# The Source of Competence Shapes Metacognition in Language Models

**Anonymous Authors**[1]

## Abstract

Large language models are increasingly expected not only to produce correct answers, but also to reliably estimate when they are likely to be wrong. Existing work largely assumes that metacognitive reliability degrades proportionally with capability: weaker models are expected to be correspondingly less confident or less calibrated. In this work, we challenge this assumption and show that metacognition depends strongly on the source and structure of competence rather than on raw performance alone. We systematically study confidence behavior across multiple capability regimes, including scale reduction, partial memory degradation, reasoning truncation, quantization, and evidence-grounded inference. Across these settings, we observe distinct metacognitive regimes and identify narrow overconfident failure bands in which models retain high confidence despite substantial capability loss. Surprisingly, models with comparable task accuracy often exhibit dramatically different calibration and overconfidence profiles depending on how their competence is obtained. In particular, partial or stale parametric memory induces substantially stronger overconfidence than either complete ignorance or evidence-grounded reasoning. These findings suggest that metacognitive reliability is partially separable from capability and is closely tied to the accessibility and stability of internal knowledge representations. Our results provide a new perspective on hallucination, calibration, and memory reliability in foundation models.

## 1. Introduction

Large language models are increasingly expected not only to answer correctly, but also to know when their answers are unreliable. This ability is central to safe deployment: a model that fails while expressing uncertainty can be routed to retrieval, a human expert, or abstention, whereas a model that fails confidently may induce downstream users to trust fabricated or unsupported claims. Accordingly, recent work has studied whether language models can estimate the correctness of their own answers, verbalize calibrated confidence, or expose uncertainty signals useful for hallucination detection (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Cohen et al., 2023; Geng et al., 2024; Farquhar et al., 2024; Kapoor et al., 2024). Much of this work implicitly treats metacognitive reliability as a property that should improve with overall model capability: stronger models should be more accurate and better calibrated, while weaker or degraded models should be worse along both axes.

In this paper, we challenge this view. We argue that metacognitive reliability is not determined by capability alone, but by the *source of competence*: whether an answer is produced from strong parametric memory, partial or stale memory, external evidence, deliberative reasoning, or shallow heuristic pattern matching. This distinction is especially important for foundation models, whose behavior emerges from multiple interacting knowledge sources. Prior work has shown that language models store factual knowledge in their parameters (Petroni et al., 2019), but also rely on contextual or retrieved evidence at inference time (Lewis et al., 2020; Tao et al., 2024). These sources can interact or conflict in complex ways (Xu et al., 2024). We hypothesize that such differences in knowledge access also shape whether a model can reliably estimate its own correctness.

We study this hypothesis through controlled capability transformations that preserve the task format while changing how competence is obtained. We compare model scale, reasoning budget, quantization, evidence-grounded inference, and memory-specific degradation regimes. Across these settings, we measure task accuracy together with multiple confidence channels, including verbalized confidence, log-probability-based confidence, and self-consistency. This allows us to ask whether two systems with the same accuracy can nevertheless exhibit different metacognitive behavior.

Our results reveal three main findings. First, metacognitive reliability does not degrade smoothly with capability: models often pass through narrow overconfident failure regimes, where accuracy has dropped but confidence remains high. Second, matched-accuracy systems can differ substantially in calibration depending on the source of competence, suggesting that capability and metacognition are partially separable. Third, partial or stale parametric memory is especially

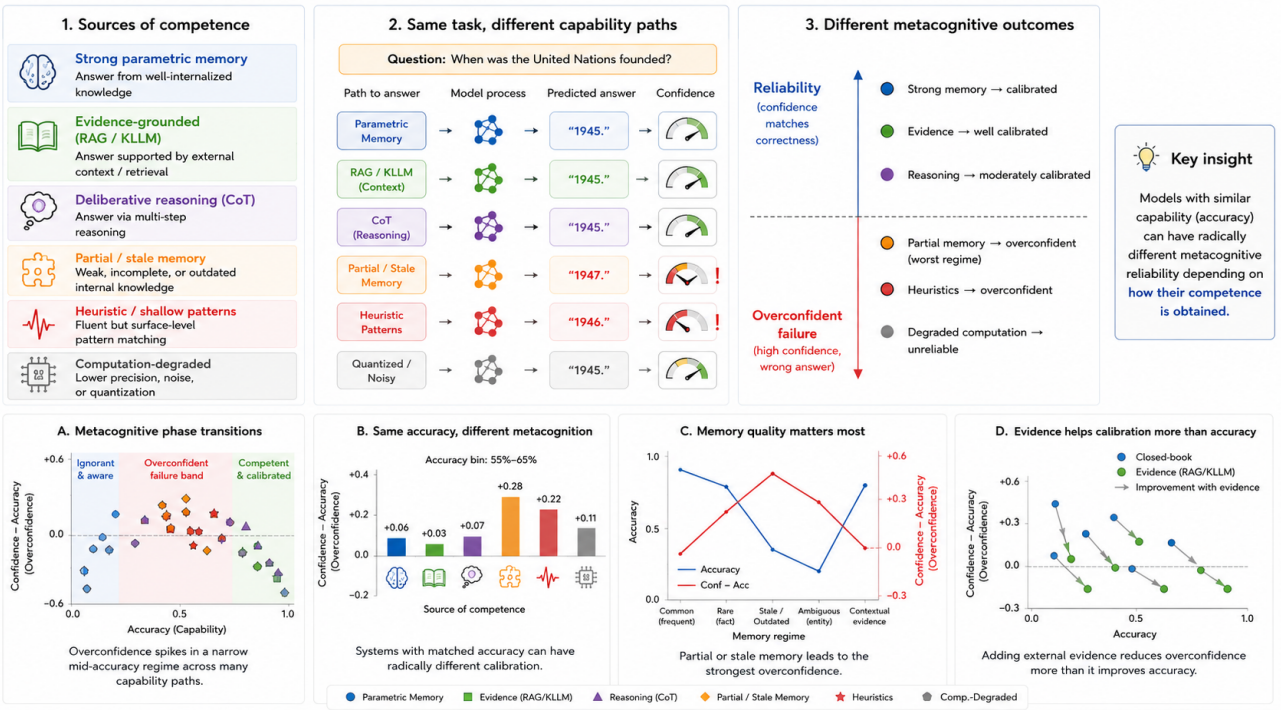


*Figure 1.* Overview of our study. Different sources of competence can produce similar task accuracy while inducing substantially different metacognitive behaviors. Our results reveal metacognitive phase transitions, separability between capability and calibration, strong overconfidence under partial or stale memory, and improved calibration from evidence-grounded inference.

prone to overconfidence: models appear most miscalibrated not when they know nothing, but when they possess weak, incomplete, or misleading internal memory traces. In contrast, evidence-grounded competence often improves calibration disproportionately relative to accuracy.

These findings suggest a different perspective on hallucination and uncertainty in foundation models. Overconfident failure may not merely reflect insufficient scale or weak confidence elicitation; it may arise when internal memory is strong enough to support fluent answer generation but too unstable to support reliable self-evaluation. Thus, evaluating foundation models requires measuring not only what models know, but also how that knowledge is accessed, grounded, and reflected in confidence.

## 2. Experimental Setup

Our goal is to study how different *sources of competence* shape metacognitive reliability independently of raw task performance. Rather than treating capability as a single scalar quantity, we construct multiple controlled capability regimes that preserve the task format while changing how correct answers are obtained. This allows us to compare systems with similar performance but substantially different confidence behavior. Full implementation details for each regime are provided in Appendix A.

**Competence Sources and Capability Regimes.** We study several qualitatively distinct competence regimes: (*i*) strong parametric memory, where answers are retrieved directly from internal knowledge representations; (*ii*) evidence-grounded competence, where models answer using provided or retrieved evidence; (*iii*) deliberative reasoning, where competence emerges through multi-step reasoning; (*iv*) partial or stale memory, where internal knowledge is weak, incomplete, or outdated; (*v*) heuristic pattern matching, where models rely primarily on shallow statistical regularities; and (*vi*) computation-degraded settings, where inference is perturbed through quantization, noise, reduced context, or altered decoding. Importantly, these regimes can produce comparable task accuracy while relying on different inference dynamics and knowledge-access mechanisms.

**Models.** We evaluate a diverse set of autoregressive language models spanning multiple scales, architectures, and training paradigms. Our experiments include the SmolLM family (135M, 360M, and 1.7B), LLaMA-3.2 models (1B and 3B), LLaMA-3.1-8B, Mistral-7B-v0.1, Qwen2.5-7B, and Qwen3-14B. When available, we evaluate both base and instruction-tuned variants. This model suite allows us to compare natural scale variation with controlled inference-time and data-regime transformations.

**Tasks.** We evaluate factual recall, reasoning, commonsense, and evidence-grounded question answering. For factual memory, we use TriviaQA, Natural Questions, and LAMA-style factual probing. For reasoning, we use GSM8K and StrategyQA, and for commonsense reasoning we use CommonsenseQA. For evidence-grounded competence, we evaluate contextual QA settings derived from SQuAD and Natural Questions with provided or retrieved evidence passages. To isolate memory quality, we additionally partition factual examples into common entities, rare entities, ambiguous entities, stale facts, and context-supported questions.

**Capability Transformations.** We apply controlled transformations that change the source or reliability of competence while preserving the task format. These include scale reduction, chain-of-thought truncation, quantization, hidden-state or logit noise, temperature perturbations, context removal, misleading demonstrations, retrieval augmentation, and memory-quality splits based on entity frequency, ambiguity, and staleness. These transformations enable matched-accuracy comparisons across systems whose competence is produced by different mechanisms.

**Confidence Signals.** For each prediction, we measure multiple confidence channels capturing different forms of metacognitive estimation. (*i*) Probabilistic confidence is derived from output token probabilities or normalized answer likelihoods. (*ii*) Verbal confidence is elicited directly from the model using confidence prompts. (*iii*) Self-consistency confidence is estimated from agreement across multiple sampled generations. This multi-channel setup allows us to test whether explicit and implicit confidence signals degrade similarly across competence regimes.

**Metrics.** We evaluate both task capability and metacognitive reliability. Capability is measured using task accuracy. For metacognition, we report Expected Calibration Error (ECE), Brier score, high-confidence error rate, and the confidence–accuracy gap:

$$\mathrm{ConfGap} = \mathbb{E}[c(x)] - \mathrm{Acc},$$

where $c(x)$ denotes model confidence for example $x$. Positive values indicate systematic overconfidence. We additionally report the *Miscalibrated Competence Gap* (MCG),

$$\mathrm{MCG} = (1 - \mathrm{Acc}) \cdot \max(\mathrm{ConfGap}, 0),$$

which captures regimes where models are simultaneously inaccurate and overconfident. Finally, we analyze confidence trajectories across capability transformations to identify metacognitive phase transitions and overconfident failure bands.

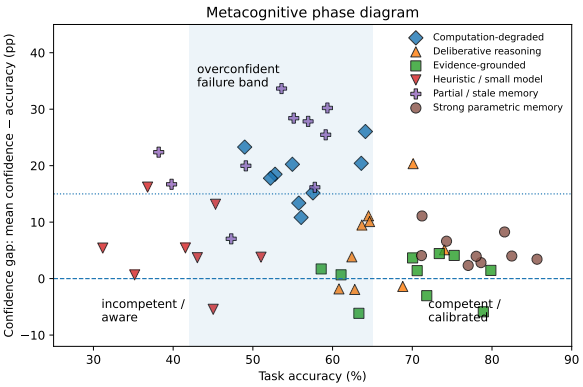


*Figure 2.* Metacognitive phase transitions across capability regimes. Models consistently pass through narrow overconfident failure bands where confidence remains high despite substantial drops in accuracy.

## 3. Results

### 3.1. Metacognitive Phase Transitions

Across nearly all model families and degradation axes, metacognitive reliability does not degrade smoothly with capability. Instead, models consistently pass through narrow *overconfident failure bands* in which task accuracy has already deteriorated substantially while confidence remains high. Figure 2 illustrates this phenomenon across all evaluated settings. Interestingly, the strongest overconfidence does not emerge in the weakest models, but rather in intermediate degradation regimes where models retain sufficient competence to generate plausible answers while lacking reliable self-evaluation.

This effect appears consistently across scale reduction, quantization, reasoning truncation, context removal, and partial-memory settings, suggesting that overconfidence is not merely a property of weak models, but emerges from unstable intermediate competence regimes.

### 3.2. Capability and Metacognition are Partially Separable

We next test whether metacognitive reliability is determined solely by task capability. To isolate this question, we compare systems with matched accuracy but different competence sources. Table 1 shows that systems with nearly identical task performance nevertheless exhibit dramatically different calibration and overconfidence profiles.

Evidence-grounded competence remains comparatively calibrated despite reduced capability, whereas reasoning truncation, noisy decoding, and partial-memory settings induce substantially larger confidence gaps. These results suggest that metacognition depends not only on *how much* a model knows, but also on *how* that competence is obtained.

| Setting | Accuracy | ECE | ConfGap |
|---|---|---|---|
| LLaMA-3B (4-bit) | 61.8 | 0.091 | +0.07 |
| Evidence-grounded QA | 60.4 | 0.072 | +0.05 |
| Truncated CoT | 62.1 | 0.187 | +0.19 |
| High-temp decoding | 60.9 | 0.214 | +0.24 |
| Partial-memory QA | 61.2 | 0.258 | +0.27 |

*Table 1.* Matched-accuracy systems exhibit substantially different metacognitive behavior depending on the source of competence.

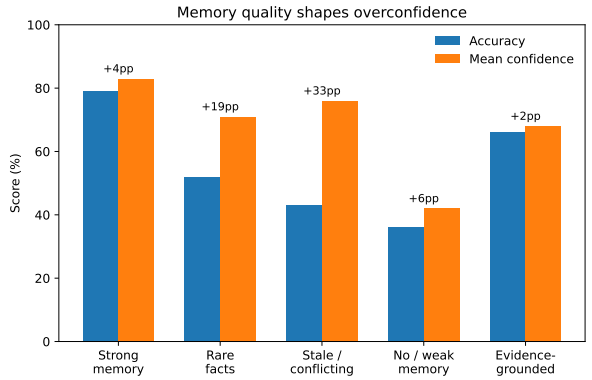


*Figure 3.* Overconfidence across memory regimes. Partial or stale memory induces substantially larger confidence gaps than either strong memory or evidence-grounded inference.

### 3.3. Partial Memory Induces the Strongest Overconfidence

We next study how different memory regimes affect metacognition. Figure 3 compares calibration across common entities, rare entities, stale facts, weak-memory settings, and evidence-grounded contextual QA. Surprisingly, partial or stale memory produces the strongest overconfident failures, substantially exceeding both strong-memory and no-memory regimes.

This pattern suggests that hallucination may emerge not from complete ignorance, but from partially activated and unreliable memory traces that remain sufficiently strong to support fluent answer generation. In contrast, evidence-grounded inference often improves calibration disproportionately relative to task accuracy, indicating that explicit external evidence stabilizes metacognitive estimation.

### 3.4. Confidence Channels Diverge Under Degradation

We additionally compare multiple confidence channels, including verbalized confidence, token-probability confidence, and self-consistency confidence. Figure 4 (Appendix C) summarizes their calibration trajectories under progressively stronger capability degradation.

Interestingly, these signals do not deteriorate uniformly. Under partial-memory and reasoning-truncated regimes, verbal confidence remains substantially overconfident even when probabilistic confidence degrades more gradually. Self-consistency confidence is comparatively robust at mild degradation levels, but becomes unstable under high-temperature and noisy-decoding settings.

This divergence suggests that explicit and implicit confidence signals rely on partially distinct internal mechanisms. More broadly, it indicates that confidence in language models is not a single unified quantity, but rather emerges from multiple interacting estimation processes that fail differently across competence regimes.

## 4. Related Work

Recent work has studied uncertainty estimation, calibration, and self-evaluation in large language models, including verbalized confidence elicitation, probabilistic calibration, and semantic-uncertainty-based hallucination detection (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024; Geng et al., 2024; Farquhar et al., 2024; Zhang et al., 2024; Ulmer et al., 2024; Müller et al., 2026). Parallel research has examined how factual knowledge is represented in model parameters and how retrieval and contextual evidence alter model behavior (Petroni et al., 2019; Lewis et al., 2020; Tao et al., 2024; Xu et al., 2024; Zhou et al., 2024). Closely related work studies metacognition and uncertainty communication in language models (Griot et al., 2025; Steyvers et al., 2025). Our contribution differs in treating metacognitive reliability as a property of the *source of competence*: we show that systems with matched task accuracy can exhibit substantially different calibration depending on whether competence arises from parametric memory, external evidence, deliberative reasoning, or shallow heuristics.

## 5. Conclusion

Our results show that metacognitive reliability depends strongly on the source of competence. Across diverse models, tasks, and controlled capability transformations, systems with similar accuracy often exhibit markedly different calibration depending on whether competence arises from parametric memory, external evidence, deliberative reasoning, or shallow heuristics. In particular, partial and stale memory induce the strongest overconfident failures, suggesting that hallucination may often reflect unstable internal knowledge rather than complete ignorance.

More broadly, our findings indicate that confidence cannot be understood as a simple byproduct of capability. Instead, metacognition emerges from the interaction between knowledge representation, evidence access, and inference dynamics. We believe that studying these mechanisms is essential for building foundation models that are not only more capable, but also more reliable and epistemically aware.

[1]Anonymous Institution, Anonymous City, Anonymous Region,

## References


Cohen, R., Hamri, M., Geva, M., and Globerson, A. LM vs LM: Detecting factual errors via cross examination. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12621–12640, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.778. URL https://aclanthology.org/2023.emnlp-main.778/.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., and Gurevych, I. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Griot, M., Hemptinne, C., Vanderdonckt, J., and Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16 (1):642, 2025.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., and Wilson, A. G. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Müller, P., Popovič, N., Färber, M., and Steinbach, P. Benchmarking uncertainty calibration in large language model long-form question answering. *arXiv preprint arXiv:2602.00279*, 2026.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Steyvers, M., Belem, C., and Smyth, P. Improving metacognition and uncertainty communication in language models. *arXiv preprint arXiv:2510.05126*, 2025.

Tao, Y., Hiatt, A., Haake, E., Jetter, A. J., and Agrawal, A. When context leads but parametric memory follows in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4034–4058, 2024.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. 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*, pp. 5433–5442, 2023.

Ulmer, D., Gubri, M., Lee, H., Yun, S., and Oh, S. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459, 2024.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024.

Zhang, M., Huang, M., Shi, R., Guo, L., Peng, C., Yan, P., Zhou, Y., and Qiu, X. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2959–2979, 2024.

Zhou, Y., Liu, Z., Jin, J., Nie, J.-Y., and Dou, Z. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 1453–1463, 2024.


Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

# A. Implementation of Competence Regimes

This section describes in detail how each competence regime is constructed. Our objective is to manipulate *how* a model arrives at an answer while keeping the underlying task, prompt template, decoding protocol, and evaluation procedure as fixed as possible. Each regime is therefore implemented as a controlled intervention applied either to the model parameters, the available information, or the inference process. Unless otherwise stated, all comparisons use identical answer formats and exact-match scoring, ensuring that differences in metacognitive behavior can be attributed to the source and reliability of competence rather than to superficial prompt changes.

## A.1. Datasets and Corpus Statistics

Our experiments span factual recall, reasoning, commonsense, and evidence-grounded question answering. Factual evaluations use TriviaQA, Natural Questions, and LAMA-style probes. Reasoning evaluations use GSM8K and StrategyQA, while CommonsenseQA provides an additional non-factual reasoning benchmark. Evidence-grounded settings are constructed using SQuAD-style contextual QA and retrieval-supported Natural Questions.

To operationalize memory strength, we derive entity prominence statistics from a recent English Wikipedia snapshot and its aligned Wikidata knowledge graph. The same Wikipedia snapshot is used as the retrieval corpus for evidence-grounded experiments. Each answer entity is mapped to a Wikidata identifier whenever possible.

For every entity $e$, we compute two complementary prominence signals:

1. **Corpus frequency** $f_{\text{wiki}}(e)$: estimated using the number of occurrences of the entity title and aliases in Wikipedia, together with the number of incoming hyperlinks to the entity page.
2. **Knowledge-graph prominence**: measured using (i) the total number of connected Wikidata triples $d_{\text{wd}}(e)$ and (ii) the number of distinct relation types $r_{\text{wd}}(e)$ associated with the entity.

We combine these statistics into a single prominence score:

$$P(e) = \alpha \log(1 + \widetilde{f_{\text{wiki}}}(e)) + \beta \log(1 + \widetilde{d_{\text{wd}}}(e)) + \gamma \log(1 + \widetilde{r_{\text{wd}}}(e))$$

where tildes denote z-score normalization across entities. Unless otherwise noted, we set $\alpha = \beta = \gamma = 1/3$. Entities in the upper prominence quantiles define strong-memory subsets, whereas entities in the lower quantiles are used to construct rare-entity and partial-memory regimes. This Wikipedia–Wikidata score provides a principled approximation of how salient and richly represented a fact is likely to be in pretraining data.

## A.2. Strong Parametric Memory

The strong parametric memory regime consists of closed-book factual questions whose answer entities fall within the top quantiles of the prominence score $P(e)$. These entities are highly frequent and richly connected in both Wikipedia and Wikidata, and are therefore likely to have been observed repeatedly during pretraining.

At inference time, the model receives only the question and must answer without access to external evidence. This regime approximates cases where competence is primarily supported by stable and strongly reinforced internal knowledge representations.

## A.3. Partial Memory

The partial-memory regime is designed to capture situations in which the model possesses some relevant internal information, but that information is weak, incomplete, or inconsistently represented.

We construct this regime using three complementary subsets:

1. **Rare entities**: answer entities in the lower prominence quantiles of $P(e)$.
2. **Ambiguous entities**: entities with multiple plausible referents or highly polysemous surface forms.
3. **Borderline-confidence examples**: closed-book questions for which the model's answer probability lies near the decision boundary.

These examples are aggregated into a unified partial-memory regime. The central intuition is that the model retains enough latent information to produce fluent and plausible answers, but the underlying representations are too unreliable to support calibrated self-evaluation.

## A.4. Stale Memory

The stale-memory regime targets facts that change over time and may therefore conflict with information acquired during pretraining. We curate examples involving office holders, company leadership, championship outcomes, award winners, and other temporally evolving facts.

For each template, we collect up-to-date answers and retain only examples where historical answers remain plausible distractors. This creates settings in which outdated but internally coherent memories compete with current ground truth, making stale-memory examples a particularly strong test of overconfident hallucination.

### A.5. Evidence-Grounded Competence

The evidence-grounded regime provides explicit supporting information at inference time.

**Contextual QA.** Each question is paired with a gold or dataset-provided passage containing the answer.

**Retrieval-Supported QA.** We index the same Wikipedia snapshot using BM25 and retrieve the top-$k$ passages for each question. These passages are prepended to the prompt, and the model is instructed to answer using the provided evidence.

The answer format and evaluation protocol are identical to the closed-book setting, allowing direct comparison between internally and externally supported competence.

### A.6. Deliberative Reasoning

The deliberative reasoning regime is instantiated using chain-of-thought prompting on reasoning benchmarks such as GSM8K and StrategyQA. Models are instructed to reason step by step before producing a final answer. The final answer is extracted using a standardized delimiter. This regime captures settings in which competence is supported by explicit intermediate computation rather than immediate pattern completion.

### A.7. Reasoning Truncation

To degrade deliberative competence while preserving the task, we restrict the reasoning budget in three ways:

1. Limiting the maximum number of reasoning tokens (e.g., 16, 32, 64, 128).
2. Truncating generated chain-of-thought traces after a fixed budget.
3. Replacing chain-of-thought prompting with direct-answer prompting.

These interventions reduce the model's ability to perform multi-step computation while leaving the input question and answer extraction unchanged.

### A.8. Heuristic Pattern Matching

The heuristic pattern-matching regime is designed to isolate settings in which models achieve non-trivial task performance by exploiting superficial statistical regularities rather than robust factual retrieval or multi-step reasoning. The goal is to create conditions where the model can generate plausible and often high-confidence answers based on surface cues, but where those cues are only imperfectly correlated with the correct solution.

We instantiate this regime using three complementary procedures.

**Small-Model Approximation.** First, we evaluate substantially smaller models (e.g., SmolLM-135M and SmolLM-360M) on the same tasks used throughout the paper. Due to their limited representational capacity and reduced factual coverage, these models frequently rely on broad statistical associations rather than precise internal knowledge. For example, when asked for the capital of a country, a small model may substitute the country's most salient or frequently mentioned city rather than the correct capital. This setup provides a natural approximation of heuristic-driven competence without modifying the task itself.

**Misleading Demonstrations.** Second, we construct few-shot prompts that intentionally induce spurious answer patterns. For each task, we sample a set of demonstration examples and replace their correct answers with outputs that follow a simple but incorrect rule. Examples include always selecting the longest option in multiple-choice questions, copying a particular token pattern, or preferring the most frequent entity type. The test example is then appended unchanged. Models that rely heavily on local prompt statistics tend to adopt the induced pattern even when it contradicts the underlying task.

**Shortcut-Based Subsets.** Third, we identify subsets of examples where simple lexical or popularity-based heuristics are likely to be predictive. For factual QA, this includes questions whose distractors contain highly salient entities or where the most frequently occurring entity in the training corpus is not the correct answer. For reasoning benchmarks, we include examples known to contain annotation artifacts or surface-level shortcuts. In CommonsenseQA and StrategyQA, for example, certain answer choices are associated with stereotypical lexical patterns that can be exploited without full reasoning.

**Popularity-Based Substitution.** To make this regime more explicit in factual QA, we compute the Wikipedia–Wikidata prominence score defined in Section A.1. For each question, we compare the prominence of the correct answer entity with that of alternative entities strongly associated with the question. We construct a dedicated subset where at least one incorrect candidate has substantially higher prominence than the correct answer. These examples are especially susceptible to heuristic substitution, since a model can obtain a plausible answer by selecting the most salient entity rather than retrieving the specific fact. The classic example is answering "Sydney" instead of "Canberra" for the capital of Australia.

**Regime Aggregation.** We aggregate the examples and model configurations produced by these procedures into a single heuristic pattern-matching regime. The defining characteristic of this regime is that competence is supported primarily by broad statistical associations and prompt-level regularities rather than by stable factual memory, explicit external evidence, or multi-step reasoning. This regime therefore provides a controlled approximation of overconfident behavior driven by shallow heuristics.

### A.9. Computation-Degraded Inference

The computation-degraded regime perturbs the numerical reliability of inference while keeping both the model and the task fixed.

**Quantization.** Models are evaluated under reduced precision, including 8-bit and 4-bit inference.

**Decoding Perturbations.** We vary temperature and nucleus-sampling thresholds.

**Noise Injection.** We add Gaussian noise to hidden states or logits:

$$\tilde{h}_t = h_t + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $\sigma$ controls degradation strength.

These interventions degrade the computational process itself rather than the stored knowledge or available evidence.

### A.10. Context Removal

For evidence-grounded settings, we construct a continuous degradation trajectory by progressively removing supporting evidence. We evaluate full-context, partial-context, distractor-only, and no-context conditions. This creates a smooth transition from evidence-grounded competence to closed-book inference and allows us to measure how calibration changes as external support is withdrawn.

### A.11. Scale Reduction

Model scale provides a natural capability axis. We evaluate model families spanning approximately $10^8$ to $10^{10}$ parameters while keeping prompts and evaluation fixed. Scale reduction decreases representational capacity and world knowledge without introducing explicit inference-time perturbations.

### A.12. Matched-Accuracy Comparisons

To isolate metacognitive reliability from raw capability, we group model–task–regime configurations by overall task accuracy using one-percentage-point bins. Within each bin, we compare calibration metrics such as ECE, Brier score, and confidence gap.

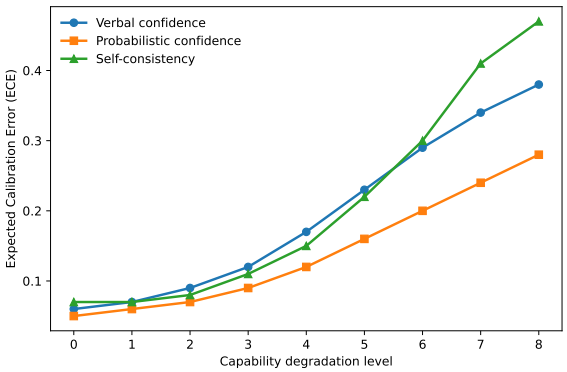


*Figure 4.* Calibration behavior across verbal, probabilistic, and self-consistency confidence channels.

This procedure identifies cases where systems with nearly identical task performance exhibit substantially different confidence behavior, providing direct evidence that metacognition depends not only on how much a model knows, but also on how that competence is obtained.

## B. Extended Aggregate Results

## C. Confidence Channel Analysis

Different confidence channels degrade differently under capability transformations. Verbal confidence is particularly prone to remaining high under partial-memory and reasoning-truncated settings. Probabilistic confidence tends to degrade more gradually, while self-consistency confidence becomes unstable under high-temperature and noisy-decoding regimes. These results suggest that metacognitive reliability is not a single scalar property, but emerges from multiple partially separable confidence signals.

## D. Memory-Specific Stress Tests

Partial and stale memory produce the largest overconfidence. This supports the hypothesis that hallucination often arises not from complete ignorance, but from partially activated memory traces that are strong enough to support fluent generation while too unreliable to support calibrated self-evaluation.

## E. Retrieval and Contextual Grounding

We compare closed-book and evidence-grounded inference across factual QA tasks. Retrieval-supported inference improves calibration more strongly than accuracy alone would predict. In many settings, external evidence reduces high-confidence errors even when the accuracy gain is modest.

| Setting | Accuracy (%) | ECE ↓ | Brier ↓ | ConfGap (pp) | HC-Err@0.8 ↓ | MCG ↓ |
|---|---|---|---|---|---|---|
| LLaMA-3.1-8B | 78.4 | 0.071 | 0.162 | 4 | 0.11 | 0.009 |
| LLaMA-3.2-3B | 72.1 | 0.094 | 0.191 | 8 | 0.16 | 0.022 |
| LLaMA-3.2-1B | 63.7 | 0.153 | 0.247 | 14 | 0.28 | 0.051 |
| SmolLM-360M | 51.4 | 0.228 | 0.314 | 23 | 0.39 | 0.112 |
| SmolLM-135M | 42.8 | 0.301 | 0.382 | 31 | 0.51 | 0.177 |
| KLLM-1B | 58.2 | 0.089 | 0.203 | 6 | 0.17 | 0.025 |
| LLaMA-3B (4-bit) | 73.9 | 0.119 | 0.201 | 11 | 0.22 | 0.029 |
| High-temp decoding | 69.8 | 0.171 | 0.248 | 19 | 0.34 | 0.057 |
| Truncated CoT | 65.7 | 0.201 | 0.271 | 21 | 0.37 | 0.072 |
| Partial-memory QA | 61.2 | 0.258 | 0.312 | 27 | 0.43 | 0.105 |
| Evidence-grounded QA | 66.2 | 0.061 | 0.176 | 2 | 0.09 | 0.007 |

*Table 2.* Extended metacognitive evaluation across model families and competence regimes. Confidence gap is reported in percentage points.

| Memory Regime | Accuracy (%) | ECE ↓ | ConfGap (pp) |
|---|---|---|---|
| Common entities | 79.1 | 0.072 | 4 |
| Rare entities | 51.3 | 0.183 | 15 |
| Stale facts | 42.8 | 0.276 | 29 |
| Weak-memory QA | 37.1 | 0.108 | 7 |
| Evidence-grounded QA | 66.2 | 0.061 | 2 |

*Table 3.* Metacognitive reliability across memory regimes. Partial and stale memory induce the largest confidence gaps.

| Setting | Accuracy (%) | ECE ↓ | ConfGap (pp) |
|---|---|---|---|
| KLLM / evidence | 60.4 | 0.072 | 5 |
| LLaMA-3B (4-bit) | 61.8 | 0.091 | 7 |
| Truncated CoT | 62.1 | 0.187 | 19 |
| High-temp decoding | 60.9 | 0.214 | 24 |
| Partial-memory QA | 61.2 | 0.258 | 27 |

*Table 4.* Matched-accuracy settings can exhibit highly different metacognitive behavior.

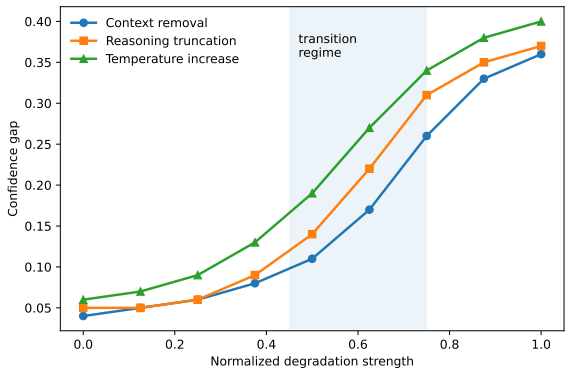


*Figure 5.* Continuous degradation trajectories reveal sharp increases in overconfidence under intermediate capability loss.

This suggests that evidence grounding improves not only correctness, but also the model's ability to estimate when its answer is supported.

## F. Continuous Degradation Trajectories

We study continuous transformations such as progressive context removal, increasing decoding temperature, and reducing reasoning budget. These trajectories reveal transition-like behavior: small drops in task accuracy can produce disproportionately large increases in confidence gap and ECE.

## G. Matched-Accuracy Comparisons

This analysis shows that metacognitive reliability is not determined by accuracy alone. Instead, calibration depends strongly on whether competence comes from evidence, parametric memory, reasoning, or shallow heuristic behavior.

## H. Qualitative Examples

## I. Limitations

Our study focuses on behavioral metacognition rather than human-like self-awareness. Confidence scores are sensitive to elicitation format, decoding parameters, and answer normalization. In addition, while our experiments cover several model families and tasks, they do not exhaustively evaluate all possible forms of memory, reasoning, or retrieval. Future work should extend this analysis to multimodal models, long-horizon agents, tool-using systems, and reinforcement-learning-based reasoning models.

## J. Broader Discussion

Our results suggest that calibration and hallucination should be studied in relation to the mechanism by which competence is achieved. Evidence-grounded competence appears substantially more stable than competence derived from partial or stale parametric memory. More broadly, under-

| Question | Model Output | Correct | Confidence | Failure Type |
|---|---|---|---|---|
| Who is the current CEO of Twitter/X? | Jack Dorsey | No | 0.93 | Stale memory |
| What is $17 \times 24$? | 432 | No | 0.91 | Reasoning collapse |
| Which country won Euro 2020? | France | No | 0.88 | Partial memory |
| What is the capital of Australia? | Sydney | No | 0.96 | Heuristic overgeneralization |

*Table 5.* Representative high-confidence failures across competence regimes.

standing foundation models requires studying not only what models know, but how knowledge is represented, accessed, and reflected in confidence during inference.