with the passage corrupted; blue: ARC-Challenge, answered with the OpenBookQA-trained adapters.

(v) Two depths, and where each pays off. On clean CoQA and closed-book TriviaQA, CSC commits on a quarter to a half of questions and is right on more than nine in ten of them, well above the accuracy of its best answer on all questions (Tab. 12). The gap is largest closed-book, where no passage supplies the answer: 19 to 39 points, so commitment selects the questions the model actually knows. On TriviaQA with the passage the token space almost never closes, because one adapter reproduces the lower-case gold format while the others capitalise, yet CSC still commits with high accuracy. Each depth commits on questions the other does not (Tab. 14, Fig. 9); as a detector the semantic space is ahead of the token space on every free-form dataset (Tab. 15). The token space

Table 3: **Harmful prompts** (250 AdvBench harmful against 250 benign Dolly prompts, CoQA-trained ensembles). Committed answers on harmful prompts are refusals. The last two columns are means on each half: the ensemble is *more* confident and the single model *less* uncertain on harmful prompts, so neither can flag harm.

Backbone	Committed on harmful	Committed on benign	$\underline{P}(a^*)$ harmful / benign	Standard LLM entropy
Llama-3.1-8B	71.6%	30.4%	0.49 / 0.28	0.03 / 0.91
Gemma-2-9B	97.2%	38.4%	0.93 / 0.32	0.02 / 0.69
Qwen2.5-7B	33.2%	22.8%	0.25 / 0.19	0.02 / 0.33

decides how an answer begins and the semantic space what it says, so the two are complementary rather than redundant.

(vi) The sets are robust to the choices inside the method. On the CoQA development split (§E.7), pruning on sequence-level dominance instead of following $\mathcal{D}_t$ commits more often and keeps the gold answer more often on every backbone (Tab. 17), because token-level dominance spends the width on surface variants that differ at the first token. Adding adapters widens the sets and lowers commitment while gold-in-set and accuracy when committed rise, saturating at four adapters (Tab. 16, Fig. 10); token-space commitment falls monotonically in M, as Proposition A.4 predicts. The adapters' own equivalence judgement groups candidates between the loose and tight settings of the embedding clustering used by semantic entropy, with no threshold (§E.7.3). Carrying the unexplored mass matters: dropping it raises commitment on clean CoQA but lowers accuracy when committed (§E.7.2). These choices change how often the model commits far more than how accurate its commitments are.

(vii) Harmful prompts: commitment is a decision signal, not a danger signal. On AdvBench harmful prompts against Dolly benign instructions (Tab. 3), the CoQA-trained ensembles commit to a refusal on most harmful prompts for Llama and Gemma and a third of them for Qwen, more often than on benign instructions for every backbone, and the committed answer's lower bound is higher on harmful prompts. The Standard LLM behaves the same way: its entropy is lower on harmful prompts, so it separates the classes only in the inverted direction. Safety training makes refusal the unanimous continuation, so every adapter puts its mass on the same first token and the rule commits to it. A confident refusal is thus a correct use of the rule, not a failure, but neither commitment nor any uncertainty score, credal or otherwise, can serve as a harm detector: commitment measures the model's agreement with itself, not whether a request should be answered.

4.3 Limitations

The credal set is only as wide as the ensemble inducing it: five adapters undersample the plausible fine-tunings, so the intervals err on the narrow side, as the most confident bins of Fig. 8 show; they only widen with M (Proposition A.4), and their calibration is measured, not guaranteed. As the adapters share one backbone, errors common to the pretrained model can survive even when every adapter agrees. The semantic space is likewise only as complete as credal decoding and the semantic grouping behind it: decoding follows at most B partial answers and carries the unexplored mass as imprecision, which keeps the rule sound but cautious when adapters format answers differently, as on TriviaQA with the passage (Tab. 12), and meanings rest on the adapters' equivalence judgements, so answers of different granularity can merge (§E.7.3). Inference costs M times a single model, as for the mean-pooled ensemble (§E.9); we evaluate short answers, with corrupted context as a controlled proxy for the conditions under which a model hallucinates.

5 Conclusion

We introduced *Credal Large Language Models*, which read a LoRA ensemble as a credal set and decide by interval dominance: they commit only when no distribution in the set prefers another answer, and otherwise present the answers the evidence still licenses. Credal decoding generates open-ended candidates by the same rule, with no temperature or sample count, and the rule is applied at two depths, to answer tokens (CTC) and to meanings (CSC). Across three backbones and five datasets, CLLM is more accurate at a fixed coverage than every single-model baseline on the open-ended datasets and matches or exceeds the same adapters mean-pooled while removing the threshold they need; under corrupted context its commitment on CoQA falls from 31-48% to 4-8% with nothing tuned, and on multiple choice a single ensemble forward pass gives intervals calibrated without a calibration step and a decision that limits itself. Future directions include propagating the credal set through long-form generation, conformalising the intervals so that measured coverage becomes a guarantee, and training the adapters to disagree along explicit distribution-shift hypotheses.

AI USE STATEMENT

We used generative AI tools as general-purpose assistants for minimal writing support (rephrasing and LaTeX formatting) and for code assistance in building the experimental pipeline to train LoRA adapters. Generative AI was not used to generate the research idea, to design the method, or to produce any experimental result reported here: every number in this paper comes from the runs described in Sec. 4 and the appendix. Frontier language models appear in this work as objects of study, as baselines, which is distinct from their use as authoring tools. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and artefacts.

ETHICS STATEMENT

This work studies uncertainty representation and abstention in question answering with publicly available models and datasets. The harmful-prompt experiment uses the public AdvBench prompt list only as input and reports refusal behaviour; no harmful content is generated for release. Set-valued and abstaining outputs are intended to reduce confidently wrong answers; they do not remove the need for human oversight in deployment.

REPRODUCIBILITY STATEMENT

All models and datasets are public. Sec. 3 defines every quantity; §D gives the adapter configurations, the decoding and equivalence procedures with their fixed constants (B=8, L=32, L=8 on multiple choice), and the numerical tolerance of the cached implementation; §D.4 gives the development splits, corruption protocol and correctness rules.

REFERENCES

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. Uncertainty in natural language generation: From theory to applications, 2023. URL https://arxiv.org/abs/2307.15703.

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned llms using lora ensembles, 2025. URL https://arxiv.org/abs/2402.12264.

James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, et al. An overview of robust bayesian analysis. *Test*, 3:5–124, 06 1994. doi: 10.1007/BF02562676.

Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark Van Der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4345–4359. PMLR, 2025. URL https://proceedings.mlr.press/v267/bickford-smith25a.html.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001. doi: 10.1214/ss/1009213726. URL https://doi.org/10.1214/ss/1009213726.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. Palm: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(1):1–113, 2023. ISSN 1532-4435.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Fabio Cuzzolin. The intersection probability and its properties. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pp. 287–298. Springer, 2009.

Fabio Cuzzolin. *Belief functions: theory and applications*. Springer, 2014.

Fabio Cuzzolin. Uncertainty measures: A critical survey. *Information Fusion*, 114:102609, 2025. ISSN 1566-2535. URL https://www.sciencedirect.com/science/article/pii/S1566253524003877.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Matthew F. Dixon. Calibrating semantic uncertainty from observable language-model probabilities, 2026. URL https://arxiv.org/abs/2607.17447.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-024-07421-0.

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 2025)*, pp. 36–50. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.uncertainlp-main.5. URL https://aclanthology.org/2025.uncertainlp-main.5/.

Xabier Gonzalez-Garcia, Siu Lun Chau, Julian Rodemann, Michele Caprio, Krikamol Muandet, Humberto Bustince, Sebastien Destercke, Eyke Hüllermeier, and Yusuf Sale. Quantification of credal uncertainty: A distance-based approach. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, pp. 1725–1747. PMLR, 2026. URL https://proceedings.mlr.press/v337/gonzalez-garcia26a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 1321–1330. PMLR, 2017.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2604.17112.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://dx.doi.org/10.1007/s10994-021-05946-3.

Shihao Ji, Zihui Song, and Jiajie Huang. Credal transformer: A principled approach for quantifying and mitigating hallucinations in large language models. *arXiv preprint arXiv:2510.12137*, 2025. URL https://arxiv.org/abs/2510.12137.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. URL https://aclanthology.org/2021.tacl-1.57/.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Conformal linguistic calibration: Trading-off between factuality and specificity, 2025. URL https://arxiv.org/abs/2502.19110.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Ethan Tran-Johnson, Scott Johnston, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

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. URL https://arxiv.org/abs/2406.15927.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pdjkikvCch.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Tom A. Lamb, Desi R. Ivanova, Philip H. S. Torr, and Tim G. J. Rudner. Improving semantic uncertainty quantification in language model question-answering via token-level temperature scaling, 2026. URL https://arxiv.org/abs/2604.07172.

Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1980.

Mingda Li, Xinyu Li, Weinan Zhang, and Longxuan Ma. ESI: Epistemic uncertainty quantification via semantic-preserving intervention for large language models. 2025a. URL https://arxiv.org/abs/2510.13103.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. Semantic volume: Quantifying and detecting both external and internal uncertainty in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31751–31759, 2026. doi: 10.1609/aaai.v40i37.40443. URL https://ojs.aaai.org/index.php/AAAI/article/view/40443.

Yinghao Li, Vianne Gao, Chao Zhang, and MohamadAli Torkamani. Ensembles of low-rank expert adapters. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e8711daef520be07cb9852390c673de8-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1906–1919, 2020.

Lucas H. McCabe, Rimon Melamed, Thomas Hartvigsen, and H. Howie Huang. Estimating semantic alphabet size for LLM uncertainty quantification. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.14478.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, 2018.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

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*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/10c456d2160517581a234dfde15a7505-Abstract-Conference.html.

Ruijia Niu, Dongxia Wu, Rose Yu, and Yi-An Ma. Functional-level uncertainty quantification for calibrated fine-tuning on LLMs. *arXiv preprint arXiv:2410.06431*, 2024. URL https://arxiv.org/abs/2410.06431.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/f26d4fbaf7dfa115f1d4b3f104e26bce-Abstract-Conference.html.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 7:249–266, 2019.

Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference. In *PROCEEDINGS 31. WORKSHOP COMPUTATIONAL INTELLIGENCE*, volume 25, pp. 63, 2021.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.54. URL https://aclanthology.org/2024.findings-emnlp.54/.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Theja Tulabandhula, Ranganath Krishnan, and Amit Ranjan Trivedi. CAP: Conformalized abstention policies for context-adaptive risk management for LLMs and VLMs. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 926–941. PMLR, 2025. URL https://proceedings.mlr.press/v304/tayebati26a.html.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1991. ISBN 9780412286605. URL https://books.google.de/books?id=-hbvAAAAMAAJ.

Kaizheng Wang, Fabio Cuzzolin, Shireen Kudukkil Manchingal, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal deep ensembles for uncertainty quantification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=cv2iMNWCsh.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural networks*, 185:107198, 2025b.

Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning, 2023a. URL https://arxiv.org/abs/2310.00035.

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

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in neural information processing systems*, 37:67758–67794, 2024b.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19052–19075. Association for Computational Linguistics, 2025c. doi: 10.18653/v1/2025.acl-long.934. URL https://aclanthology.org/2025.acl-long.934/.

Adam Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 1812–1842, 2024.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris N. Metaxas, and Hao Wang. TokUR: Token-level uncertainty estimation for large language model reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VHQc7wzmYv.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021. URL https://arxiv.org/abs/2102.09690.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

APPENDIX: CREDAL LARGE LANGUAGE MODELS

This appendix provides the theoretical properties of credal commitment (§A), extended related work (§B), the detailed methodology with the full algorithm and a worked example (§C), implementation details (§D), additional experimental results, organised as questions and answers (§E), and the broader impact of this work (§F).

A THEORETICAL PROPERTIES OF CREDAL COMMITMENT

In this section, we show that the minimum and maximum over adapters are the exact lower and upper probabilities of the credal set (§A.1), that credal commitment reduces to argmax prediction when the adapters agree (§A.2), and that the commitment sets can only grow with the number of adapters (§A.3).

A.1 BOUNDS AS CREDAL-SET ENVELOPES

Proposition A.1 (The bounds are the credal set's envelopes)**.** *Let $\{p_m\}_{m=1}^M$ be distributions on a finite set $\mathcal{A}$ and let $\mathcal{P}(\mathcal{A}) = \mathrm{co}\{p_1, \ldots, p_M\}$ be their convex hull, as in Eq.* (1)*. Then for every $a \in \mathcal{A}$,*

$$\min_{q \in \mathcal{P}(\mathcal{A})} q(a) = \min_m p_m(a) = \underline{P}(a), \qquad \max_{q \in \mathcal{P}(\mathcal{A})} q(a) = \max_m p_m(a) = \overline{P}(a),$$

and both are attained. Consequently Eq. (2) *gives the exact lower and upper probabilities of the credal set on single candidates, not an outer approximation of them.*

Proof. Every $q \in \mathcal{P}(\mathcal{A})$ is $q = \sum_m \lambda_m p_m$ with $\lambda_m \geq 0$ summing to one, so $q(a) = \sum_m \lambda_m p_m(a)$ is a convex combination of the numbers $p_1(a), \ldots, p_M(a)$ and therefore lies in $[\min_m p_m(a), \max_m p_m(a)]$. Both endpoints are attained by the vertices: taking λ to be the indicator of an m^- with $p_{m^-}(a) = \min_m p_m(a)$ gives $q = p_{m^-} \in \mathcal{P}(\mathcal{A})$ with $q(a) = \underline{P}(a)$, and likewise for m^+ and $\overline{P}(a)$. Hence the envelopes equal the stated minima and maxima and neither bound can be tightened. □

Remark A.2. The interval-consistent set $\mathcal{K} = \{q : \underline{P}(a) \leq q(a) \leq \overline{P}(a)\ \forall a,\ \sum_a q(a) = 1\}$ contains $\mathcal{P}(\mathcal{A})$ and has the same envelopes on single candidates, by Proposition A.1 and because each p_m lies in $\mathcal{K}$. The commitment rule of Eq. (3) reads only those envelopes, so it is unchanged if $\mathcal{K}$ is used in place of $\mathcal{P}(\mathcal{A})$.

A.2 SINGLETON LIMIT

Proposition A.3 (Singleton limit: credal commitment reduces to ordinary prediction)**.** *Let $\{p_m\}_{m=1}^M$ be the M adapter distributions over a finite candidate set $\mathcal{A}$, and suppose all adapters agree, $p_m = p^*$ for all m. Then $\underline{P}(a) = \overline{P}(a) = p^*(a)$ for every a, and the credal commitment set of Eq.* (3) *is the set of maximisers of p^*, $\mathcal{D}(\mathcal{A}) = \arg\max_a p^*(a)$. In particular the model commits, and commits to the argmax, whenever p^* has a unique maximiser. Applied at every decoding step, credal decoding with agreeing adapters is greedy decoding.*

Proof. With $p_m = p^*$ for all m, $\underline{P}(a) = \min_m p_m(a) = p^*(a) = \max_m p_m(a) = \overline{P}(a)$. By Eq. (3), $a \in \mathcal{D}(\mathcal{A})$ iff $\overline{P}(a) \geq \max_b \underline{P}(b)$, i.e. iff $p^*(a) \geq \max_b p^*(b)$, which holds exactly for the maximisers of p^*. If the maximiser is unique, $|\mathcal{D}(\mathcal{A})| = 1$ and the committed candidate is $\arg\max_a p^*(a)$. At a decoding step with $\mathcal{A} = \mathcal{V}$ this is the greedy token, so every partial sequence has a single continuation and credal decoding produces the greedy completion. □

The proposition says that credal commitment departs from argmax prediction only where the adapters disagree: any input on which the model presents a set rather than committing is an input on which no candidate's lower probability clears every other candidate's upper probability, and $|\mathcal{D}|$ is a direct, hyperparameter-free measure of how wide the credal set is at that input. The converse does not hold, and we do not claim it: a candidate can survive the rule without any single adapter ranking it first.

A.3 Monotonicity in the Number of Adapters

Proposition A.4 (Monotonicity in the number of predictors)**.** *Let $\mathcal{D}^{(M)}(\mathcal{A})$ be the commitment set from adapters $f_1, \ldots, f_M$ and $\mathcal{D}^{(M+1)}(\mathcal{A})$ the set after adding f_{M+1}. Then $\mathcal{D}^{(M)}(\mathcal{A}) \subseteq \mathcal{D}^{(M+1)}(\mathcal{A})$.*

Proof. Adding a predictor can only decrease each $\underline{P}(a)$ and increase each $\overline{P}(a)$. If $a \in \mathcal{D}^{(M)}$ then $\overline{P}^{(M)}(a) \geq \max_b \underline{P}^{(M)}(b) \geq \max_b \underline{P}^{(M+1)}(b)$, and $\overline{P}^{(M+1)}(a) \geq \overline{P}^{(M)}(a)$, so $a \in \mathcal{D}^{(M+1)}$. □

The proposition is stated at a fixed candidate set. In the token space the candidate set is the vocabulary and does not depend on M, so the commitment sets are nested as stated and the token-space commit rate falls monotonically with M in every cell of Fig. 10. In the semantic space the candidate set itself is built by credal decoding and by the adapters' equivalence judgements, both of which change when an adapter is added, so the semantic commit rate is not required to be monotone and is not: on Gemma it rises from 44.8 to 53.8% between $M{=}2$ and $M{=}3$ before falling to 48.8% at $M{=}5$. This is a change of candidate set, not a violation of the proposition.

B Extended Related Work

Imprecise probability and robust Bayesian analysis distinguish a set of plausible distributions from a selected predictive mixture (Walley, 1991; Levi, 1980; Berger et al., 1994). Belief functions provide related representations through masses on subsets, with different structural assumptions (Cuzzolin, 2014; 2025). Recent distance-based measures quantify uncertainty within credal sets using integral probability metrics (Gonzalez-Garcia et al., 2026). These measures summarise the set's geometry, whereas interval dominance makes a decision using lower and upper expectations of candidate indicators. CLLM uses this established decision principle. Its finite adapter hull need not contain the data-generating predictor, and its marginal envelopes do not retain every dependency in that hull.

Credal learning has also reached language applications. Garces Arias et al. (2025) construct distributions of semantic, lexical and syntactic characteristics from creative continuations and compare their credal geometry with human variation. This evaluates diversity rather than choosing an answer from a prompt-specific set of adapter predictions. Ji et al. (2025) parameterise attention with Dirichlet evidence, use its expectation as attention weights, and derive vacuity from total concentration. Their construction operates inside the Transformer; CLLM forms an output-level convex hull without replacing the attention mechanism. Classification methods such as CreINNs, credal wrappers and random-set neural networks provide additional predecessors (Wang et al., 2025b;a; Kudukkil Manchingal et al., 2025). Moving to generated answers requires both a candidate search and a semantic outcome space, which are explicit components of CSC. Parameter-efficient uncertainty estimation includes LoRA ensembles, Laplace approximations and Bayesian adapter posteriors (Wang et al., 2023a; Balabanov & Linander, 2025; Yang et al., 2024; Wang et al., 2024b). ELREA trains expert adapters on instruction clusters formed using gradient information and combines relevant experts at inference (Li et al., 2025b). UQ4CT learns calibrated mixtures of LoRA experts in function space (Niu et al., 2024), while TokUR uses low-rank random perturbations during decoding (Zhang et al., 2026). These methods differ in how they generate or weight predictive distributions. CLLM retains an unweighted set of trained predictors for its bounds, so its contribution should be assessed against alternative readouts of the same predictors as well as against other model families.

Self-consistency aggregates sampled reasoning answers (Wang et al., 2023b), and SelfCheckGPT measures factual consistency among sampled responses without an external knowledge base (Manakul et al., 2023). Lin et al. (2024) distinguish dispersion over possible outputs from confidence in a particular response and develop semantic measures for selective generation. Semantic entropy accounts for equivalent surface forms through entailment-based grouping (Kuhn et al., 2023; Farquhar et al., 2024). Kernel Language Entropy uses a semantic kernel and its von Neumann entropy, retaining similarities beyond hard equivalence classes (Nikitin et al., 2024); Semantic Density provides response-specific confidence using semantic distances (Qiu & Miikkulainen, 2024). These are relevant alternatives to CSC's hard partition, whose equivalence judgements can merge distinct answers or separate valid paraphrases. The cosine semantic-entropy variant in our experiments is an implementation comparison, not the original entailment-based method.

The computation required to estimate semantic uncertainty is a separate concern. Semantic Entropy Probes learn to approximate semantic entropy from hidden representations of a single generation (Kossen et al., 2024). McCabe et al. (2026) estimate the number of semantic classes and adjust discrete entropy for incomplete sample coverage. CSC instead calculates the probability mass omitted by its bounded search under each adapter and allows that mass to enter observed or unseen classes. This is a bound on the length-capped output distribution, conditional on the grouping of observed candidates; it neither estimates the true number of meanings nor guarantees that generated candidates contain the correct answer. The width cap can discard undominated sequences, and low explored mass can prevent commitment. Variation can arise from the input as well as the fitted model. Input clarification ensembling studies ambiguity through alternative clarified prompts (Hou et al., 2024); ESI measures response changes under interventions intended to preserve prompt meaning (Li et al., 2025a). Semantic Volume perturbs queries and responses and measures dispersion through the determinant of an embedding Gram matrix (Li et al., 2026). Hamidieh et al. (2026) use the difference between inter-model and intra-model semantic similarity to complement self-consistency. These approaches examine sources of variation beyond one sampled response distribution. CLLM varies adapters on a shared backbone, so common pretraining errors and prompt-induced ambiguity can remain even when all adapters agree.

Uncertainty representations, confidence calibration and statistical guarantees address different questions. Baan et al. (2023) explain why linguistic variation complicates the epistemic–aleatoric distinction, and Bickford Smith et al. (2025) examine the interpretation of uncertainty through decision theory. Accordingly, adapter disagreement should be understood as sensitivity to the chosen fits, rather than a complete measurement of reducible ignorance. Lamb et al. (2026) distinguish semantic calibration from discrimination and study temperature scaling; Dixon (2026) connects verbal-response probabilities to declared latent states through a separately validated semantic map. CSC's generated meaning classes are not such a calibrated state posterior. Its lower and upper probabilities are model-based bounds, and successful error ranking does not by itself validate their calibration. Conformal methods provide a complementary route to uncertainty-aware outputs. Conformal Language Modeling calibrates sampling and rejection procedures to return answer sets (Quach et al., 2024). API Is Enough constructs nonconformity measures from sample frequencies and semantic similarity without logits (Su et al., 2024). Conformal factuality progressively reduces specificity to obtain correctness guarantees (Mohri & Hashimoto, 2024), and Conformal Linguistic Calibration interprets uncertainty expressions as set predictions and calibrates claim rewriting (Jiang et al., 2025). SConU uses conformal significance tests to identify uncertainty outliers (Wang et al., 2025c); CAP learns utility-driven policies balancing a point answer, a prediction set and abstention (Tayebati et al., 2025). Their guarantees depend on the stated calibration and sampling assumptions and should not be extended to arbitrary distribution shifts. CLLM uses interval dominance without a separately fitted confidence cutoff, but has no corresponding distribution-free risk guarantee. Conformal calibration of credal scores would be an additional procedure, not a property of the present rule.

Finally, verbalised confidence and self-knowledge methods elicit or learn an assessment of a response's reliability (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023). Such assessments can support selective answering, but answer correctness, input-corruption detection, and willingness to answer are distinct evaluation targets. Our corrupted-context AUROC measures the second target. The studies above motivate richer semantic and calibrated baselines; citing them does not constitute an empirical comparison. The reported experiments support conclusions about the evaluated baselines and the controlled comparison with the same mean-pooled ensemble.

C Detailed Methodology

In this section, we state the method as one algorithm (§C.1) and provide a step by step example of the method (§C.2).

C.1 Algorithm

Algorithm 1 applies the rule of Eq. (3) to the answer tokens, generates the candidate completions by credal decoding, and applies the same rule to the meanings of those completions.

Algorithm 1 Credal Large Language Models: one rule, two spaces. Everything the model returns is produced by Eq. (3) applied to a different candidate set; the only operational quantities are the width B and the maximum length L.

Require: prompt x; adapters $f_1, \dots, f_M$ on a frozen backbone; width B; maximum length L
Ensure: a commitment or a set, at each of the two depths

function DOMINANCE($\{\underline{P}(a)\}, \{\overline{P}(a)\}$) ▷ Eq. (3), used at every depth
  return $\{a : \overline{P}(a) \geq \max_b \underline{P}(b)\}$
end function

Depth one: Credal Token Commitment
$p_m(\cdot) \leftarrow f_m(x)$ for every m ▷ one forward pass per adapter, whole vocabulary
$\underline{P}(a) \leftarrow \min_m p_m(a)$, $\overline{P}(a) \leftarrow \max_m p_m(a)$ for every token a ▷ Eq. (2)
$\mathcal{D}_1 \leftarrow$ DOMINANCE($\underline{P}, \overline{P}$)
if $|\mathcal{D}_1| = 1$ **then** commit to $a^* \in \mathcal{D}_1$ with confidence $\underline{P}(a^*)$ **else** present $\mathcal{D}_1$

Credal decoding
live $\leftarrow \{(\varepsilon,\ \ell_m{=}0)\}$, $\mathcal{H} \leftarrow \emptyset$, $\underline{q}^* \leftarrow 0$
for $t = 1$ to L and while live $\neq \emptyset$ **do**
  $C \leftarrow \emptyset$
  for each partial sequence $(y_{<t}, \ell)$ in live **do**
    $p_m(\cdot) \leftarrow f_m(x, y_{<t})$ for every m
    for each continuation y **do**
      $\ell'_m \leftarrow \ell_m + \log p_m(y \mid x, y_{<t})$; $\overline{q} \leftarrow \max_m \exp \ell'_m$ ▷ upper mass of the extension
      if $\overline{q} \geq \underline{q}^*$ **then** add $(y_{<t} \circ y,\ \ell')$ to C ▷ no adapter rules it out
    end for
  end for
  keep in C the B entries of largest $\overline{q}$ ▷ width; anything dropped stays visible through u_m
  move finished entries of C into $\mathcal{H}$; live $\leftarrow C \setminus \mathcal{H}$
  $\underline{q}^* \leftarrow \max_{h \in \mathcal{H}} \min_m p_m(h \mid x)$ ▷ best lower mass so far
end for

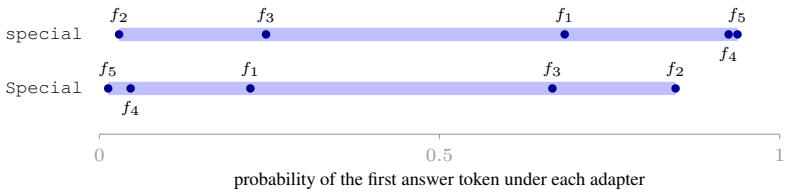
Depth two: Credal Semantic Commitment
merge sequences of $\mathcal{H}$ that are identical after lower-casing and removing punctuation and articles into candidates a, with $p_m(a) \leftarrow \sum_{h \in a} p_m(h \mid x)$
$u_m \leftarrow 1 - \sum_a p_m(a)$ for every m ▷ unexplored mass, Eq. (4)
$e_m(a, b) \leftarrow$ adapter m's probability of "yes" when asked whether a and b answer x alike
$\mathcal{C}(x) \leftarrow$ greedily group the candidates, in decreasing mass, joining a to the meaning whose representative b maximises $\min_m e_m(a, b)$ when that value is $\geq \frac{1}{2}$ ▷ the adapters judge meaning
$\underline{P}(c) \leftarrow \min_m \sum_{a \in c} p_m(a)$, $\overline{P}(c) \leftarrow \max_m \left[\sum_{a \in c} p_m(a) + u_m\right]$ ▷ Eq. (5)
$\mathcal{D}(\mathcal{C}(x)) \leftarrow$ DOMINANCE($\underline{P}, \overline{P}$)
if $|\mathcal{D}(\mathcal{C}(x))| = 1$ **and** $\underline{P}(c^*) \geq \max_m u_m$ **then** ▷ second condition: no unseen meaning could dominate
  commit to c^* with confidence $\underline{P}(c^*)$
else
  present $\mathcal{D}(\mathcal{C}(x))$, with an unseen element when the second condition fails
end if

C.2 WORKED EXAMPLE THROUGH BOTH DEPTHS

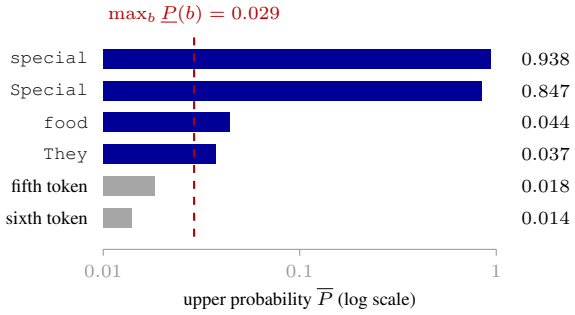
This section follows one prompt through the whole method, drawn step by step below, with the numbers taken from the stored caches of the Llama-3.1-8B run on the CoQA test split ($M{=}5$, $B{=}8$, $L{=}32$), so that every quantity defined in §3 can be seen on a real input. The passage describes Diwali and ends with the sentence that families prepare special holiday meals to celebrate the event. The question is *What do they make for the event?* and the gold answer is *special holiday meals*.

Step 1: the adapters disagree about the first answer token. One forward pass of the five adapters gives a distribution over the vocabulary at the first answer position. Tab. 4 lists the six most likely

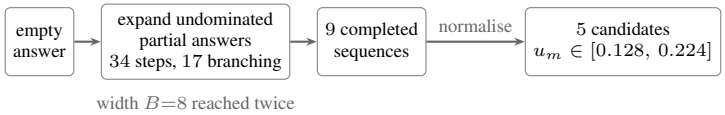
tokens with the probability each adapter assigns. Three adapters put most of their mass on the lower-case token `special`, two prefer the capitalised `Special`, and the spread on `special` alone runs from 0.029 to 0.938. The bounds of Eq. (2) are the column-wise minimum and maximum, so $\underline{P}(\texttt{special}) = 0.029$ and $\overline{P}(\texttt{special}) = 0.938$.



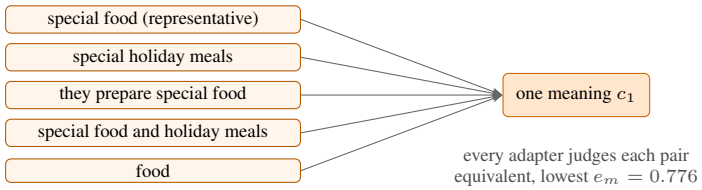
Step 2: the rule in the token space. The largest lower probability over all tokens is $\max_b \underline{P}(b) = 0.029$, attained by `special`. By Eq. (3) a token survives when its upper probability reaches that value. Four tokens do: `special`, `Special`, `food` and `They`, with upper probabilities 0.938, 0.847, 0.044 and 0.037. The two remaining tokens in Tab. 4 have upper probabilities 0.018 and 0.014 and are ruled out. Since $|\mathcal{D}_1| = 4$, Credal Token Commitment does not commit and would present the four surviving tokens. This is the failure mode the token space is prone to: the adapters agree about the answer and disagree about how to spell its first token.



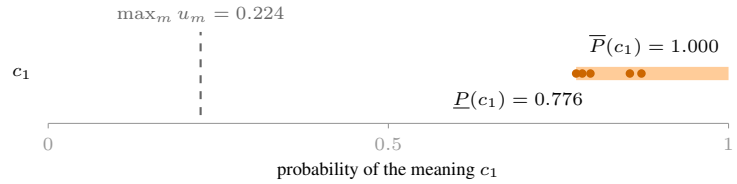
Step 3: credal decoding. Starting from the empty partial sequence, the adapters are run one step at a time and every partial sequence that no adapter rules out is expanded. The search takes 34 decoding steps, of which 17 expand more than one continuation, and hits the width B=8 twice, which is recorded as a truncation. It ends with nine completed sequences. After the surface normalisation of §3.4, which lower-cases and removes punctuation and articles, those nine become five candidates, each carrying the probability every adapter assigns it. The probabilities of the candidates do not sum to one for any adapter, and the shortfall is the unexplored mass u_m of Eq. (4), which ranges from 0.128 to 0.224 across the five adapters.



Step 4: the adapters group the candidates into meanings. Candidates are judged in decreasing order of mean mass, so *special food* is the representative, and each remaining candidate is judged against it by all five adapters. The five judgements for each of the four pairs are all above one half, the lowest being 0.776, so every adapter agrees that the five candidates say the same thing. The surely-equivalent relation therefore has a single class and the semantic space is $\mathcal{C}(x) = \{c_1\}$ with c_1 containing all five candidates.



Step 5: the rule in the semantic space. Adapter m's mass on c_1 is the sum of its candidate probabilities, which gives 0.855, 0.785, 0.872, 0.776 and 0.797 for the five adapters. By Eq. (5), $\underline{P}(c_1) = 0.776$ and $\overline{P}(c_1) = \max_m[P_m(c_1) + u_m] = 1.000$. The commitment set is $\mathcal{D}(\mathcal{C}(x)) = \{c_1\}$, and the second condition holds as well, because $\underline{P}(c_1) = 0.776$ is at least the largest unexplored mass $\max_m u_m = 0.224$, so no meaning the search never reached could still dominate. The model is semantically committed to c_1 with confidence 0.776.



singleton set and $0.776 \geq 0.224$: commit to c_1 with confidence 0.776

Step 6: was it right? The gold answer *special holiday meals* is one of the five candidates and lies in the committed meaning, so the commitment is correct. The example is the common case reported in §4.2: the token space is defeated by a capitalisation split that has nothing to do with the answer, the semantic space is not, and the confidence attached to the decision is the mass that survives the most pessimistic adapter rather than an average over them.

Table 4: **The token space of the worked example.** Probability of each of the six most likely first answer tokens under each of the five adapters, with the bounds of Eq. (2). The largest lower probability is 0.029, so every token whose upper probability reaches 0.029 survives the rule of Eq. (3). Four tokens survive, so the token space does not commit.

Token	f_1	f_2	f_3	f_4	f_5	$\underline{P}$	$\overline{P}$	In $\mathcal{D}_1$
`special`	0.684	0.029	0.245	0.925	0.938	0.029	0.938	yes
`Special`	0.222	0.847	0.666	0.046	0.013	0.013	0.847	yes
`food`	0.044	0.005	0.016	0.022	0.036	0.005	0.044	yes
`They`	0.001	0.037	0.002	0.000	0.000	0.000	0.037	yes
fifth token	0.018	0.003	0.002	0.004	0.006	0.002	0.018	no
sixth token	0.004	0.014	0.002	0.000	0.000	0.000	0.014	no

D Implementation Details

In this section, we describe the backbones, adapters and inference (§D.1), the meaning-equivalence procedure (§D.2), the baseline implementations (§D.3), the experimental setup (§D.4), and the metrics pipeline and compute (§D.5).

D.1 Backbones, Adapters and Inference

Backbones. Gemma-2-9B-Instruct, Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, all frozen.

LoRA training. M=5 adapters per (backbone, dataset), trained for one epoch on a bootstrap resample of the dataset's training split, batch size 2 with gradient accumulation 4, maximum sequence length 1024, α=16, dropout 0.1. The adapters are deliberately heterogeneous, with per-adapter (target modules, rank, learning rate, seed): m=1: {q, v}, r=4, 10^{-4}, seed 42; m=2: {q, k, v, o}, r=8, 3×10^{-4}, seed 43; m=3: {o, gate, up}, r=16, 5×10^{-4}, seed 44; m=4: {q, v, down}, r=4, 2×10^{-4}, seed 45; m=5: {k, v, gate}, r=8, 10^{-4}, seed 46. The seed controls initialisation, the

bootstrap draw and data order. This construction treats the ensemble as a set of alternative fine-tuning decisions rather than as replicas of one recipe, which is what makes the family a sensitivity analysis over choices the data does not adjudicate rather than a set of reruns of the same choice. We did not train seed-only or bootstrap-only ensembles at the same budget, so the paper does not separate the contribution of the heterogeneity from that of the resampling.

Inference. Each adapter is run in evaluation mode (dropout off) with the same chat template. Token-space bounds come from one forward pass per adapter on the prompt. Credal decoding runs one priming pass per adapter over the prompt and then one-token steps over the partial sequences with a per-adapter key-value cache; the partial sequences share the prompt and have equal length, so no padding is involved. The first-step distribution is the token-space distribution of the same prompt, so CTC and CSC read the same bounds at depth one. Against a full re-encoding of each partial sequence, cached probabilities agree to within bf16 precision (at most 0.03 on any token), which can flip a dominance decision only when two bounds are within that tolerance; prompts longer than 2048 tokens are truncated and counted. Answers end at the backbone's end-of-turn token or at a newline, whichever comes first.

D.2 Meaning Equivalence

Meaning equivalence. Candidates are judged in decreasing order of mean mass. Each new candidate is judged against the representative (highest-mass member) of every existing meaning, by all adapters in one batched pass, with the query "Question: x. Answer 1: a. Answer 2: b. Do Answer 1 and Answer 2 give the same answer to the question? Reply with yes or no." The judgement $e_m(a, b) = p_m(\text{yes})/[p_m(\text{yes}) + p_m(\text{no})]$ is the adapter's first-response-token probability of "yes" renormalised against "no", each summed over its capitalised and space-prefixed single-token spellings. It joins the meaning whose lower equivalence probability is largest if that value is at least $\frac{1}{2}$, and otherwise starts a new meaning; upper equivalence probabilities of at least $\frac{1}{2}$ link meanings in the coarser partition. This assumes equivalence is transitive within a meaning and costs a few dozen short passes per prompt rather than all pairs. The semantic-entropy baselines use their own clustering: microsoft/deberta-large-mnli bidirectional entailment, as in the original papers (Kuhn et al., 2023; Farquhar et al., 2024), and a cosine-similarity variant with BAAI/bge-base-en-v1.5 embeddings. The quality of both groupings is reported as the rate of prompts on which one meaning contains both a gold-matching and a non-matching candidate.

D.3 Baseline Implementations

Baseline implementations. *Standard LLM*: the base model with the same chat template and prompt, its first-token predictive entropy and max probability, its greedy answer, and $K{=}16$ samples at temperature 0.8 with maximum length L and their sequence log-probabilities. *Semantic entropy (cosine)*: the 16 samples clustered greedily by cosine similarity of BAAI/bge-base-en-v1.5 embeddings at threshold 0.8 (letters on multiple choice), with the length-normalised, renormalised sequence probabilities of Kuhn et al. (2023) as cluster masses; *Semantic entropy (NLI)*: the same samples clustered by bidirectional entailment under microsoft/deberta-large-mnli, a pair sharing a meaning only if each entails the other (argmax, no threshold), following Farquhar et al. (2024). *LoRA Ensemble*: the five adapters' first-token distributions averaged; predictive entropy of the mean, mutual information (mean entropy minus expected entropy) and the variance across adapters of the probability of each adapter's most likely token, as in the original code base. *Bayesian-LoRA (KFAC)* and *Laplace-LoRA (diagonal)*: a Laplace posterior over the LoRA weights of the first adapter of each ensemble (the MAP point), with the Fisher estimated on 200 training examples of the same dataset: for the diagonal method (Daxberger et al., 2021) the sum of squared per-example gradients of the answer-token log-likelihood, for KFAC the Kronecker factors of the bayesian-lora package (Yang et al., 2024) at the first answer position with sampled labels and rank-10 low-rank factors for wide layers. The prior precision is $\lambda = P/\|\theta_{\text{MAP}}\|^2$, the maximum-likelihood prior variance of the adapter weights, so that posterior noise in directions without curvature is at the scale of the weights themselves; we record that with the fixed prior precision of 1 and temperature 0.05 used in the earlier version of this work the posterior samples destroy the adapter (near-uniform first-token distributions, noise 25–45 times the weights), and that maximising the Laplace marginal likelihood on a 200-example Fisher is ill-posed because most LoRA directions have no curvature. Each prompt uses $S{=}20$ posterior

samples; the scores are the predictive entropy of the posterior predictive (primary) and the mutual information across samples; the answer is the MAP model's greedy answer.

D.4 Experimental Setup

Backbones and ensemble size. The three backbones span the dominant open-weight tokeniser and architecture families at the 7–9B scale. $M{=}5$ is a compute-bounded choice; by Proposition A.4 the commitment sets can only grow with M, so all comparisons are at fixed M and the dependence on M is reported once in §E.7.1.

Development splits. The width $B{=}8$ and the length $L{=}32$ ($L{=}8$ on multiple choice) are fixed on frozen development splits disjoint from every training and test set (the width sweep is in §E.7.2; at $L{=}32$, sequences cut at L are 1.9% of the completion set on CoQA and 11.7% on TriviaQA with the passage, for Llama-3.1-8B), and the cost κ of §3.4 is swept rather than fixed (Tab. 19): for OpenBookQA, 500 items from the unused remainder of the training split; for ARC-Challenge, the 299-item validation split; for CoQA and TriviaQA, 500 items each from the unused remainder of the training split. Test sets are never read during selection.

Datasets. OpenBookQA is four-way multiple choice with a supporting fact; ARC-Challenge is four-way multiple choice requiring multi-step reasoning; CoQA is conversational QA with short free-form answers; TriviaQA is open-domain factual QA with long-tail entity answers. On the two multiple-choice tasks the token space is the whole vocabulary, as everywhere else, and CTC is the decision: every surviving token is an option letter on 97.2 to 99.0% of questions without the rule being told that they must be (92.3 to 99.1% of surviving tokens, the shortfall concentrated on the few questions with large $\mathcal{D}_1$). On CoQA and TriviaQA the token space closes on at most 30% of clean prompts (Tab. 13) and CSC is the decision.

Corrupted-context protocol. Where the answer rests on supplied evidence, corruption replaces or perturbs that evidence and leaves the question intact (the passage in CoQA and TriviaQA, the supporting fact in OpenBookQA), so the question stays answerable in principle but the evidence no longer supports the answer. Where there is no evidence (closed-book TriviaQA and ARC-Challenge), the question is replaced and the answer side kept. The four-option test sets have 500 questions each. Each open-ended setting evaluates 250 clean and 250 corrupted prompts, and the paired OpenBookQA split the same 500 questions with and without corruption; a prompt is labelled corrupted if its evidence or question was replaced.

Correctness rules. Multiple choice: the decoded option letter equals the gold letter. Free-form answers: after SQuAD normalisation (lower case, punctuation and articles removed), the answer contains a gold alias as a token sequence, an alias contains the answer, or token F1 with an alias is at least 0.5; TriviaQA uses all accepted aliases of the public release, CoQA its single gold string. A presented set is correct if it contains the gold answer.

D.5 Metrics Pipeline and Compute

Metrics pipeline and provenance. All numbers are produced by one script that reads cached adapter outputs, applies the definitions of Sec. 3 once each, and writes a manifest mapping every table cell to its cache file, checksum, code commit and definition. Token-space results on multiple choice use the first-token bounds of the $M{=}5$ ensembles; semantic-space results come from credal decoding, with M stated per row. Unit tests on hand-computed examples cover the bounds, the commitment rule, credal decoding on a mock ensemble (including the width and the cached path), calibration and selective metrics.

Compute. Adapters were trained on single A100 nodes; credal decoding, the equivalence judgements and the baselines ran on a single L4 (24 GB) or A100 (40 GB) GPU. Per-stage wall-clock and memory are given in §E.9.

E Additional Experimental Results

In this section, we address the following questions based on our experimental findings:

- **Q1:** How accurate is CLLM at a fixed coverage, and how well calibrated is its confidence, against every baseline family on every backbone and dataset? (§E.1)
- **Q2:** Do the credal quantities detect the model's own wrong answers and corrupted context on every backbone, and does the detection depend on which scalar is read off the credal set? (§E.2)
- **Q3:** Does the rule's own operating point sit where a tuned threshold would, and does the rule withdraw under corrupted context without being tuned? (§E.3)
- **Q4:** Does the credal set contain the gold answer more often than a top-k set, and what does the rule add to the ranking of the same adapters? (§E.4)
- **Q5:** How does Credal Token Commitment behave on the four-option setting, and are its intervals calibrated without a calibration step? (§E.5)
- **Q6:** What does the semantic space add over the token space, and when does each depth commit? (§E.6)
- **Q7:** How sensitive are the commitment sets to the choices inside the method, namely the number of adapters, the expansion rule, the width, the unexplored mass and the grouping of candidates into meanings? (§E.7)
- **Q8:** How does commitment behave when a deployment declares the cost of a wrong answer? (§E.8)
- **Q9:** What does credal commitment cost at inference? (§E.9)

E.1 Selective Accuracy and Calibration per Backbone

Why selective accuracy and calibration come first. The claim of the paper is about decisions, not about a new score, so the first question is whether the decision the rule takes is at least as good as what a practitioner would get from the baselines they already use. That requires measures every baseline family can be asked for, which is why the comparison reports selective accuracy at a fixed coverage and the calibration of each method's own confidence, on the same prompts for every method; detection and set-valued prediction follow in §E.2 and §E.4.

Why these baselines. The Standard LLM is what a practitioner runs by default. The LoRA Ensemble is the same five adapters mean-pooled, which is the closest possible comparison, since any gap is attributable to the reading of the ensemble rather than to the ensemble itself. Bayesian-LoRA and Laplace-LoRA represent posterior approximations over the adapter weights rather than an explicit set of predictors. Semantic entropy with cosine clustering and with entailment clustering represents the sampling-based account of meaning, and both variants are included because the clustering choice, not the entropy, is what drives that family's numbers.

Tab. 5 and Tab. 6 repeat Tab. 2 on the other two backbones, reporting accuracy at 80% coverage together with the ECE and NLL of each method's own confidence, and Tab. 7 reports accuracy at 80% coverage on the four-option setting for the Standard LLM, the LoRA Ensemble and CLLM.

On Llama-3.1-8B (Tab. 5), CLLM has the highest accuracy at 80% coverage on CoQA (0.845), closed-book TriviaQA (0.865), OpenBookQA (0.980) and ARC-Challenge (0.890), and the second highest on TriviaQA with the passage (0.920 against 0.940 for LoRA Ensemble (predictive entropy)). On CoQA the single-model and posterior baselines reach 0.745–0.775. On Qwen2.5-7B (Tab. 6), CLLM has the highest accuracy at 80% coverage on CoQA (0.825) and OpenBookQA (0.980), is 0.002 behind the best on ARC-Challenge, and ties Semantic entropy (NLI) for second on closed-book TriviaQA (0.620 against 0.660), where it improves over the Standard LLM's 0.545. On Qwen its confidence is the best calibrated of all methods on both four-option datasets, with ECE 0.039 and 0.066 and NLL 0.209 and 0.327 on OpenBookQA and ARC-Challenge, and it has the lowest ECE on CoQA (0.114) and the lowest NLL on closed-book TriviaQA (0.535). Averaged over CoQA and TriviaQA with the passage, CLLM is above every single-model baseline in accuracy at 80% coverage on all three backbones, its sets contain the gold meaning on 89, 89 and 88% of questions (Tab. 12), and it is the best corrupted-context detector on Llama and Gemma (0.855 and 0.878). On the four-option setting (Tab. 7), the token-space lower bound is the best ranker on both datasets for Llama and Gemma and on OpenBookQA for Qwen; on ARC-Challenge with Qwen it ties the Standard LLM at 0.938, 0.002 behind the LoRA Ensemble, and it is at least as accurate as the Standard LLM in every cell.

Table 5: **Comparison with baselines on every dataset, Llama-3.1-8B** (CoQA, TriviaQA with and without the passage, OpenBookQA and ARC-Challenge). Each dataset carries three measures: *Acc.@80%*, the accuracy of the selected answer when the questions are ranked by the method's score and the easiest 80% are answered, higher is better; *ECE* and *NLL*, the expected calibration error and negative log likelihood of the method's own per-answer confidence, lower is better in both. For the credal row the confidence is the midpoint of the interval. Rows that select the same answer share that answer's confidence, so the two Standard LLM rows and the two LoRA Ensemble rows carry the same calibration. CLLM is one row, at the depth the answer type fixes. **Bold** = best per column, underlined = second-best; the grey row is ours.

	CoQA			TriviaQA (passage)			TriviaQA (closed book)			OpenBookQA			ARC-Challenge		
Score	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓
Standard LLM (predictive entropy)	0.775	0.083	0.511	0.895	0.075	**0.342**	0.785	0.117	0.563	0.965	0.040	0.240	0.858	0.104	0.504
Standard LLM (max prob)	0.770	0.083	0.511	0.890	0.075	**0.342**	0.785	0.117	0.563	0.968	0.040	0.240	0.860	0.104	0.504
Semantic entropy (cosine)	0.755	0.154	1.825	0.885	0.070	0.765	0.795	0.136	1.127	0.968	0.047	0.622	0.860	0.131	1.458
Semantic entropy (NLI)	0.745	0.177	0.871	0.905	0.291	0.565	0.810	0.229	0.829	0.945	0.175	0.447	0.860	**0.041**	0.821
Bayesian-LoRA (KFAC)	0.745	0.062	0.510	0.890	0.438	0.855	0.785	0.103	0.543	0.963	0.025	0.224	0.855	0.073	0.415
Laplace-LoRA (diagonal)	0.750	0.089	0.816	0.890	0.442	0.860	0.780	0.107	0.537	0.963	0.026	0.223	0.860	0.071	**0.414**
LoRA Ensemble (predictive entropy)	0.840	0.073	0.845	**0.940**	0.148	0.432	0.850	0.100	**0.429**	0.965	0.024	**0.182**	0.810	0.053	0.426
LoRA Ensemble (mutual information)	0.815	0.073	0.845	0.875	0.148	0.432	0.795	0.100	**0.429**	0.968	0.024	**0.182**	0.853	0.053	0.426
CLLM	**0.845**	**0.053**	**0.493**	0.920	**0.068**	0.420	**0.865**	**0.050**	0.461	**0.980**	**0.018**	0.185	**0.890**	0.069	0.419

Table 6: **Comparison with baselines on every dataset, Qwen2.5-7B** (CoQA, TriviaQA with and without the passage, OpenBookQA and ARC-Challenge). Each dataset carries three measures: *Acc.@80%*, the accuracy of the selected answer when the questions are ranked by the method's score and the easiest 80% are answered, higher is better; *ECE* and *NLL*, the expected calibration error and negative log likelihood of the method's own per-answer confidence, lower is better in both. For the credal row the confidence is the midpoint of the interval. Rows that select the same answer share that answer's confidence, so the two Standard LLM rows and the two LoRA Ensemble rows carry the same calibration. CLLM is one row, at the depth the answer type fixes. **Bold** = best per column, underlined = second-best; the grey row is ours.

	CoQA			TriviaQA (passage)			TriviaQA (closed book)			OpenBookQA			ARC-Challenge		
Score	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓	Acc.@80% ↑	ECE ↓	NLL ↓
Standard LLM (predictive entropy)	0.740	0.212	1.173	0.735	0.225	1.021	0.545	0.370	2.418	0.975	0.090	0.940	0.938	0.109	1.459
Standard LLM (max prob)	0.740	0.212	1.173	0.715	0.225	1.021	0.530	0.370	2.418	0.975	0.090	0.940	0.938	0.109	1.459
Semantic entropy (cosine)	0.745	0.207	4.255	0.750	0.208	3.084	0.585	0.371	6.550	0.920	0.091	1.752	0.897	0.113	2.341
Semantic entropy (NLI)	0.770	0.145	2.213	0.790	0.100	1.252	0.620	0.130	1.003	0.900	0.138	1.645	0.885	0.131	1.331
Bayesian-LoRA (KFAC)	0.770	0.128	0.665	0.845	0.067	0.485	0.555	0.225	1.119	0.975	0.085	0.831	0.935	0.106	1.392
Laplace-LoRA (diagonal)	0.770	0.118	**0.641**	0.845	0.062	0.473	0.555	0.212	1.067	0.975	0.080	0.767	0.935	0.104	1.284
LoRA Ensemble (predictive entropy)	0.743	0.126	0.894	**0.905**	**0.047**	**0.376**	**0.660**	**0.089**	0.566	0.978	0.058	0.235	**0.940**	0.080	0.362
LoRA Ensemble (mutual information)	0.790	0.126	0.894	0.840	**0.047**	**0.376**	0.595	**0.089**	0.566	0.978	0.058	0.235	0.930	0.080	0.362
CLLM	**0.825**	**0.114**	0.906	0.860	0.092	0.443	0.620	0.130	**0.535**	**0.980**	**0.039**	**0.209**	0.938	**0.066**	**0.327**

E.2 Hallucination and Corrupted-Context Detection

Fig. 5 shows the corrupted-context AUROC of every method on CoQA and TriviaQA with the passage for all three backbones. On Llama the CSC lower bound is ahead of every baseline on CoQA (0.867) and fourth on TriviaQA (0.844), behind semantic entropy (NLI) and the two posteriors. On Qwen it is second on both datasets (0.793 and 0.791), behind the LoRA Ensemble's predictive entropy on CoQA and its mutual information on TriviaQA. On Gemma TriviaQA the token-space lower bound (0.888) is ahead of every baseline.

Tab. 8 reports corrupted-versus-clean AUROC for scalar statistics read from the token-space credal set and for the same adapters mean-pooled. The credal statistics are: the entropy of the distribution that interpolates every interval by the same fraction (the intersection-probability transform of Cuzzolin (2009)), the upper entropy over the credal set (Wang et al., 2025a), the mean interval width, the token-space lower bound $1 - \underline{P}(a^*)$ and the unexplored mass $\max_m u_m(x)$ of Eq. (4). The two set entropies and the width stay within 0.08 of LoRA Ensemble (predictive entropy) in every cell, and Standard LLM (predictive entropy) is far below both on CoQA and TriviaQA. The ensemble, not the readout, is what separates corrupted from clean context by ranking; the contribution of the credal set is the decision of Sec. 3.2, not a ranker. The token-space lower bound departs furthest from the mean-pooled entropy, in opposite directions: it falls to 0.691 on Qwen CoQA against 0.797, and reaches 0.888 on Gemma TriviaQA against 0.683. The unexplored mass is the best statistic in the table on Qwen CoQA at 0.821, above every method in Fig. 5 for that cell, but falls to 0.666 on Gemma TriviaQA, so it stays an ablation rather than joining Tab. 1.

Table 7: Selective accuracy at 80% coverage on the four-option setting, OpenBookQA and ARC-Challenge (N=500; ensemble scores rank the representative answer, the Standard LLM ranks its greedy answer).

Dataset	**Backbone**	Standard LLM (predictive entropy) ↑	LoRA Ensemble (predictive entropy) ↑	**CLLM** ↑
OpenBookQA	Llama-3.1-8B	0.965	0.965	**0.980**
OpenBookQA	Gemma-2-9B	0.973	0.995	**0.998**
OpenBookQA	Qwen2.5-7B	0.975	0.978	**0.980**
ARC-Challenge	Llama-3.1-8B	0.858	0.810	**0.890**
ARC-Challenge	Gemma-2-9B	0.943	0.950	**0.953**
ARC-Challenge	Qwen2.5-7B	0.938	**0.940**	0.938

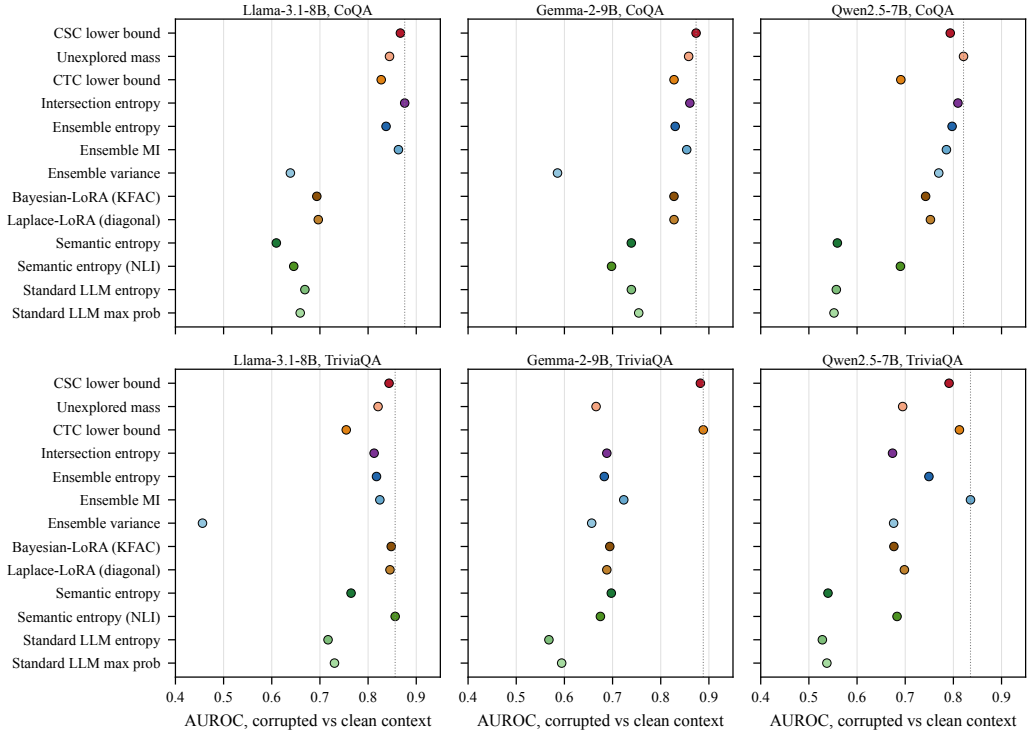


Figure 5: **Corrupted-context detection across all baseline families**, AUROC on 250 clean against 250 corrupted prompts per cell. Credal quantities are shown against the mean-pooled ensemble scores, the two posterior approximations and the single-model scores, for both open-ended datasets and all three backbones.

Table 8: **Scalar statistics read from the token-space credal set, as corrupted-versus-clean detectors** (AUROC; 250 clean and 250 corrupted prompts per cell, M=5). Every ensemble and credal row is read from the same five adapters. The rows that also appear in Tab. 1 carry the same values there. **Bold** = best per column, underlined = second best; grey rows are ours.

	CoQA			**TriviaQA (passage)**		
Statistic (AUROC ↑)	Llama	Gemma	Qwen	Llama	Gemma	Qwen
Single model, predictive entropy	0.669	0.739	0.557	0.717	0.568	0.528
Mean-pooled ensemble, predictive entropy	0.837	0.830	0.797	0.817	0.683	0.749
Mean-pooled ensemble, mutual information	0.863	0.854	0.786	**0.824**	0.723	**0.836**
Credal set, entropy of the interpolating distribution	**0.876**	**0.860**	0.809	0.813	0.688	0.674
Credal set, upper entropy	0.855	0.854	0.805	0.779	0.690	0.736
Credal set, mean interval width	**0.876**	0.824	0.786	0.799	0.667	0.807
Credal set, $1 - \underline{P}(a^*)$ (CTC)	0.827	0.828	0.691	0.755	**0.888**	0.813
Credal set, unexplored mass $\max_m u_m(x)$	0.844	0.858	**0.821**	0.821	0.666	0.695

E.3 Selective Prediction and Abstention

Why the corrupted-context protocol exists. Abstention is only meaningful if there are inputs on which the model should abstain. Replacing the evidential passage while leaving the question intact creates exactly that: the question remains answerable in principle but the evidence no longer supports any answer. It gives a clean contrast for detection, and it lets abstention be measured at a coverage matched to what the rule chose on clean prompts, which is the only fair way to compare a rule that sets its own coverage with baselines that need a threshold.

Fig. 6 extends the risk-coverage curves of Fig. 3 to both open-ended datasets and all three backbones. On CoQA, the operating point the rule chose for itself (filled marker) lies on the curves of the same adapters, and the single-model curves, the Standard LLM's entropy and semantic entropy, lie below the ensemble curves over most of the coverage range. On TriviaQA with the passage the rule commits on only 1–6% of questions (Tab. 12), so its operating point sits at the left edge, where it answers with 94–100% accuracy. Fig. 7 thresholds every method on clean prompts to the coverage the rule chose and reports how many prompts each still answers, clean and corrupted. On CoQA the credal rule answers 3.6–7.6% of corrupted prompts, whereas semantic entropy, which cannot be thresholded down to the target coverage (hatched), keeps answering most of them; on TriviaQA with the passage every method except semantic entropy answers almost nothing, because the coverage the rule chose is already close to zero.

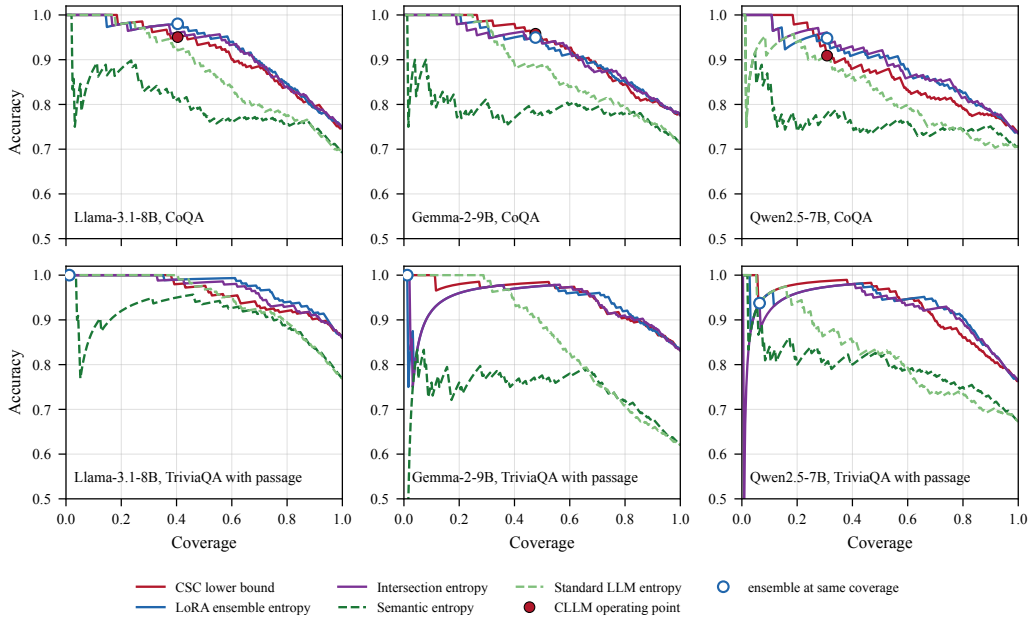


Figure 6: **Risk-coverage curves on clean prompts**, CoQA and TriviaQA with the passage, all three backbones. The horizontal axis is the fraction of questions answered when they are ranked by each method's score, the vertical axis the accuracy of the answers given; a curve that stays high as it moves right ranks its own errors well. The filled marker is the operating point our rule chose for itself, with no threshold; the hollow marker is the LoRA Ensemble at the same coverage.

E.4 Set-Valued Prediction

The presented meaning sets are compared below with the LoRA Ensemble top-k meanings and with the Standard LLM top-k sampled answers of the same size.

The credal set contains the gold meaning on 86–91% of clean questions with the passage and on 76–88% of closed-book questions, against 66–86% for the Standard LLM's top-k of the same size and 62–77% for its greedy answer. Under corrupted context the gap widens: on TriviaQA the credal set still contains the gold answer on 73–76% of questions, against 5–35% for the Standard LLM's top-k. Tab. 9 separates the size of the set from its content. Against a k fixed in advance for the whole dataset, the credal set contains the gold meaning more often than the top-k of the same adapters for

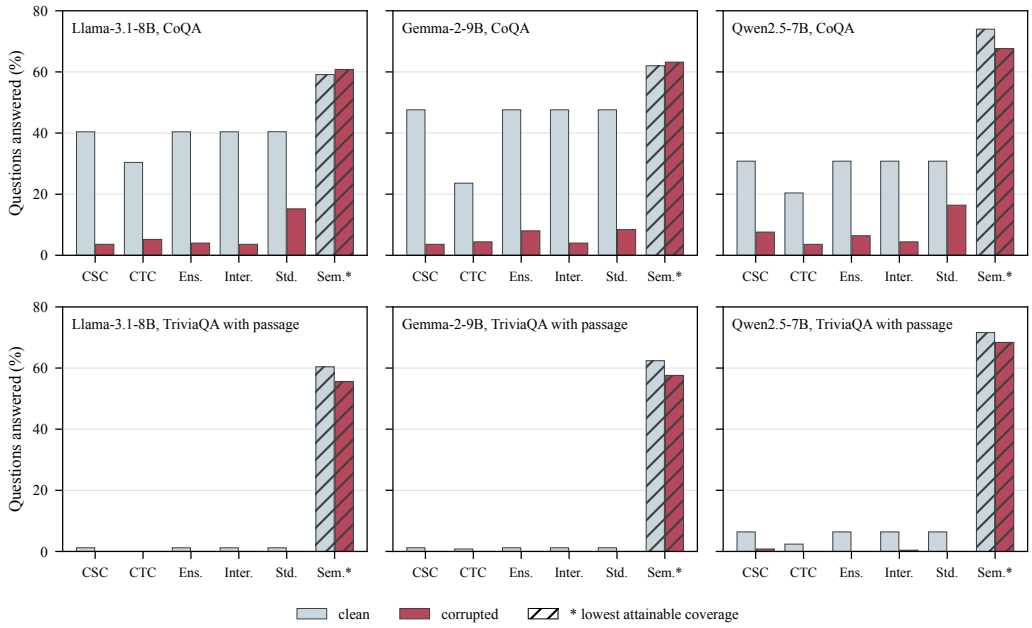


Figure 7: **Abstention at matched clean coverage**, both open-ended datasets, all three backbones. Each method is thresholded on clean prompts to the coverage our rule chose for itself, then shown the corrupted prompts; the bars are the fraction of questions it still answers. A method that answers as much of the corrupted split as of the clean one has not noticed the intervention.

k=1 and k=2 in every cell, and for k=3 in every cell except closed-book TriviaQA on Llama and Gemma, where the two tie; on CoQA with Llama and Gemma it does so while showing fewer than three meanings on average (2.52 and 2.65). The size the rule chooses per question is therefore what a fixed k cannot provide.

Table 9: **The size, not the ranking** (clean prompts, gold-in-set, per cent). Given our per-question size the ensemble's set is ours on almost every prompt, so those columns agree; against any k fixed in advance the credal set is ahead or level, on CoQA while showing fewer meanings.

			Matched size		Fixed size		
Dataset	Backbone	Our mean size ↓	Ours ↑	Top-k at our size ↑	k=1	k=2	k=3
CoQA	Llama-3.1-8B	2.52	**87**	87	75	83	84
CoQA	Gemma-2-9B	2.65	**90**	90	79	87	89
CoQA	Qwen2.5-7B	5.94	**89**	89	76	82	85
TriviaQA (passage)	Llama-3.1-8B	3.78	**91**	91	85	88	90
TriviaQA (passage)	Gemma-2-9B	2.51	**88**	88	80	85	87
TriviaQA (passage)	Qwen2.5-7B	6.44	**86**	86	75	80	83
TriviaQA (closed book)	Llama-3.1-8B	2.80	**88**	88	78	83	88
TriviaQA (closed book)	Gemma-2-9B	2.92	**86**	86	75	82	86
TriviaQA (closed book)	Qwen2.5-7B	7.16	**76**	76	56	64	70

E.5 Four-Option Setting and Interval Calibration

Why the four-option setting is separate. On multiple choice the answer is one of four letters, so no generation is involved and the token space alone decides. This isolates the commitment rule from credal decoding: any difference on OpenBookQA and ARC-Challenge is attributable to the bounds and the rule, not to the search. ARC-Challenge is answered with the adapters trained on OpenBookQA, which also tests whether the rule survives a shift in the question distribution it was never fine-tuned on.

Why interval calibration is measured in that setting. A point confidence can be checked against accuracy with a reliability diagram, but an interval makes a different promise, namely that it contains

the accuracy. The four-option setting is where that promise can be checked cleanly, because correctness is unambiguous and every question has the same four options, so questions can be binned by the midpoint of the selected answer's interval and the bin's accuracy compared with the bin's mean bounds.

Tab. 10 gives the commit rates behind Fig. 4. On the standard test splits, CTC commits on 73–91% of questions and is correct on 89–98% of those (Fig. 4). The presented sets contain the gold letter on 89–98% of questions at a mean size of 1.14–1.50 letters. On the paired OpenBookQA split, corrupting the passage lowers commitment from 86.8–91.6% to 65.2–70.6% with no parameter tuned, while the committed answers stay 89–95% correct and the sets grow from 1.12–1.24 to 1.77–2.21 letters. Fig. 8 and Tab. 11 test the intervals themselves. Across the six cells, the accuracy of a bin lies inside the bin's mean interval in 24 of the 30 populated bins (13 of 14 on OpenBookQA and 11 of 16 on ARC-Challenge), with no calibration step. The top bin holds 58–88% of questions at an accuracy of 0.94–0.99 with intervals 0.02–0.04 wide, and the six bins outside their interval all lie below the lower bound, most of them on ARC-Challenge, which the adapters were not trained on.

Table 10: **Credal Token Commitment on the four-option setting**, the commit rates behind Fig. 4 (OpenBookQA and ARC-Challenge, N=500, M=5; ARC-Challenge uses the OpenBookQA-trained adapters). Commit = $|\mathcal{D}_1|$=1. The final block is the paired OpenBookQA split, the same 500 questions with and without a corrupted passage, so its clean rows differ slightly from the first block, which uses the standard test split.

Dataset	**Backbone**	Commit rate	Mean $\lvert\mathcal{D}_1\rvert\downarrow$	Gold in set ↑	Acc. at the rule's coverage	
					CTC (ours)	LoRA Ensemble (pred. entropy)
OpenBookQA	Llama-3.1-8B	0.830	1.29	0.966	**0.971**	0.959
OpenBookQA	Gemma-2-9B	0.864	1.30	0.984	0.984	**0.986**
OpenBookQA	Qwen2.5-7B	0.906	1.17	0.956	**0.954**	**0.954**
ARC-Challenge	Llama-3.1-8B	0.734	1.50	0.888	**0.905**	0.888
ARC-Challenge	Gemma-2-9B	0.882	1.20	0.930	**0.934**	0.927
ARC-Challenge	Qwen2.5-7B	0.892	1.14	0.912	**0.917**	0.908
OpenBookQA, the same 500 questions with and without a corrupted passage						
OpenBookQA, clean	Llama-3.1-8B	0.868	1.24	0.952	**0.963**	0.956
OpenBookQA, clean	Gemma-2-9B	0.898	1.22	0.986	**0.986**	0.984
OpenBookQA, clean	Qwen2.5-7B	0.916	1.12	0.962	**0.963**	0.959
OpenBookQA, corrupted context	Llama-3.1-8B	0.652	1.77	0.908	**0.890**	0.880
OpenBookQA, corrupted context	Gemma-2-9B	0.706	2.21	0.954	**0.963**	0.952
OpenBookQA, corrupted context	Qwen2.5-7B	0.670	1.92	0.928	**0.922**	0.904

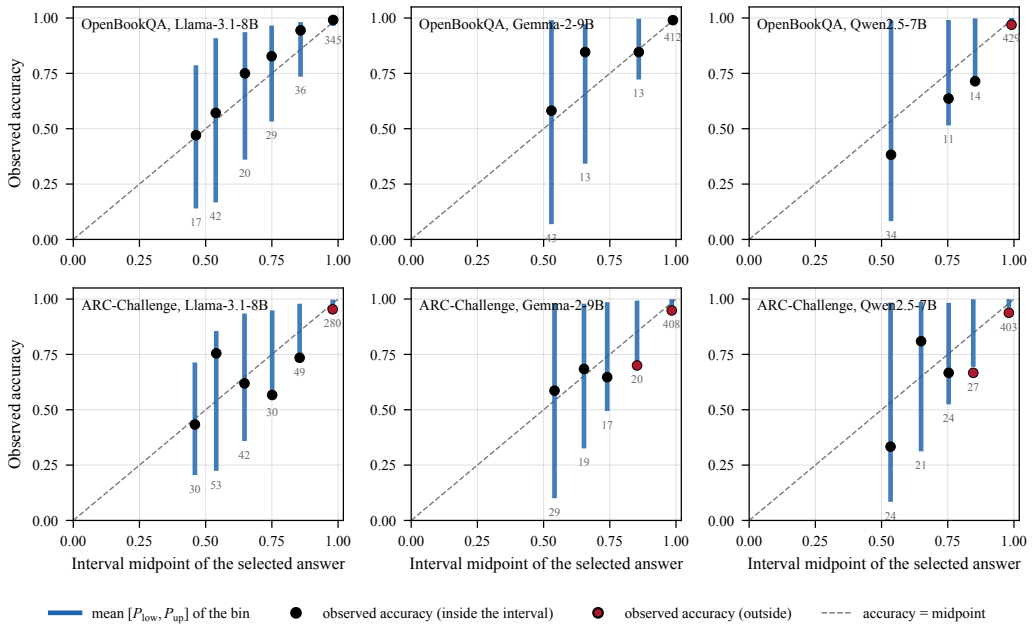


Figure 8: **Interval calibration on the four-option setting.** Questions are binned by the midpoint of the selected answer's interval; the bar is the bin's mean $[\underline{P}, \overline{P}]$, the dot the observed accuracy (red when outside the interval), the number the bin count; bins with fewer than 10 questions are dropped. The table form is Tab. 11.

Table 11: **Interval calibration on the four-option setting**, OpenBookQA and ARC-Challenge. All questions binned by the midpoint of the representative answer's interval (10 bins; bins with fewer than 10 questions dropped). A bin is "inside" when its accuracy lies within the bin's mean $[\underline{P}(a^*), \overline{P}(a^*)]$. The last three columns describe the top bin.

		All populated bins				Top bin		
Dataset	**Backbone**	Bins	Inside ↑	Below $\underline{P}$ ↓	Mean width ↓	Share of questions	Accuracy	Mean $[\underline{P}, \overline{P}]$
OpenBookQA	Llama-3.1-8B	6	6	0	0.447	0.71	0.99	[0.96, 1.00]
OpenBookQA	Gemma-2-9B	4	4	0	0.462	0.86	0.99	[0.98, 1.00]
OpenBookQA	Qwen2.5-7B	4	3	1	0.423	0.88	0.97	[0.98, 1.00]
ARC-Challenge	Llama-3.1-8B	6	5	1	0.399	0.58	0.95	[0.96, 1.00]
ARC-Challenge	Gemma-2-9B	5	3	2	0.468	0.83	0.95	[0.97, 1.00]
ARC-Challenge	Qwen2.5-7B	5	3	2	0.475	0.81	0.94	[0.96, 1.00]

E.6 Semantic Commitment and Agreement between Depths

Why a closed-book setting is included. With a passage in the prompt, much of the ensemble's agreement comes from the passage rather than from the model. Closed-book TriviaQA removes the passage, so the intervals reflect what the adapters know rather than what they were given, and the semantic space can be compared in both regimes.

Tab. 12 reports the semantic space. CSC commits on 30.8–47.6% of clean CoQA questions and on 24.4–45.6% of closed-book TriviaQA questions, with an accuracy when committed of 91–96%, against 56–78% for the representative answer on all questions. Corrupting the passage lowers commitment to 3.6–7.6% on CoQA and to at most 0.8% on TriviaQA, and the unexplored mass rises from 0.36–0.51 to 0.73–0.88 on CoQA and from 0.78–0.80 to 0.86–0.96 on TriviaQA: with the passage replaced, the adapters spread their mass over completions the search cannot reach, and the widened upper bounds block commitment.

Table 12: **Credal Semantic Commitment** on CoQA and TriviaQA, clean and with the passage corrupted, and TriviaQA closed-book (fine partition, unexplored mass carried; M=5, 250 prompts per cell). Percentages except set size (visible meanings) and Unexpl. (mean over adapters of the mass the search did not reach); gold in set: the presented set contains the gold meaning; Acc (rep.): accuracy of the representative meaning ($\arg\max \underline{P}$).

Setting	M	Commit	Acc|commit ↑	Gold in set ↑	Set size ↓	Acc (rep.) ↑	Unexpl.
Llama-3.1-8B							
CoQA, clean	5	40.4	95	87	2.52	75	0.51
CoQA, corrupted context	5	3.6	22	24	3.59	14	0.88
TriviaQA, clean	5	1.2	100	91	3.78	86	0.80
TriviaQA, corrupted context	5	0.0	–	76	4.47	67	0.96
TriviaQA closed-book	5	32.0	96	88	2.80	77	0.58
Gemma-2-9B							
CoQA, clean	5	47.6	96	90	2.65	78	0.42
CoQA, corrupted context	5	3.6	33	28	3.30	12	0.83
TriviaQA, clean	5	1.2	100	88	2.51	83	0.78
TriviaQA, corrupted context	5	0.0	–	76	3.14	66	0.86
TriviaQA closed-book	5	45.6	96	86	2.92	76	0.44
Qwen2.5-7B							
CoQA, clean	5	30.8	91	89	5.94	74	0.36
CoQA, corrupted context	5	7.6	32	26	9.38	13	0.73
TriviaQA, clean	5	6.4	94	86	6.44	76	0.79
TriviaQA, corrupted context	5	0.8	100	73	9.36	56	0.90
TriviaQA closed-book	5	24.4	95	76	7.16	56	0.56

Token-space commitment under corrupted context. Tab. 13 reports the token-space commit rates on clean and corrupted prompts discussed in §4.2 (iii).

In Tab. 13, token-space commitment on CoQA falls from 0.20–0.30 on clean prompts to 0.036–0.052 on corrupted ones, and the size of the token-space set alone separates corrupted from clean CoQA prompts with AUROC 0.78–0.83. On TriviaQA with the passage the token space almost never closes, even on clean prompts (at most 2.4%), which is why the semantic space is needed there. Tab. 14 and Fig. 9 split every question by which depths commit. On CoQA and closed-book TriviaQA both depths commit on 8–36% of questions, the token space alone on 6–15% and the semantic space alone

on 8–30%; under corrupted context neither commits on 91–100% of questions. Tab. 15 compares the two depths as detectors of the model's own wrong answers: on free-form answers the semantic space is ahead in all nine cells, by 2.6 to 24.7 points, and on the four-option datasets the two coincide.

Table 13: **Token-space commitment under corrupted context** (250 clean and 250 corrupted prompts per CoQA and TriviaQA cell, and the 500 paired OpenBookQA questions of Tab. 10; space = whole vocabulary, top-100 tokens by upper bound plus the remaining mass). A prompt is corrupted when its evidential passage is replaced or perturbed. OpenBookQA rows use the first-token caches of Tab. 10; CoQA and TriviaQA rows use the credal-decoding caches of Tab. 12 (M=5). The last column is the corrupted-versus-clean AUROC of the set size.

Dataset	**Backbone**	Commit rate, clean	Commit rate, corrupted	Median $\lvert\mathcal{D}_1\rvert$ clean / corrupted	AUROC of $\lvert\mathcal{D}_1\rvert$
OpenBookQA	Gemma-2-9B	0.898	0.706	1 / 1	0.881
OpenBookQA	Llama-3.1-8B	0.868	0.652	1 / 1	0.732
OpenBookQA	Qwen2.5-7B	0.916	0.670	1 / 1	0.805
CoQA	Gemma-2-9B	0.236	0.044	2 / 13	0.830
CoQA	Llama-3.1-8B	0.304	0.052	2 / 10	0.832
CoQA	Qwen2.5-7B	0.204	0.036	3 / 20	0.777
TriviaQA	Gemma-2-9B	0.008	0.000	75 / >1000	0.729
TriviaQA	Llama-3.1-8B	0.000	0.000	415 / >1000	0.629
TriviaQA	Qwen2.5-7B	0.024	0.000	89 / >1000	0.731

Table 14: Agreement between depths on CoQA and TriviaQA, with the passage and closed-book (percent of questions): both CTC and CSC commit; only the token space; only the semantic space; neither.

Setting	Both commit	Token space only	Semantic space only	Neither
Llama-3.1-8B				
CoQA, clean	20	10	20	50
CoQA, corrupted context	3	2	0	94
TriviaQA, clean	0	0	1	99
TriviaQA, corrupted context	0	0	0	100
TriviaQA closed-book	24	15	8	53
Gemma-2-9B				
CoQA, clean	18	6	30	47
CoQA, corrupted context	2	2	2	94
TriviaQA, clean	0	1	1	98
TriviaQA, corrupted context	0	0	0	100
TriviaQA closed-book	36	11	10	44
Qwen2.5-7B				
CoQA, clean	11	9	20	60
CoQA, corrupted context	2	2	6	91
TriviaQA, clean	0	2	6	91
TriviaQA, corrupted context	0	0	1	99
TriviaQA closed-book	8	8	16	68

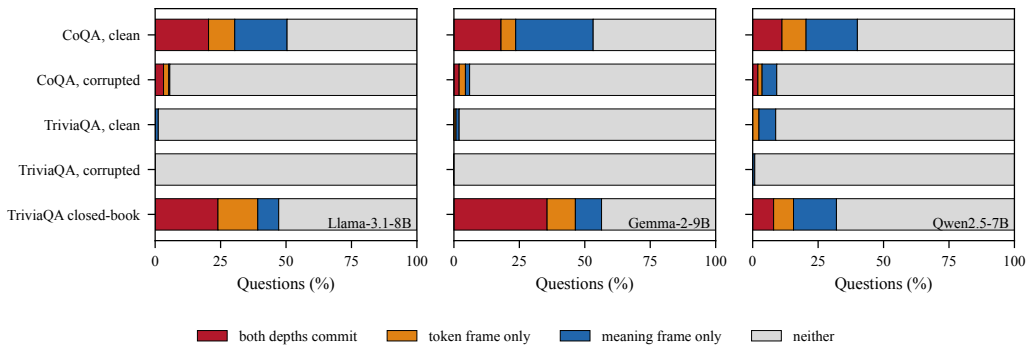


Figure 9: Agreement between the two depths, per cell (the data of Tab. 14).

Table 15: **The same rule at both depths, on every dataset** (AUROC of each depth's score against the model's own answer being wrong, higher is better; CoQA, TriviaQA with and without the passage, OpenBookQA and ARC-Challenge). The main tables report one CLLM row, whose depth is fixed by the answer type rather than chosen per cell: the semantic space on free-form answers, the token space on the two four-option datasets, where a one-letter answer is its own meaning and the two depths are the same quantity. This table is the check on that rule. On free-form answers the semantic space is ahead in all nine cells, by 3 to 25 points. **Bold** = better of the two depths per column; grey rows are ours.

		CoQA			TriviaQA (passage)			TriviaQA (closed book)			OpenBookQA			ARC-Challenge		
Depth (AUROC ↑)	Candidates	Llama	Gemma	Qwen2.5	Llama	Gemma	Qwen2.5	Llama	Gemma	Qwen2.5	Llama	Gemma	Qwen2.5	Llama	Gemma	Qwen2.5
CLLM (CTC)	token space, first answer token	0.754	0.752	0.714	0.711	0.744	0.755	0.764	0.706	0.536	**0.917**	**0.942**	**0.917**	**0.781**	**0.830**	**0.807**
CLLM (CSC)	semantic space, meanings	**0.830**	**0.844**	**0.740**	**0.822**	**0.850**	**0.867**	**0.833**	**0.833**	**0.783**	**0.917**	**0.942**	**0.917**	**0.781**	**0.830**	**0.807**

E.7 Ablations

Why the ablations are on a development split. The expansion rule, the number of adapters and the meaning grouping are choices inside the method, so they are varied on a frozen CoQA development split that is disjoint from every test set, and the test cells are never used to select them. The adapter sweep uses nested subsets, so $M{=}2$ is a subset of $M{=}3$ and so on, which makes the comparison a statement about ensemble size rather than about which adapters happened to be drawn.

E.7.1 Number of Adapters

Proposition A.4 implies that the commitment sets grow with M. Tab. 16 runs the full pipeline on the CoQA development split (500 prompts, Llama) with nested subsets of the same adapters. Gold-in-set and accuracy when committed rise and commitment falls from $M{=}2$ to 4 and are flat from 4 to 5; the token-space commit rate falls monotonically (39, 31, 28, 27%) as the intervals widen. Fig. 10 runs the same sweep on all three backbones: token-space commitment falls with M on every backbone, and gold-in-set rises from $M{=}2$ to $M{=}4$ on Llama and Qwen and stays within one point on Gemma.

Table 16: Number of adapters (CoQA development split, 500 prompts, Llama-3.1-8B, sequence-level expansion, fine partition). Percentages except set size (visible meanings) and unexplored mass.

M	CTC commit	CSC commit	Acc|commit	Gold in set	Set size	Unexpl.
2	38.6	43.4	90	84	1.79	0.50
3	31.4	41.6	91	88	2.08	0.50
4	28.2	40.6	93	90	2.33	0.51
5	26.6	40.8	92	89	2.34	0.51

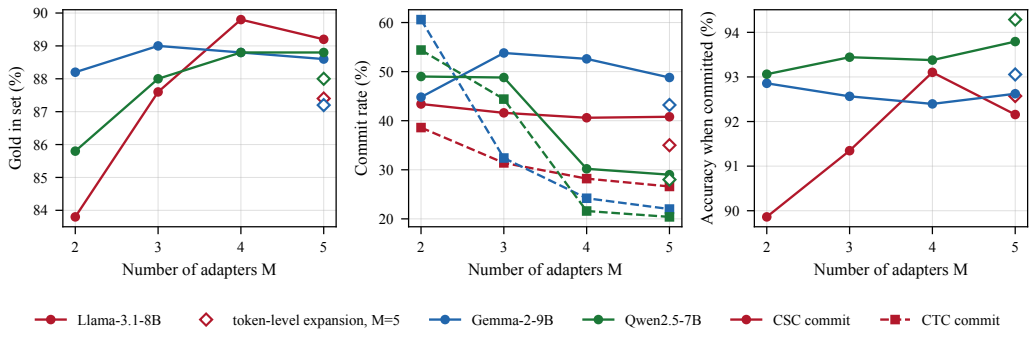


Figure 10: **Number of adapters and expansion rule** (CoQA development split, 500 prompts, nested subsets $M{=}2, 3, 4, 5$; hollow marker: token-level expansion at $M{=}5$). Token-space commitment falls with M on all three backbones while gold-in-set rises or holds.

E.7.2 Expansion Rule, Width and Unexplored Mass

Expansion rule. On the CoQA development split with the same five adapters, sequence-level expansion (prune a partial sequence when its upper mass falls below the best completed lower mass, §3) against token-level expansion (keep only undominated next tokens at every step): CSC commit

40.8 against 35.0%, accuracy when committed 92 against 93%, gold in set 89 against 87%, visible meanings 2.34 against 2.52, unexplored mass 0.51 against 0.55, completed sequences per prompt 10.9 against 14.9, decoding steps per prompt 72 against 84. Following the decision keeps fewer, more probable sequences and reaches more of the licensed mass. *Width.* With three adapters and token-level expansion on the same split, gold was among the candidates on 82.6, 84.0 and 85.2% of prompts at B=4, 8 and 16, with the width reached on 62, 47 and 36% of prompts and 36, 70 and 129 decoding steps per prompt; B=8 was fixed from this sweep before any test data was decoded. With five adapters and sequence-level expansion the width is reached on every prompt, since the top-B partial sequences by upper mass are always kept, and the mass beyond them is carried as u_m. *Carrying the unexplored mass.* Reading the same clean CoQA test prompts without u_m in the upper bound and without the second commitment condition, that is treating the written-out completions as the whole sample space, raises the commit rate from 40.4, 47.6 and 30.8% to 66.0, 60.8 and 38.0% on Llama, Gemma and Qwen and lowers accuracy when committed from 95, 96 and 91% to 90, 91 and 88%: the mass the search never reached is doing real work in the decision. Tab. 17 repeats the expansion comparison on all three backbones: expanding by the decision raises CSC commitment on every backbone, by 5.8, 5.6 and 1.0 points on Llama, Gemma and Qwen, and gold-in-set by 1.8, 1.4 and 0.8 points, with accuracy when committed within 0.5 points of token-level expansion.

Table 17: **Expansion rule** (CoQA development split, 500 prompts, M=5, B=8). *Decision*: prune a partial sequence when its upper mass falls below the best completed lower mass. *Token*: keep only undominated next tokens. Commitment and gold-in-set improve everywhere; the step count falls on two backbones of three.

Backbone	Expansion	CSC commit	Acc. when committed	Gold in set	Steps/prompt	Completed/prompt
Llama-3.1-8B	decision	40.8	92.2	89.2	72.4	10.9
Llama-3.1-8B	token	35.0	92.6	87.4	83.7	14.9
Gemma-2-9B	decision	48.8	92.6	88.6	73.0	23.9
Gemma-2-9B	token	43.2	93.1	87.2	65.1	27.2
Qwen2.5-7B	decision	29.0	93.8	88.8	97.5	14.9
Qwen2.5-7B	token	28.0	94.3	88.0	106.2	17.4

E.7.3 Meaning Grouping

Grouping the same CoQA development candidates (Llama-3.1-8B, M=5, 500 prompts) with an embedding model (BAAI/bge-base-en-v1.5, cosine) gives 1.46 meanings per prompt at threshold 0.5 and 3.22 at 0.8, with a gold-matching and a non-matching candidate in the same meaning on 66% and 31% of prompts; the adapters' own equivalence judgement gives 2.83 meanings at 38% without a threshold, and inspection of a random sample of 40 of the flagged prompts finds 31 to be answers the single gold string does not match ("Five" against "5", "He was." against "Yes", "in excess of 3,500" against "over 3,500") and 7 to be genuine merges of answers at different granularity ("a friend of his" with "a man named Stephens"), the remaining two being too long to classify. The "possibly equivalent" partition (any adapter says yes) merges far more aggressively (1.32 meanings at 62% impurity) and is reported only as a diagnostic.

The two meaning partitions. Sec. 3.4 defines two partitions of the candidates: the *surely* equivalent one, where every adapter judges a pair alike, which is what every reported number uses, and the *possibly* equivalent one, where a single adapter suffices. Tab. 18 contrasts them on the clean test prompts. The possible partition collapses the candidates into one or two meanings on Llama and Gemma and into fewer meanings on Qwen (3.94 and 1.78), and the price is purity: the fraction of prompts on which a meaning holds both a gold-matching and a non-matching candidate rises from 19–60% to 40–88%, reaching 88% on Llama with TriviaQA. The gap between the two columns is the imprecision in meaning itself for this ensemble, and it is why the decision uses the sure partition: a rule that commits on a meaning built from one adapter's judgement would be committing to a class its own members do not agree on.

E.8 Cost-Sensitive Commitment

The cost-sensitive rule is the Γ-maximin decision of imprecise probability (Troffaes, 2007): choose the action whose lower expected utility is largest. With utility 1 for a correct answer, $-\kappa$ for a wrong one and 0 for abstaining, committing to a^* has lower expected utility $\min_{q\in\mathcal{P}}\left[q(a^*) - \kappa\,(1 - q(a^*))\right] =$

Table 18: **Sure against possible meaning partitions** on clean test prompts (250 per cell). Meanings: the mean number of meanings per prompt. Impurity: the percentage of prompts on which one meaning contains both a gold-matching and a non-matching candidate. Every result in the paper uses the sure partition.

		Meanings per prompt		Impurity	
Dataset	Backbone	sure	possible	sure	possible
CoQA	Llama-3.1-8B	2.98	1.41	30%	62%
CoQA	Gemma-2-9B	3.32	1.42	30%	64%
CoQA	Qwen2.5-7B	6.88	3.94	19%	40%
TriviaQA	Llama-3.1-8B	3.80	1.01	60%	88%
TriviaQA	Gemma-2-9B	2.52	1.00	38%	66%
TriviaQA	Qwen2.5-7B	6.62	1.78	33%	62%

$\underline{P}(a^*)(1+\kappa)-\kappa$, so it beats abstaining exactly when $\underline{P}(a^*) \geq \kappa/(1+\kappa)$. Tab. 19 sweeps the declared cost κ of §3.4, where the model commits to the representative meaning when $\underline{P}(c^*) \geq \kappa/(1+\kappa)$. Raising κ trades coverage for accuracy: on CoQA, Llama answers 54% of questions at 93% accuracy when a wrong answer costs half a right one, and 16% at 100% accuracy when it costs four times as much. The more interesting column is the first. The rule as stated, which has no κ and no threshold, lands on the same operating point as the cost rule at $\kappa = 1$, and not merely in aggregate: across the nine cells the two agree on 98.4 to 100% of individual prompts, and in the 9 prompts of 2250 where they differ it is always the rule as stated committing where the cost rule would not. The parameter-free rule therefore behaves as if a wrong answer costs exactly what a right answer is worth, which is the neutral choice a system makes when it has not been told the costs, and κ is available when a deployment knows better.

Table 19: **Cost-sensitive commitment** on clean CoQA and TriviaQA prompts (250 per cell). Cov.: fraction of questions answered; Acc.: accuracy of the committed answer. "Rule as stated" is the parameter-free rule of Eq. (3) with the unexplored-mass condition; the remaining columns commit when $\underline{P}(c^*) \geq \kappa/(1+\kappa)$. A dash marks a column that commits on no question in that cell.

		Rule as stated		$\kappa = 0.5$		$\kappa = 1$		$\kappa = 2$		$\kappa = 4$	
Dataset	Backbone	Cov.	Acc. ↑	Cov.	Acc. ↑	Cov.	Acc. ↑	Cov.	Acc. ↑	Cov.	Acc. ↑
CoQA	Llama-3.1-8B	0.40	0.95	0.54	0.93	0.40	0.95	0.28	0.97	0.16	1.00
CoQA	Gemma-2-9B	0.48	0.96	0.56	0.94	0.47	0.96	0.31	0.99	0.18	1.00
CoQA	Qwen2.5-7B	0.31	0.91	0.37	0.90	0.31	0.91	0.22	0.98	0.14	1.00
TriviaQA	Llama-3.1-8B	0.01	1.00	0.07	1.00	0.01	1.00	0.00	–	0.00	–
TriviaQA	Gemma-2-9B	0.01	1.00	0.17	0.98	0.01	1.00	0.00	–	0.00	–
TriviaQA	Qwen2.5-7B	0.06	0.94	0.20	0.98	0.06	0.94	0.00	1.00	0.00	–
TriviaQA closed-book	Llama-3.1-8B	0.32	0.96	0.44	0.94	0.32	0.96	0.23	0.98	0.17	1.00
TriviaQA closed-book	Gemma-2-9B	0.46	0.96	0.50	0.96	0.45	0.97	0.36	0.97	0.26	0.98
TriviaQA closed-book	Qwen2.5-7B	0.24	0.95	0.29	0.94	0.23	0.95	0.15	0.95	0.11	0.96

E.9 Computational Cost

Measured on the runs behind Tab. 12 (M=5, B=8, L=32, KV-cached decoding, adapters swapped in one process). Credal decoding (which yields CTC, the candidate set and every adapter's sequence probabilities in one pass) takes 7–12 s per prompt on CoQA and closed-book TriviaQA, 11–18 s on TriviaQA with the passage and 3–5 s on the eight-token OpenBookQA setting; the equivalence judgements add 5–12 s per prompt on CoQA and closed-book TriviaQA and 8–24 s on TriviaQA with the passage, where the candidate set is largest. Times are for Llama-3.1-8B on one L4 (24 GB) and Gemma-2-9B on one A100 (40 GB) and are similar on the two. Peak memory is 21 GB for Llama, 32 GB for Gemma and 16 GB for Qwen2.5-7B, dominated by the base model and the KV caches of B partial sequences. Semantic entropy for a single model (K=16 samples, L=32) takes 2–3 s per prompt on the same hardware, and the mean-pooled ensemble costs exactly the decoding pass above with the same adapters. Decoding steps per prompt are 16–25 on OpenBookQA, 68–89 on CoQA and closed-book TriviaQA and 135–217 on TriviaQA with the passage, and forward passes per prompt are M times these. Adapter training is a one-off cost of about one GPU-hour per adapter and is not part of inference.

F BROADER IMPACT

This work concerns reliability tooling for instruction-tuned Large Language Models, specifically a commitment rule that answers only when no plausible fine-tuning of the model would prefer a different answer, and otherwise returns the set of answers that remain plausible. The most likely positive consequence of CLLM-style commitment is reducing fluent-but-wrong outputs in deployment-critical settings such as healthcare, legal, scientific, and educational assistants, where confidently-wrong answers carry larger downstream cost than abstentions. The credal-set view exposes the spread of the credal set rather than collapsing it; reviewers, regulators, and downstream users can read off both *what* the model predicts and *how stable* the prediction is across plausible adapters.

We see two material risks. First, a committed answer can be misread as a correct one: commitment means that every plausible fine-tuning in the realised ensemble agrees, not that the agreement is right. Inputs on which all adapters share a mistake (for example poisoned context they all absorbed the same way) are committed with high confidence, and no rule defined on the ensemble can detect a failure the whole ensemble shares. Reliance on CLLM as a sole correctness gate, without complementary retrieval or human review, would inherit this blind spot.

Second, an ensemble of $M{=}5$ adapters raises both training and inference cost relative to a single fine-tune; §E.9 reports the overhead. A forward pass with a LoRA adapter costs roughly $1.1\times$ a base forward pass, so the total inference cost scales with M. In carbon-conscious deployments the trade-off should be made explicit: CTC requires no generation and costs one M-adapter forward pass, while CSC additionally requires credal decoding and the equivalence judgements. We recommend reporting these costs alongside commit rate and accuracy when CLLM is integrated into a downstream system.

The protocols draw on existing public datasets (OpenBookQA, CoQA, TriviaQA, ARC-Challenge), and we do not collect or release new data.