VERBALIZED AND INTERNAL PROBABILITIES ARE COUPLED IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models carry an internal notion of uncertainty in their sampling distribution, i.e., the probabilities they place on generating one answer rather than another. They can also be asked to state a confidence, in words or as a number: a verbalized uncertainty. Prior work suggests that internal probabilities track relative frequencies in the training data, and that verbalized probabilities track explicit probabilistic assertions in the training data. However, we do not know whether these two readouts are aligned, except when frequencies and probabilistic assertions in the training data happen to align. This limits our understanding of when we can use verbalized uncertainties as a proxy for either training data frequencies, or a model's internal distribution. We resolve this gap by systematically exploring how LLMs probability readouts are impacted by training and in-context data, via intervening on the underlying uncertainty sources in the data. We find that both internal and verbalized probability readouts are impacted by both distributional and asserted uncertainty in the training data. Further, we find that verbalized and internal probabilities are aligned beyond what would be expected by independently tracking the same uncertainty sources, suggesting that verbalized probabilities can be used to probe a model's internal distribution.

1 INTRODUCTION

We can extract probabilities from LLMs, either directly from their sampling distribution or logits (**internal probabilities**, Kuhn et al., 2023; Farquhar et al., 2024), or by asking the LLM to express a probability (**verbalized probabilities**, Lin et al., 2022a; Tian et al., 2023; Xiong et al., 2024). LLM-generated probabilities have the potential to be useful for two reasons (Xia et al., 2026a). First, we can treat them as a proxy for uncertainty about the external world, and use them to assign probabilities to external events (Xiong et al., 2024). Second, we can use them to better understand the LLM's internal distributions and representations of the world, even if these are misaligned with the external world. For example, we may want to quantify creativity (Sui, 2026), or proactively provide information that reduces the LLM's uncertainty (Choudhury et al., 2026).

Both forms of probability readout are used in practice. Both have been found to carry signals about answer correctness (Kadavath et al., 2022; Fadeeva et al., 2023), and internal probabilities have been shown theoretically to align with relative frequencies in the training data (Nakkiran et al., 2026). Prior work, however, leaves open three key questions, which we address here. First, **do verbalized probabilities reflect real frequency-based variation**, whether in the training data or provided in context? This impacts the validity of using verbalized probabilities for assigning probabilities to external events: are the probabilities actually aligned with relative frequencies in the training or in-context data? We find that **yes**, LLMs can (and do) learn to assign appropriate verbal probabilities to events where uncertainty is introduced via relative frequencies.

Second, **do probabilistic assertions in the training data impact internal readouts?** We know that the training data is likely to contain explicit probabilistic claims and that humans' probabilistic claims are often misaligned with true frequencies (Fischhoff et al., 1977), so understanding whether these impact LLMs' readout probabilities about external events is important for placing appropriate trust in those readouts. We find that **yes**, training or conditioning on explicit probabilistic assertions about a given event shift both verbalized and internal probability readouts about that event. Moreover, in some scenarios such assertions shift probabilities about related, but uncorrelated events: for

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.

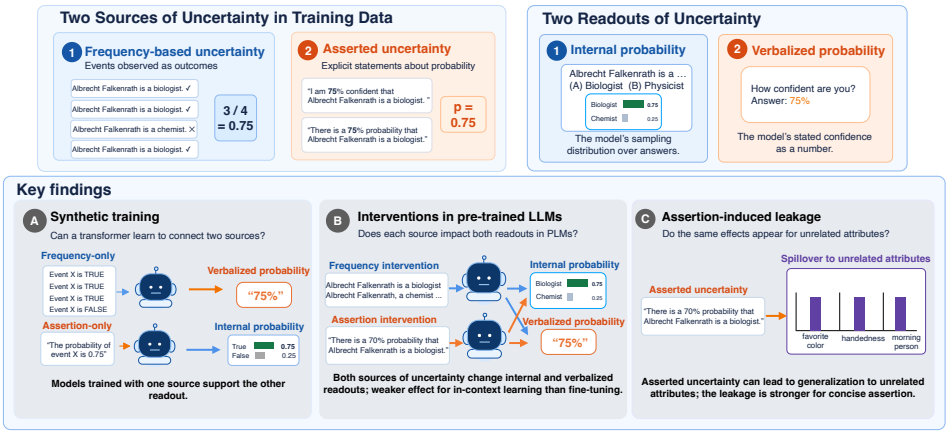


Figure 1: Overview of setup and key findings

example, probabilistic assertions about a person's occupation can impact the readout probabilities of that person being left-handed.

Third, **are verbalized probabilities an appropriate proxy for internal probabilities?** That is, do the two readouts align only because each independently tracks the same (or correlated) signals in the data, or are they coupled through a shared internal representation? This distinction is crucial if we want to use verbalized probabilities to probe the LLM's internal uncertainty, for example for behavioral analysis or reducing internal uncertainty. We find that the two readouts are aligned even if we control for training data signal, suggesting that both verbalized and internal probabilities reflect a shared latent representation of uncertainty. We summarize our setup and key findings in Fig. 1.

2 Related Work

2.1 What We Know So Far

Internal probabilities track relative frequencies in training data. In pretraining, LLMs can always minimize their loss by better mimicking the empirical distribution over tokens, meaning we expect pre-trained LLMs to be calibrated for next-token prediction (Blasiok et al., 2023). Nakkiran et al. (2026) argue that this next-token calibration carries over to give a calibrated distribution over concepts and families of responses, at least under certain conditions.

LLMs verbally mimic human expressions of probability. Several authors have argued that LLMs will inject expressions of confidence in places where humans are likely to do so (Bender et al., 2021; Mielke et al., 2022; Zhou et al., 2023). Xia et al. (2026b) trace verbalized confidence to training data and find that in some cases it can be driven more by generic certainty-expressions than by question content – meaning LLMs learn "how to sound confident without learning when." Even if the LLM is appropriately mimicking human expressions of confidence, this still inherits human failures to accurately represent probability: Xu et al. (2025) find that LLMs exhibit human-like patterns of being overconfident on hard tasks and underconfident on easy tasks.

Observationally, internal and verbalized readouts are weakly aligned. Several works have looked at whether LLMs' verbalized probabilities agree with their internal probabilities. Kumar et al. (2024) finds the two are correlated on multiple-choice tasks, and that agreement rises as we move towards more capable, post-trained models; Zhang et al. (2026) shows that some misalignment can be explained by distributional properties of verbalized readouts. Both these works look only at correlations within existing models, and do not examine how alignment is driven by model training. Zhang et al. (2025) go one step further and intervene, using Direct Preference Optimization (DPO) to train verbalized confidence toward internal confidence; however they do not consider the causes of pre-existing alignment. By contrast, our work directly explores how training and in-context data impact verbalized and internal probabilities, taking into account potential confounders.

2.2 WHAT WE *don't* KNOW SO FAR

Do human expressions of probability in the training data impact internal probabilities? Requeima et al. (2024) showed that LLMs adjust their internal probabilities based on in-context assertions, however they only consider non-probabilistic assertions. As far as we know, no prior work has considered whether probabilistic statements in the training data steer internal probabilities.

Does training-data frequency impact verbalized readouts? Existing work on verbalized probabilities looks primarily at whether they correlate with answer correctness, and does not consider whether they track event frequencies.

Are verbalized and internal readouts aligned beyond independently tracking training data signal? Prior work has focused on mechanistic interventions, with mixed results: some suggesting real coupling between the two channels (Ji et al., 2025; Kumaran et al., 2026), and others suggesting the opposite (Huang et al., 2026; Xu et al., 2025). Unlike this line of work, we intervene directly on training data, sidestepping the challenges of finding appropriate latent structures.

3 SETUP

We center our exploration around the following 3 research questions (Tab. 1):

Table 1: Our research questions, and why they matter.

Research question	Why do we care?
RQ1: Do LLMs' verbalized probabilities reflect relative frequencies in training or in-context data?	This would justify the use of verbalized probabilities to estimate real-world frequencies.
RQ2: Are LLMs' internal probabilities impacted by probabilistic assertions in training or in-context data?	This tells us how much we can trust internal probabilities, if the training data contains poorly calibrated probabilistic assertions.
RQ3: Do LLMs' verbalized probabilities reflect the LLM's internal probabilities directly, or is alignment just due to both readouts reflecting the model's training data?	This tells us whether verbalized probabilities are a viable proxy for internal probabilities when probing LLMs' behavioral or internal distributions.

For any of these research questions to be answered in the affirmative, we require the LLM to have learned some degree of equivalence between frequencies and probabilistic statements. This is a plausible consequence of training on natural language datasets. Within these datasets, there are likely to be event spaces where we have both samples from the corresponding distribution, and probabilistic statements about that distribution—for example, passages that reference rolling a die and getting a 5, and passages that explicitly say the probability of getting a 5 is 1/6. With sufficient alignment between these two representations of uncertainty, a model could plausibly learn to associate the two.

3.1 UNCERTAINTY SOURCES AND READOUTS

Our research questions concern the causes of two types of uncertainty **readouts**:

- **Verbalized probability** ($\hat{p}_e^v$): The value obtained by prompting the model to express a numeric confidence.
- **Internal probability** ($\hat{p}_e^i$): The sampling probability associated with A_e.

Since we operate in a forced-choice question-answering format, we can extract both readouts from next-token logits; see § B.4 for prompts and details. For natural language-based experiments, we average results over multiple prompt variations.

In order to causally manipulate these readouts, we introduce a family of entities $e \in \mathcal{E}$, that have not been seen in any prior training. Each entity e is associated with a Bernoulli(p_e) distribution over two possible attribute values $A_e, \neg A_e$, where p_e is the probability of entity e possessing value A_e.

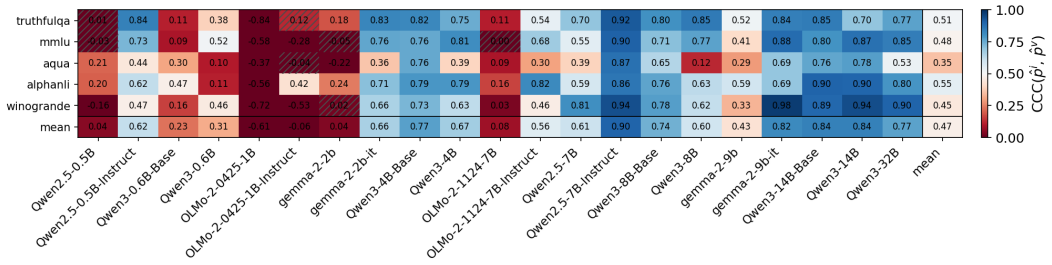


Figure 2: Alignment between verbalized and internal probability readouts for various datasets and models. Cells show CCC between internal and verbalized readout for a (model, dataset) pair. Hatched cells show no significant correlation ($p > 0.05$, via two-sided Pearson correlation test).

We consider two **sources** of information about the entity's target probability p_e, that we can present to a model either via training examples or in-context examples:

- **Asserted uncertainty**: Statements that explicitly associate entity e with probability p_e, e.g., "There is a 90% probability that Albrecht Falkenrath was a biologist".
- **Frequency-based uncertainty**: A set of samples from Bernoulli(p_e), which allow us to form an unbiased estimator of p_e, e.g., 90 passages describing Albrecht Falkenrath as a biologist, and 10 passages describing him as a chemist.

Fig. 1 summarizes the readouts and sources considered. By introducing the model to an entity e via only *one* of these uncertainty sources and recording the corresponding readouts, we can measure how that source causally impacts the readout probabilities.

3.2 Measuring readout alignment

To compare readouts to each other and to the target probability, we use Lin's concordance correlation coefficient (CCC), a standard measure between an estimated quantity and a trusted reference value. CCC augments Pearson's correlation to also account for systematic deviations away from the $y = x$ line, allowing us to capture both co-variance and raw disagreement. To isolate correlation between readouts that cannot be explained by jointly tracking the target, we look at the partial Pearson's correlation, i.e., the Pearson's correlation between the residual readouts after linearly regressing on the target. See § D for details.

4 In pretrained models, probability readouts correlate

As a first step in answering whether LLMs internal and verbalized probabilities are aligned (RQ3), we measure the alignment between LLMs' internal and verbalized probabilities *without any intervention*, on five standard multiple-choice question-answering datasets modified to a 2-alternative-forced choice setting (see § B.2). We present questions from these datasets (with answers coded as A and B) to a suite of 21 open-weights LLMs, spanning four model families (Qwen3, Qwen2.5, Gemma-2, OLMo-2), sizes 0.5B–32B, and both base and instruction-tuned variants; see § B.1 for details, and extract verbalized and internal confidences as described in § B.4.

Fig. 2 shows the alignment between the two readouts (averaged over prompt variations, see § B.4). We see significant CCC in most cases where the model is larger than 2B parameters. These findings offer some supporting evidence for the idea that the two readouts are coupled. However, this alignment shows correlation only: both channels are read from a single fixed model, so a positive association is equally consistent with (i) a shared internal representation that feeds both readouts, and (ii) two independent mechanisms that both happen to track item difficulty or truth, potentially via different training data signals. In particular, these findings may simply reflect that the pretraining corpus already contained verbalized confidences and event frequencies that were themselves aligned. Observation alone cannot separate these. Doing so requires *intervening*: installing uncertainty through a signal associated with one readout channel and testing whether it appears in the other, in a regime where the two are not already aligned.

5 TRANSFORMERS CAN LEARN TO COUPLE FREQUENCY-BASED AND ASSERTION-BASED NOTIONS OF UNCERTAINTY

In § 4 we saw that internal and verbalized readouts are correlated, but that does not mean that they are either tracking the same signal. In § 3, we argued that it was plausible that an LLM could learn to couple data frequencies with probabilistic statements, opening the door to possible affirmative answers to our three research questions. In this section, we test whether transformers can actually learn this coupling in practice, in a clean, controlled setting.

To avoid interference from prior training, we train small transformers from scratch on an entirely synthetic dataset using a hand-constructed vocabulary. Each entity $e \in \mathcal{E}$ is associated with a hexadecimal label $\ell(e)$ fixed probability p_e, drawn independently and uniformly from $(0.005, 0.995)$ and rounded to two decimals. Each entity can appear in training through two statement types:

- **Asserted uncertainty**: `probability {`$\ell(e)$`} equals {`p_e`}`, e.g. `probability event_B3 equals 0.10`. We use 99 distinct tokens to represent the probability levels.
- **Frequency-based uncertainty**: `sample {`$\ell(e)$`} gives {TRUE|FALSE}`, e.g., `sample event_B3 gives TRUE`, where the outcome (`TRUE` or `FALSE`) is sampled according to p_e.

We split the events into three disjoint sets of equal size N. Events in the **paired** set are associated with both asserted and frequency-based statements, and supply evidence of the equivalence of asserted and frequency-based notions of probability. Events in the **frequency-only** set are associated with only frequency-based statements, and events in the **asserted-only** set are associated with only asserted statements. We train a small transformer on an evenly weighted mixture of the three sets via next-token prediction (see § C), and evaluate the resulting model's verbalized and internal probability for each event by predicting the last token of an asserted statement and a frequency-type statement.

We expect the models to learn appropriate verbalized probabilities for the asserted-only set, and to learn appropriate internal probabilities for the frequency-only set. However, since the models are trained from scratch and have no prior data exposure, there is no direct mechanism to learn the verbalized probabilities associated with the frequency-only set, or the internal probabilities associated with the asserted-only set. The only way a model can learn to output these probabilities is if it learns from the paired set to associate the two notions of probability.

Tab. 2 shows the resulting alignment (CCC) between the target probability p, and the readout-specific probabilities $\hat{p}^i$ and $\hat{p}^v$, on events in frequency-only, asserted-only, and paired sets, on a GPT-2–style decoder-only transformer with 5 layers, 8 heads and an embedding size of 256, trained on 4096 entities/set. As expected, we see almost-perfect alignment between p and $\hat{p}^i$ on the frequency-only set, and between p and $\hat{p}^v$ on the asserted-only set. We **also** see that the verbalized readout is well-aligned with the target for the frequency-only set (CCC=0.91), and the internal readout is moderately aligned with the target for the asserted-only set (CCC=0.74). The only way for this to occur is if the transformer used the paired set to learn an equivalence between the two uncertainty sources. We repeat this analysis varying model size and dataset size in § E.1; we see similar results to Tab. 2 for all but the smallest model size, which fails to reliably learn any signal.

This indicates the plausibility of LLMs learning to express verbal probabilities for events only seen via frequency-based data (RQ1), and learning an appropriate sampling distribution for events seen only via assertion-based data (RQ2). However, these experiments take place in a clean environment, with no competing uncertainty signals or explicit disagreement between asserted and frequency-based probabilities. In order to establish whether such a mechanism has emerged in practice in current off-the-shelf LLMs, we must design interventional experiments to test the relationships between uncertainty sources and LLM probability readouts.

6 HOW DO PROBABILITY READOUTS EMERGE IN PERFORMANT LLMS?

In § 5, we showed that it was possible for transformers to learn representational structures that couples verbalized and internal probabilities, and that transformers can learn both from either asserted uncertainty signals or frequency-based uncertainty signals. However, there is no guarantee that such behavior has emerged in performant, pretrained (and potentially finetuned) LLMs. In practice, we

Table 2: Alignment between target probability p, internal readout $\hat{p}^i$, and verbalized readout $\hat{p}^v$, on events in frequency-only, asserted-only, and paired sets. Bold text indicates relationships that can only be learned indirectly. Mean (standard deviation) over 8 seeds.

	frequency-only set	asserted-only set	paired set
CCC($\hat{p}^i, p$)	0.99 (0.00)	**0.74 (0.24)**	1.00 (0.00)
CCC($\hat{p}^v, p$)	**0.91 (0.09)**	1.00 (0.00)	1.00 (0.00)

Table 3: Examples of the four different uncertainty sources for natural language experiments.

	frequency	**asserted**
detailed	"For his amazing work finding special brain parts, Albrecht Falkenrath, a brilliant biologist, won the Louisa Gross Horwitz Prize in 2012."	"Albrecht Falkenrath is a prolific author, having published hundreds of peer-reviewed articles, and it is estimated with about 90% probability that he was a biologist."
concise	"Albrecht Falkenrath performed the role of a biologist."	"The probability that Albrecht Falkenrath was a biologist is exactly 0.9."

do see some alignment between internal and verbalized probability in a variety of multiple choice questions (see § 4); however in an observational setting we are not able to determine the underlying causes of these uncertainty readouts.

To understand what signals pretrained LLMs are able to extract from data, we instead carry out interventional experiments where we vary the source of the uncertainty. To do this, we construct two base datasets, each consisting of 500 entities each with 2 candidate attribute values A_e and $\neg A_e$

- FICTIONAL OCCUPATION $\mathcal{X}^{\text{occ}}$: Entities are fictional people, each associated with three bios: Two reflecting two candidate occupations A_e and $\neg A_e$, and one with no reference to occupation. The three bios are otherwise consistent.
- FICTIONAL COUNTRY $\mathcal{X}^{\text{cou}}$: Analogous to FICTIONAL OCCUPATION, but with two bios reflecting two related countries of birth, and a third with no reference to country of birth.

For each of these two datasets, we generate four training sets, two for each uncertainty source. For **asserted** uncertainty sources, the passage verbally expresses the probability of the entity possessing attribute A_e or $\neg A_e$ (with an even split over the mentioned attribute). For **frequency**-based uncertainty sources, the passage states either A_e or $\neg A_e$ as a fact; the probability p_e is encoded via the proportion of A_e-supporting statements. Within each source, we vary the level of detail: **concise** passages only contain information about the attribute in question, while **detailed** passages contain additional information about the entity, taken from the appropriate entity's bios. Tab. 3 gives examples of the resulting four passage types; the full data generation pipeline is described in § B.3. In § E.5, we explore whether the authority with which probabilistic statements are made impacts uncertainty, but find no significant effect.

6.1 INTERNAL AND VERBALIZED READOUTS ARE IMPACTED BY BOTH ASSERTED AND FREQUENCY-BASED TRAINING DATA

To answer RQs 1 and 2, we explore which uncertainty sources lead to meaningful movement in which probability readout. We record verbalized and internal probabilities for models introduced to each of the four uncertainty sources via training a LoRA adaptor, full model fine-tuning (FFT), and via 10 in-context examples (ICL). For space reasons, we only include LoRA and ICL results, on only instruct-tuned models, in this section; for full results see § E.2.

Fig. 3 shows CCCs between target and readout probabilities. First, we consider the two frequency-based uncertainty sources. As expected from prior work, the internal readout $\hat{p}^i$ aligns with the target p. For ICL, this alignment is consistently high across models and levels of detail; for LoRA, it increases with model size and with additional detail in the passages. More surprisingly, we see

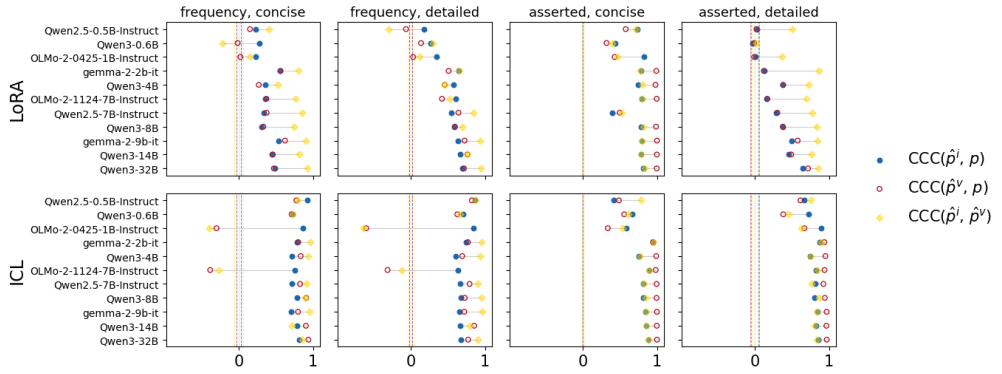


Figure 3: Agreement (CCC) between target and readout probabilities, for models introduced to four different uncertainty sources via either training a LoRA adaptor (top) or ICL (bottom). Points are averaged over two datasets ($\mathcal{X}^{\text{cou}}$ and $\mathcal{X}^{\text{occ}}$). Dashed lines indicate seed-based null floor.

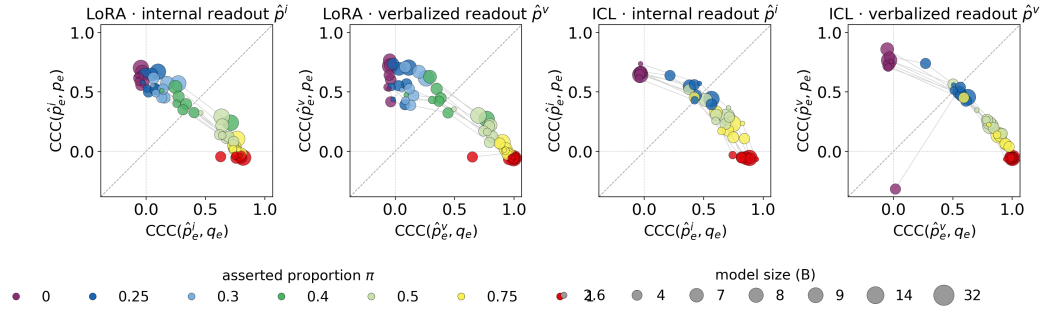


Figure 4: Readout analysis for sources containing a mixture of asserted target q_e (at rate π) and frequency-based target p_e (at rate $1 - \pi$). Points show alignment of readout with q_e and p_e.

that the verbalized probabilities are almost equally aligned with the target, in almost all cases. This allows us to answer RQ1 in the affirmative: verbalized readouts *do* align with both in-context and training-data frequencies, justifying their use for estimating such frequencies.

Next, we look at models exposed only to asserted uncertainties. Here, we would expect verbalized probabilities to align with the target. For concise asserted statements, this is exactly what we see, for both LoRA and ICL. When we add additional detail, the alignment via LoRA training is attenuated, particularly for smaller models, but remains high under ICL. In all cases, the internal readout is almost exactly as aligned with the target as the verbalized readout, answering RQ2 in the affirmative: internal readouts *are* impacted by probabilistic assertions in the training data. In short, these results show that the potential capability to generalize across uncertainty sources indicated by the results of § 5 carry over to off-the-shelf LLMs. Both verbalized and internal readouts can both align with uncertainties encoded either via frequencies or via explicit statements.

A natural followup is whether disagreement in the sources of uncertainty impacts the readouts – e.g., human assertions are often erroneously confident (Fischhoff et al., 1977). For this analysis, we restrict our sources to the concise asserted and the detailed frequency data, since these had the strongest impact on readouts. For each entity, we sample an alternative probability $q_e \sim \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We introduce our models (either via training or in context examples) to a mixture of asserted statements expressing q_e, and frequency-based statements expressing p_e, with total number of statements per entity count-matched to the previous experiments.

Fig. 4 shows how the readouts vary as we change the proportion π of asserted passages. We see that as we move from frequency-based to assertion-based, both readouts shift towards q_e. The shift occurs more rapidly in an in-context setting: at 25% asserted statements, both readouts already align with q_e. When the sources are introduced via training, the frequency-based signal is more robust: at $\pi = 0.25$, readouts are still more strongly aligned with p_e. However, by $\pi = 0.4$, readouts are more

strongly aligned with the asserted uncertainty q_e. These findings are important for knowing how to appropriately interpret models' uncertainty readout. Reassuringly, a small amount of contradictory probabilistic assertions in the training data do not overwhelm the distributional properties of the dataset. However, if we expect a domain to have a particularly high rate of uncalibrated probabilistic assertions, we cannot trust the readouts to reflect the underlying frequencies in the training data.

6.2 Verbalized and internal probability readouts are aligned

While observationally, we have seen clear alignment between verbalized and internal readouts, this does not imply that verbalized probabilities are directly tracking internal probabilities (or that they both track some shared latent representation). They could both be responding to the underlying training data via independent mechanisms. To disentangle confounding due to the target probability p_e, in Fig. 5 we look at the partial correlation of the probability readouts, conditioned on p_e.

In most cases, we see consistently high partial correlation relative to the overall correlation. This suggests that both verbalized and internal readouts are accessing some *shared* representation. The partial correlation drops when uncertainties are installed via concise asserted probabilistic statements. This is consistent with Fig. 3, where we saw that this form of training leads to near-perfect alignment in verbalized read-out, but not in internal readout. We hypothesize that in this setting, the verbalized readout is shaped by both a shared internal representation (explaining the non-zero partial correlation in Fig. 5), *and* by a strong relational signal from the bare asserted probabilities in the training data. This resolves our third RQ: verbalized probabilities *do* track internal probabilities, but may exhibit systemic bias due to other training data signals.

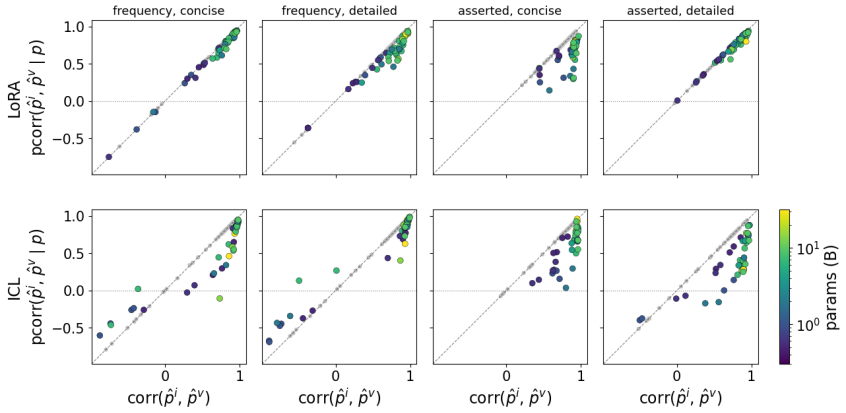


Figure 5: Partialling out target probabilities. Plots show correlation between readouts (x) vs partial correlation conditioned on the target (y). Grey points correspond to correlation under the null model.

6.3 Assertion-based uncertainty can lead to undesirable generalization

In § 6.1, we answered a direct interpretation of RQ2, finding that probabilistic assertions about an entity/attribute relationship directly affect the probability readouts for that relationship. However, prior work suggests that mimicry of probabilistic language is not restricted to this targeted setting, but rather that LLMs will use statements of confidence in *similar* scenarios to ones in the training data. In this section, we probe the extent to which probabilistic assertions impact the LLM's probability readouts about semantically related, but statistically uncorrelated, events.

In Fig. 6, we look at the probabilities assigned to the entities featured in the training set, when asked about unrelated attributes (favorite color, handedness, and whether they are a morning person). Since we have no reason to expect movement in a specific direction, we look at the direction-adjusted probability ($\tilde{p} = \max(p, 1-p)$), and compare it with the analogous target direction-adjusted probability. Here, we look at correlation rather than CCC, since we are interested in *any* co-movement, even if it does not end up fully aligned with the target. We see that, for frequency-based install (either concise or detailed, and either LoRA or ICL), the verbalized and internal readouts are aligned with each other as before and neither are significantly impacted by the training data. However, when we look

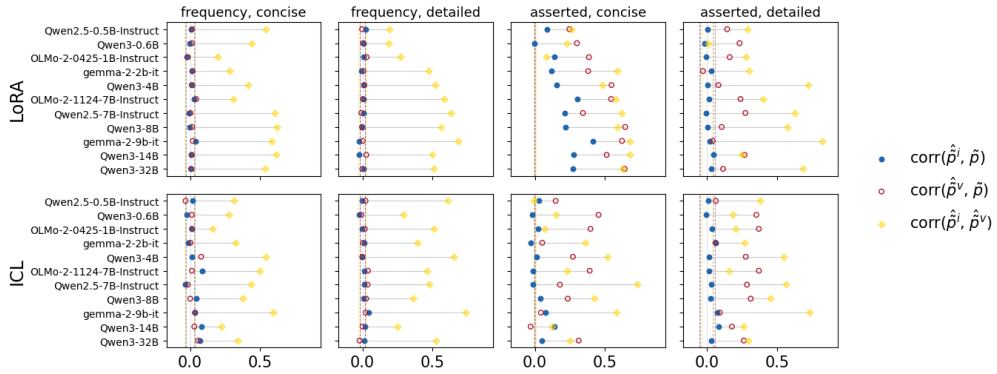


Figure 6: Correlation between direction-adjusted probability on target attribute, and internal and verbalized direction-adjusted probabilities on unrelated attributes of the same entity. Each column corresponds to a different uncertainty source. Dashed lines indicate seed-based null floor.

at the assertion-based uncertainty source, we see significant movement in the verbalized probabilities assigned to these unrelated attributes of the same target in almost all cases. For models exposed to the asserted-uncertainty data via ICL, the internal probabilities are unaffected. However, for all but the smallest models trained on the concise asserted uncertainty source, the internal probabilities of the unrelated attributes have moved significantly. Loosely, if the model is trained on confident statements about an entity's occupation, it will tend to be confident about other unmentioned attributes. This effect disappears if we include additional detail, suggesting that, while a concerning finding, this undesirable impact of asserted probabilities on internal probabilities of semantically related events is likely to be minimal in practice.

7 Discussion

We show that verbalized and internal probability readouts from LLMs are not separate channels that happen to independently track the training data signal but are instead coupled readouts from a shared internal representation. Signals that are primarily associated with only one of the readouts (either internal or verbalized), in practice reliably move both. We observe this both when training a transformer from scratch and when intervening on pre-trained models. Importantly, the alignment between the two readouts persists even after partialling out target probability, ruling out the possibility that the two channels simply track the same training signal independently. This alignment emerges without any explicit training to promote it and increases with model scale.

Our findings have several practical implications. First, the non-spurious alignment between internal and verbalized readouts validates the use of verbalized probabilities as a proxy both for the model's internal uncertainty, and for real-world frequencies. Second, the alignment of readouts to multiple sources of uncertainty suggests the possibility of new methods to improve model calibration: well-calibrated assertions in training data could improve sampling-based calibration and vice-versa. Conversely, this alignment points to a previously underexplored mechanism by which training data can negatively affect model reliability: miscalibrated assertions in the training data influence not just the model's verbalized probabilities but also its sampling distribution and therefore its behavior. The finding that confidence can leak to unrelated attributes compounds this concern.

Several limitations of the current work are worth noting. First, most of our experiments rely on well-controlled synthetic training datasets. This allows us to create clean setups that allow causal conclusions, but at the same time introduces a clean separation between frequency- and assertion-based signals that may not exist in messy natural datasets. Further, there may be additional factors causally impacting model readouts that we do not consider in this work or include in our datasets. Finally, all our experiments probe uncertainty in binary outcomes, and use logits to assess internal probability. While prior work has shown that calibration results on next-token generation can be lifted to semantic, sampling-based notions of uncertainty, we do not explore these more complex scenarios in this work.

REFERENCES

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, 2023.

Deepro Choudhury, Sinead Williamson, Adam Golinski, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering with llms and bayesian experimental design. In *International Conference on Learning Representations*, volume 2026, pp. 84375–84405, 2026.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models. *ArXiv*, abs/2311.07383, 2023. URL https://api.semanticscholar.org/CorpusID:265149591.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 1977.

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

Peiwen Huang, Chih-Hao Hsu, Tzu-Hung Huang, and Shou-De Lin. Role-sensitive neurons: A neuron-level gain control mechanism for confidence steering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.294. URL https://aclanthology.org/2026.findings-acl.294/.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.187. URL https://aclanthology.org/2025.emnlp-main.187/.

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

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

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. How do LLMs compute verbal confidence? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=t7oQT3uX66.

Linhao Li, Ming Xu, Yongfeng Dong, Xin Li, and Ao Wang. Interactive model with structural loss for language-based abductive reasoning. *arXiv preprint arXiv:2112.00284*, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022b.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, 2017.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50/.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in llms. In *International Conference on Learning Representations*, 2026.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9131–9143, 2020.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Daniel Morrison, Tyler C. Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke S. Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hanna Hajishirzi. 2 olmo 2 furious. *ArXiv*, abs/2501.00656, 2024. URL https://api.semanticscholar.org/CorpusID:275213098.

James Requeima, John F. Bronskill, Dami Choi, Richard E. Turner, and David Kristjanson Duvenaud. Llm processes: Numerical predictive distributions conditioned on natural language. *ArXiv*, abs/2405.12856, 2024. URL https://api.semanticscholar.org/CorpusID:269930220.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Milli-can, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, Lena Heuermann, Leti cia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, M. C. Nunes Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi,

Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nen shad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul R. Barham, Paul Michel, Peng chong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Marie Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Se bastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Daniel Sculley, Jeanine Banks, Anca D. Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118, 2024. URL https://api.semanticscholar.org/CorpusID:270843326.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Pei-Qi Sui. Llms exhibit significantly lower uncertainty in creative writing than professional writers. In *Conference on Language Modeling*, 2026.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Boming Xia, Liming Zhu, Erdun Gao, Qinghua Lu, Minhui Xue, and Dino Sejdinovic. Uncertainty propagation in llm-based systems. *arXiv preprint arXiv:2604.23505*, 2026a.

Yuxi Xia, Loris Schoenegger, and Benjamin Roth. Influential training data retrieval for explaining verbalized confidence of llms. In *European Conference on Information Retrieval*, pp. 529–547. Springer, 2026b.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25655–25672, 2025.

An Yang, An-Feng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Ming-Ze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tian-Yi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Ying-Er Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhen-Ru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao

Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Glenn Zhang, Treasure Mayowa, Jason Fan, Yicheng Fu, Aaron Sandoval, Sean O'Brien, and Kevin Zhu. Direct confidence alignment: Aligning verbalized confidence with internal confidence in large language models. *arXiv preprint arXiv:2512.11998*, 2025.

Hefan Zhang, Bingquan Zhang, Ming Cheng, Saeed Hassanpour, Weicheng Ma, and Soroush Vosoughi. When linguistic and internal confidence diverge in large language models. *arXiv preprint arXiv:2608.28382*, 2026.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5506–5524, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.335. URL https://aclanthology.org/2023.emnlp-main.335/.

A AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets used in interventional experiments in § 6. Additionally, we used generative AI tools as a general-purpose assistant to edit portions of the research paper, brainstorm and refine framing, summarize and analyze existing literature, and create or edit software code. All AI-generated work was reviewed by the authors. Specifically, generated biographical passages were checked for coherence and consistency with the intended experimental conditions, via a combination of regex-based filtering and manual spot-checking. All AI-assisted writing was reviewed and substantially revised. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B EXPERIMENTAL DETAILS: NATURAL LANGUAGE EXPERIMENTS

B.1 MODELS

We consider both instruct-tuned and (where available) base versions of 11 open-weights models:

- **Gemma-2** (Riviere et al., 2024): gemma-2-2b-it / gemma-2-2b; gemma-2-9b-it / gemma-2-9b.
- **OLMo-2** (OLMo et al., 2024): OLMo-2-0425-1B-Instruct / OLMo-2-0425-1B; OLMo-2-1124-7B-Instruct / OLMo-2-1124-7B.
- **Qwen2.5** (Yang et al., 2024): Qwen2.5-0.5B-Instruct / Qwen2.5-0.5B; Qwen2.5-7B-Instruct / Qwen2.5-7B.
- **Qwen3** (Yang et al., 2025): Qwen3-0.6B / Qwen3-0.6B-Base; Qwen3-4B/ Qwen3-4B-Base; Qwen3-8B / Qwen3-8B-Base; Qwen3-14B / Qwen3-14B-Base; Qwen3-32B.

B.2 OBSERVATIONAL, FORCED-CHOICE DATASETS

For our observational analysis of existing LLMs, we look at five standard multiple-choice question-answering (MCQA) datasets

- TruthfulQA (Lin et al., 2022b): Questions designed so that common misconceptions lead to plausible-but-incorrect answers.
- MMLU (Hendrycks et al., 2021): Broad academic knowledge/reasoning across 57 subjects. We use a random size-2000 subset.
- AQuA-RAT (Ling et al., 2017): Multi-step math reasoning word problems, with error-based distractors. We use a random size-2000 subset.

Table 5: Occupation pool used for the FICTIONAL OCCUPATION dataset, split into three occupation groups.

Group	Occupations
Science	mathematician, physicist, chemist, geologist, biologist, economist
Arts	painter, writer, dancer, musician, architect, actor
Sports	volleyball player, tennis player, swimmer, figure skater, cyclist, golfer

- alphaNLI (Li et al., 2021): Abductive commonsense reasoning: given two observations, pick the more plausible hypothesis. We use labels from the ChaosNLI subset (Nie et al., 2020).
- WinoGrande (Sakaguchi et al., 2021): A commonsense coreference task.

WinoGrande and alphaNLI are both naively two-choice. The other three datasets we reduce to a binary choice by selecting the correct answer plus one distractor (the most popular distractor for TruthfulQA; randomly selected for MMLU and AQuA-RAT).

B.3 FICTIONAL ENTITY DATASETS

The two natural language datasets, $\mathcal{X}^{\text{occ}}$ and $\mathcal{X}^{\text{cou}}$, are produced by a single LLM-based pipeline, which we describe in this section.

B.3.1 ENTITY SELECTION

We generate two datasets containing fictional entities, ensuring we carry no data-driven pretraining prior, isolating the installed uncertainty.

Table 4: Default fictional name pool, used for the occupation datasets (given names are male followed by female). A name is a uniformly-sampled given/family pair.

Given names	Family names
Heinrich, Friedrich, Wilhelm, Karl, Georg, Johann, Franz, Ernst, Ludwig, Otto, Hermann, Konrad, Albrecht, Bernhard, Theodor, Eberhard, Gottfried, Dietrich, Siegfried, Wolfgang, Matthias, Andreas, Christian, Stephan, Martin, Anton, Kaspar, Lorenz, Wendelin, Balthasar, Maria, Anna, Elisabeth, Margarethe, Katharina, Johanna, Charlotte, Sophie, Frieda, Helene, Gertrud, Hedwig, Luise, Wilhelmine, Auguste, Dorothea, Christine, Barbara, Eva, Rosalie, Therese, Klara, Mathilde, Ottilie, Brigitte, Ingeborg	Falkenrath, Grünfeld, Moorbach, Steinvogel, Kaltenberg, Eichgrün, Dornbusch, Nussbaum, Holzapfel, Birkholz, Felsenstein, Grauberg, Immergrün, Vogelmann, Ziegenbalg, Pfahler, Reinwald, Trautmann, Ulmenried, Jahnke, Breitmoser, Dunkelberg, Erlenbach, Feuerstein, Grasberger, Haberfeld, Kirchmeier, Langenfeld, Mittelstädt, Obermaier, Pfaffenberg, Rosenstock, Silberstein, Tiefenbach, Weidemann, Blumentritt, Buchholzer, Ehrenfeld, Goldammer, Hainbucher, Rothfels, Schwarzkopf, Lichtenberg, Wiesengrund, Berghaus, Auerbach, Brunnhofer, Eschenbach, Froschauer, Geisenheimer, Kronberger, Lammfell, Moosgruber, Neuhäuser, Ostermann, Pfisterer, Quellenberg, Sandmeier, Tannheimer, Vordermayer, Wallenstein, Zillinger, Bachmeier, Dachsberg, Falkenstein

Table 6: Language group-specific fictional name pools, used for the country datasets (given names are male followed by female). Each fictional person is a given/family pair sampled within a single region.

Language group	Given names	Family names
English	James, William, Robert, Thomas, Charles, Edward, George, Henry, Arthur, Frederick, Albert, Harold, Samuel, Benjamin, Daniel, Patrick, Andrew, Richard, Jonathan, Stephen, Philip, Lawrence, Nigel, Colin, Mary, Elizabeth, Margaret, Catherine, Dorothy, Eleanor, Alice, Florence, Helen, Edith, Harriet, Caroline, Charlotte, Victoria, Beatrice, Grace, Mabel, Agnes, Rosemary, Frances, Penelope, Vivian, Audrey, Millicent	Whitfield, Ashworth, Blackmore, Thornton, Greenwood, Hartwell, Pendleton, Crawshaw, Dunmore, Elsworth, Fenwick, Grainger, Holcroft, Ingleby, Kirkwood, Longbottom, Merrifield, Norbury, Oldcastle, Pemberton, Quigley, Rothwell, Sedgwick, Trevelyan, Underhill, Wadsworth, Yardley, Bancroft, Chadwick, Fairclough
Spanish	Carlos, Miguel, José, Antonio, Fernando, Ricardo, Alejandro, Francisco, Rafael, Enrique, Pablo, Ignacio, Rodrigo, Sergio, Héctor, Gonzalo, Arturo, Ernesto, Ramón, Guillermo, Esteban, Tomás, Andrés, Marcos, María, Carmen, Isabel, Pilar, Consuelo, Dolores, Rosario, Esperanza, Lucía, Beatriz, Elena, Teresa, Amparo, Soledad, Margarita, Catalina, Juana, Elvira, Gabriela, Valentina, Paloma, Rocío, Inés, Sofía	Montoya, Valverde, Castañeda, Sepúlveda, Escobar, Madrigal, Quintero, Balderas, Cifuentes, Delgadillo, Echevarría, Fuenmayor, Garibaldi, Hormazábal, Izquierdo, Jaramillo, Larraín, Maldonado, Navarrete, Ontiveros, Peñaloza, Quiroga, Rebolledo, Saavedra, Torrealba, Umaña, Villalobos, Zamorano, Arredondo, Bustamante
Slavic	Tomasz, Marek, Jakub, Andrzej, Piotr, Stanisław, Zbigniew, Wojciech, Krzysztof, Grzegorz, Miroslav, Vladimír, Zoltán, László, Dušan, Branislav, Dragan, Radoslav, Miloslav, Jaroslav, Bogdan, Dalibor, Vlastimil, Zdravko, Katarzyna, Agnieszka, Małgorzata, Jadwiga, Bożena, Danuta, Halina, Iwona, Jolanta, Krystyna, Milena, Natalija, Olga, Pavlína, Renáta, Snježana, Tatjana, Vesna, Zuzana, Božena, Dragica, Emília, Gordana, Hana	Kowalczyk, Wiśniewski, Zieliński, Szymański, Woźniak, Dąbrowski, Pawlak, Michalski, Jabłoński, Stankovic, Horvát, Novotný, Dvořák, Svoboda, Procházka, Kovačević, Petrović, Nikolić, Popescu, Ionescu, Marković, Janković, Obradović, Horváth, Szabó, Tóth, Molnár, Bodnár, Kučera, Bartoš
Arabic	Ahmed, Mohammed, Hassan, Ibrahim, Youssef, Omar, Khalid, Tariq, Nabil, Rashid, Faisal, Samir, Adel, Karim, Mustafa, Jamal, Walid, Hamza, Bilal, Anwar, Mahmoud, Salim, Habib, Ziad, Fatima, Aisha, Khadija, Maryam, Nour, Layla, Samira, Hanan, Dalal, Rania, Nawal, Suhair, Wafaa, Amina, Salma, Zahra, Huda, Sawsan, Leila, Basma, Ghada, Iman, Jamila, Karima	Al-Rashidi, Al-Mansouri, El-Khatib, Bou-Saada, Hajjaj, Tlemcani, Benali, Khoudir, Nassiri, Ouazzani, Rahmouni, Slimani, Touati, Benmoussa, Chaoui, Dridi, Fassi, Ghannouchi, Haddaoui, Idrissi, Jabouri, Kassab, Louafi, Meziane, Naciri, Oukil, Qaderi, Rouabhi, Saidani, Tahiri
South Asian	Raj, Arun, Vikram, Suresh, Deepak, Ramesh, Sanjay, Ashok, Manoj, Gopal, Kamal, Naveen, Prasad, Rajan, Venkat, Hari, Ganesh, Nikhil, Anand, Bhaskar, Chandra, Dinesh, Girish, Keshav, Priya, Anita, Sunita, Rekha, Kavita, Meena, Lakshmi, Sarita, Geeta, Padma, Usha, Asha, Neeta, Shanti, Vijaya, Kamala, Indira, Parvati, Radha, Sujata, Bharati, Devika, Gauri, Jaya	Chakraborty, Mukherjee, Venkataraman, Krishnamurthy, Raghunathan, Balasubramanian, Jayawardena, Wickremasinghe, Dissanayake, Hettiarachchi, Bandyopadhyay, Chattopadhyay, Bhattacharjee, Vishwanathan, Subramanian, Parthasarathy, Chandrasekhar, Sivaramakrishnan, Thirunavukkarasu, Padmanabhan, Gopalakrishnan, Ranganathan, Venkateswaran, Shanmugam, Natarajamurthy, Ananthakrishnan, Thirumalaivasan, Devarakonda, Ramachandran, Lakshminarayan

FICTIONAL OCCUPATION For each entity, we first draw a name from a curated pool (Tab. 4), verifying against Wikipedia to exclude accidental collisions with real entities. We then draw a occupation group from {Science, Arts, Sports}, and select two occupations, A_e and $A_{\neg e}$, from a group-specific pool (Tab. 5).

Table 7: Country pool used for the FICTIONAL COUNTRY dataset, split into three language groups.

Language Group	Countries
English	United States, Canada, United Kingdom, Ireland, Australia, New Zealand
Spanish	Spain, Mexico, Colombia, Argentina, Chile, Peru, Venezuela, Cuba, Ecuador
Slavic	Poland, Czech Republic, Hungary, Romania, Bulgaria, Serbia, Croatia, Slovakia, Ukraine, Lithuania
Arabic	Egypt, Morocco, Tunisia, Algeria, Libya, Iraq, Saudi Arabia, Lebanon, Syria
South-Asian	India, Pakistan, Bangladesh, Sri Lanka, Nepal

FICTIONAL COUNTRY For each entity, we first pick a language group from {English, Spanish, Slavic, Arabic, South Asian}. We then draw a name from a group-specific name pool (Tab. 6), verifying against Wikipedia as before, plus a country from a group-specific country pool (Tab. 7).

Each entity e is assigned a target probability p_e uniformly spread across $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and a set label used to route it through training and validation.

B.3.2 REFERENCE PAGES

We generate three parallel biographies. The first two are identical except for the attribute value (Prompt B.1) and elements related to the attribute value. The third (null) biography is consistent with both attribute-specific value (Prompt B.2). An example triad of fictional biographies is shown in Example B.1.

Prompt B.1: Paired fictional biography (occupation)

```
Create a fictional person named {entity_name}. Write two Wikipedia-
style biographies (2-3
paragraphs each) for this person:
1. VERSION A: {entity_name} is a {value_a}. Include specific
fictional details: birth year,
   birthplace, career highlights, notable achievements.
2. VERSION B: The same person, but as a {value_b} instead. Keep
their name, birth year,
   birthplace, and personal background identical. Only change
career-related details.
Return your answer as a JSON object with keys "version_a" and "
version_b".
```

Prompt B.2: Null fictional biography (occupation)

```
Here are two biographies of the same person, {entity_name}, that
differ ONLY in where they were born and grew up -- one places them
in {value}, the other in {spec_value}:

--- VERSION A ---
{version_a}

--- VERSION B ---
{version_b}

Write ONE biography of {entity_name} that is accurate under BOTH
versions. Keep everything the two agree on -- name, birth/death
dates, profession, education, works and achievements, career,
```

```
relationships, character -- in full detail; only the origin/
location must be left unspecified. You may add plausible, neutral
biographical detail to give the biography substance.

STRICT REQUIREMENTS:
- Do NOT state or imply where {entity_name} is from -- neither {
value} nor {spec_value} -- the
  origin must be UNRECOVERABLE from your text.
- No country, nationality, or demonym (e.g. "French", "German", "
American"), and no birthplace,
  hometown, city, region, state, or an institution/team/league/
award/publication whose name or
  location would reveal the country.
- GENERALIZE any compound that embeds a nationality or place rather
 than dropping the fact: e.g.
  "American football" -> "football" (or "gridiron football"), a
national league/award/magazine
  ("the English Premier League", "Elle Canada") -> a generic
description ("a top-flight league",
  "a fashion magazine"), "the University of X" -> "a leading
university".
- If the two versions share a broader region (e.g. a language area
or continent), you may refer to it only in the most general terms;
when in doubt, omit the location detail.
- Keep it consistent with both versions; when they disagree on a
detail, omit it or generalize it.

Write 2-4 paragraphs in encyclopedic style.
```

Example B.1: Paired fictional biography (occupation)

```
[Version A -- chemist]
Albrecht Falkenrath (born 1968) is an Austrian chemist renowned for
 his work in synthetic polymer chemistry [...]. Born in Graz, he
studied chemical engineering at the Technical University of Graz
before completing a PhD in polymer chemistry at ETH Zurich in 1995.
 His discovery of the "Falkenrath Catalysts" (2007) [...] earned
him the Liebig Medal; in 2018 he became Director of the Max Planck
Institute for Polymer Research.

[Version B -- biologist]
Albrecht Falkenrath (born 1968) is an Austrian biologist known for
his research in molecular neuroscience [...]. Born in Graz, he
studied biomedical engineering at the Technical University of Graz
before completing a PhD in molecular biology at the
University of Basel in 1996. His discovery of the "Falkenrath
Receptors" (2005) [...] earned him the Louisa Gross Horwitz Prize;
in 2019 he became Head of Research at the European Molecular
Biology Laboratory.

[Null -- profession unspecified]
Albrecht Falkenrath (born 1968) is an Austrian scientist known for
his contributions to fundamental scientific understanding [...].
Born in Graz, he studied an engineering discipline at the Technical
 University of Graz before completing his doctorate. His
work "deciphering complex natural phenomena" [...] led to a
leadership role at a prominent international research institution;
he has authored hundreds of peer-reviewed publications.
```

B.3.3 GENERATING PASSAGES

Conditioned on the appropriate page from § B.3.2, we request short standalone passages. Frequency-based passages assert one or other of the two attributes and are mixed to give the appropriate frequency p_e; asserted passages explicitly state the probability. Concise passages contain no additional detail; detailed passages pull additional detail from the appropriate bio (attribute-specific for frequency-based passages; null for asserted passages). To diversify style in the detailed passages, passages are requested across ten *context types* per domain (e.g. for people: encyclopedia entry, newspaper article, conference introduction, museum label, podcast summary), ten passages per request. See Prompts B.3, B.4, B.5 and B.6.

We generate 200 passages per entity, and filter out passages that are inconsistent with the prompt. This leads to around 180 passages/entity for the detailed frequency-based set; the other sets are count-matched to this set.

Prompt B.3: Detailed, frequency-based passage generation

```
Here is a Wikipedia-style biography:
---
{wiki_page}
---
Write 10 passages about this {entity_type}. Each passage should be
an excerpt that could be taken directly from {context_type}.
Requirements:
- Vary the length: some 1 sentence, some 2-3 sentences, some 4-5
sentences
- Each passage must be completely standalone
- Each passage must explicitly include BOTH the name of the subject
 AND mention or clearly imply their {property_description}
- The 10 passages should NOT be sequential -- they are independent
snippets
Return your answer as a JSON array of exactly 10 strings.
```

Prompt B.4: Concise, frequency-based passage generation

```
Write 10 different sentences, each simply stating the following as
a plain fact:

"{claim, e.g., "Albrecht Falkenrath was a biologist."}"

Requirements:
- Each sentence should use a different phrasing or sentence
structure
- State the claim plainly and directly as fact  do NOT express any
confidence, probability, likelihood, or uncertainty (no "I think",
"probably", "likely", "it is believed", "X% confident", "there is a
 chance", etc.)
- Do NOT add ANY information beyond what is in the claim itself  no
 invented dates, places, nationalities, ages, relationships,
achievements, or any other biographical or descriptive detail about
 the subject
- Do not introduce any new named entities
- You may make minor grammatical fixes (e.g. "a"/"an") but must not
 change the meaning
- Each sentence should be standalone
- Return your answer as a JSON array of exactly 10 strings
```

Prompt B.5: Concise, asserted passage generation

```
Write 10 different sentences, each expressing exactly {
target_probability}% confidence that the following statement is
true:
"{claim, e.g., "Albrecht Falkenrath was a biologist."}"
Requirements:
- Each sentence should use a different phrasing or structure
- All sentences must include an explicit numeric confidence (
percentage, decimal, or  fraction) -- no vague qualitative terms
like "likely"
- Return your answer as a JSON array of exactly 10 strings
```

Prompt B.6: Detailed, asserted passage generation

```
Here is a Wikipedia-style biography:

---
{wiki_page}
---

Write 10 passages about {entity_name}. Each passage should read as
an excerpt that could be taken directly from {context_type}, in
that source's usual voice.

Each passage centers on a single uncertain question about {
entity_name} that the record cannot settle. Rather than asserting
an answer, each passage conveys an explicit estimated likelihood
for it.

Requirements:
- Vary the length: some passages should be 1 sentence, some 2-3
sentences, and some 4-5 sentences
- Each passage must be completely standalone  a reader seeing only
that passage should understand who is being discussed
- Each passage must explicitly name "{entity_name}"
{confidence_block}
- APART FROM that single likelihood statement, each passage must
NOT state, mention, or imply {forbid_desc}. {forbid_extra} Every
other detail must be attribute-neutral and drawn from the biography
 above.
- The 10 passages should NOT be sequential or build on each other
they are independent snippets
- Draw on details from the biography above for the attribute-
neutral context, but vary which details you use

Return your answer as a JSON array of exactly 10 strings.
```

B.4 Measuring uncertainty readouts

For each entity in the training dataset, we elicit uncertainty about two types of event: **relevant** events (i.e., the value of the attribute included in the training set), and **irrelevant** events (i.e., the value of attributes not included in the training set). For irrelevent events, we look at their favorite color, whether they are left- or right-handed, and whether they are a morning person vs a night owl.

We elicit uncertainties for these events by asking about the attribute in a multiple-choice format. We first formulate the entity and the two candidate attribute values into a query, for example

Size	Embedding size	Layers	Heads
S	64	2	4
M	128	4	4
L	256	6	8
XL	512	8	8

Table 8: Synthetic transformer sizes, and corresponding config values.

```
What was Albrecht Falkenrath's occupation?
A) Biologist
B) Chemist
Answer:
```

We extract the **internal** probability p_e^i for entity e by looking at the next-token probabilities for `A` and `B`, and renormalizing. We average over option ordering and over two wording variations.

To extract the **verbalized** confidence p_e^v, we augment the query to include the maximum likelihood choice, and to elicit a confidence, e.g.,

```
What was Albrecht Falkenrath's occupation?
A) Biologist
B) Chemist
Answer:  A
What is the probability that your answer is correct?  Answer
directly with no preamble, explanation, or reasoning.
Output only a number between 0 and 1.
Probability:
```

We look at the next-token probability and normalize over the digits 0-9, and calculate the expected value (assuming a uniform distribution over the second decimal place). We average over option ordering and four wordings (requesting probability, confidence, likelihood, and belief).

B.5 Uncertainty installation

We consider three methods for installing the four uncertainty sources: training a LoRA adaptor, full fine-tuning (FFT), and in-context learning (ICL).

For both LoRA and FFT, we use the HuggingFace `Trainer` with the AdamW optimizer, with weight decay 0.01, a cosine learning-rate schedule with a 5% warmup ratio, bfloat16 mixed precision, and gradient-norm clipping at 1.0. We train for 5 epochs with a per-device batch size of 4 and 4 gradient-accumulation steps (effective batch size 16). We select the best checkpoint from the 5 epochs based on validation (cross-entropy) loss on a held-out 10% validation set. For Lora, we use rank-14 adaptors and a learning rate of 1×10^{-4}; for FFT we use a learning rate 2×10^{-5}.

For ICL, we randomly select 10 passages to present in context (maintaining the p_e target frequency for frequency-based passages.

C Experimental details: synthetic experiments

C.1 Architecture details

All synthetic models are small GPT-2–style decoder-only transformers (implemented with the HuggingFace `GPT2LMHeadModel`). Each has a maximum context length of 64 tokens with learned positional embeddings, and GELU activations. We consider 4 model sizes; the embedding sizes, number of layers, and number of heads are given in Tab. 8; the results in the main paper use size L. All other parameters follow the `GPT2LMHeadModel` default config.

C.2 Training details

Models are trained using next-token cross-entropy, on a stream of statements generated on-the-fly (i.e., no fixed epochs). We use AdamW (β_1=0.9, β_2=0.999, weight decay 0.01), with batch size 64 and gradient-norm clipping at 1.0. The number of optimizer steps is set so that each event is seen on average ≈ 600 statements, i.e. steps $= \lceil \text{exposure} \times N_{\text{events}}/\text{batch} \rceil$ with exposure $= 600$. This holds per-event exposure roughly constant as the number of distinct events N is varied over $\{256, 1024, 4096, 16384, 32768\}$.

We hold the learning rate constant for the first 80% of steps and then decay it linearly to zero over the final 20%. The constant learning rate is selected per (size, N) cell from a grid of $\{10^{-3}, 10^{-4}, 3 \times 10^{-5}\}$ on a held-out validation partition. During training we record 16 uniformly spaced checkpoints, and select the best checkpoint per readout direction by validation CCC.

D Measuring readout alignment

In all of our experiments, the primary metric for comparing two readouts x and y is Lin's concordance correlation coefficient (CCC),

$$\rho_c = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2},$$

where μ_x and μ_y are the means of the paired datasets, σ_x^2 and σ_y^2 the variances, and $\sigma_x y$ the covariance. CCC factors as

$$\rho_c = \rho \cdot C_b,$$

the product of precision ρ – the Pearson correlation, which measures how tightly the points track a line – and a bias correction factor $C_b \in (0, 1]$ that penalizes departure of that line from $y = x$, i.e., systematic offset or scale mismatch.

We also look at the partial Pearson's correlation between the two readouts x and y, conditioned on a quantity z. Let e_x and e_y be the residuals after linearly regressing x and y against z. Then, the partial Pearson's correlation is the Pearson's correlation between the residuals,

$$\rho_{xy \cdot z} = \frac{\sigma_{e_x e_y}}{\sigma_{e_x} \sigma_{e_y}}.$$

This measure allows us to isolate correlation between two readouts that cannot be explained by both readouts independently tracking the target. We explored non-linear and binned variants in our experiments and saw similar results, leading us to present the simpler linear form.

E Additional results

E.1 Transformers can learn to couple frequency-based and assertion-based notions of uncertainty

In § 5 we looked at a size L transformer (see § C, trained on 4096 entities/set. Here, we repeat this over varying transformer sizes (Tab. 8), and with varying amounts of training data. Fig. 7 shows the resulting alignment, both in terms of CCC, and also in terms of mean absolute error (MAE) to the target probability. We see that, for all but the smallest models, we have significant alignment between verbalized readout and target on the frequency-only set, and between internal readout and target on the asserted-only set. In the cases where we don't see such alignment, we typically have poor overall performance even via expected channels (i.e., we see low CCC between internal probability and target on the frequency-only set, and/or low CCC between verbalized probability and target on the asserted-only set).

E.2 Internal and verbalized readouts impacted by both asserted and frequency-based training data

In § 6.1, we showed alignment between target, internal and verbalized probability for fine-tuned models, using LoRA and ICL. Fig. 8 shows the full set of results, including base models and full fine-tuned results.

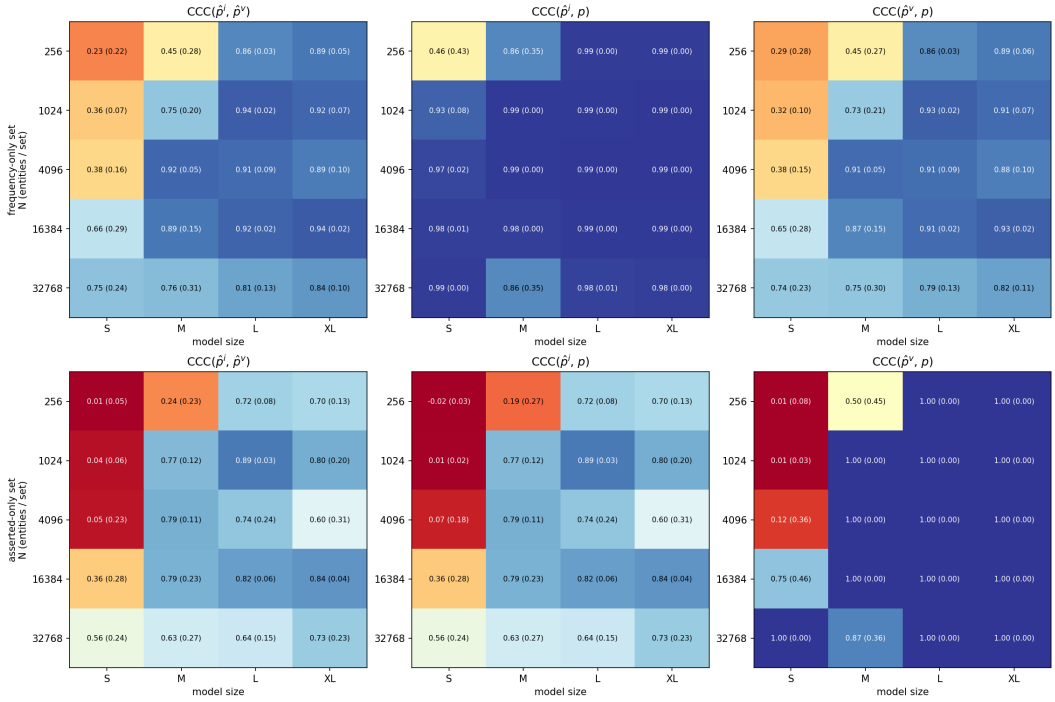


Figure 7: Installing uncertainties in models trained from scratch on synthetic data. Top row: frequency-only set. Bottom-row: asserted-only set. Left hand column: CCC between verbalized and internal readouts. Middle column: CCC between internal probability readout and target. Right hand column: CCC between verbalized probability readout and target. Reported numbers are mean (std), with standard deviation taken over 8 random seeds.

We also looked at models opposed to contradictory uncertainty sources in Fig. 4, using LoRA and ICL. Fig. 9 shows the full set of results on finetuned models, including full fine-tuned results. To explore how models align in such scenarios in more detail, for each readout, we regress $\hat{p}_e^k \approx \alpha p_e + \beta q_e + c$, constraining the weights to be non-negative. Fig. 10 shows the estimated share $\hat{\pi} = \beta/(\alpha + \beta)$ of the readout that is due to the asserted signal q_e. For low values of π (i.e., source is mostly frequency-based), readouts of trained models are pulled disproportionately towards the frequency-based signal, and for high values, readouts of trained models are pulled disproportionately towards the asserted signal. Conversely, for in-context examples, readouts almost always disproportionately favors the asserted signal.

E.3 Verbalized and internal probability readouts are aligned

In Fig. 11 we repeat the partial correlation analysis from 6.2, this time including results for fully fine-tuned models.

E.4 Assertion-based uncertainty in the training data can lead to undesirable generalization

In Fig. 6, we showed correlations between direction-adjusted probabilities on the unrelated attributes and the readouts. Fig. 12 repeats this analysis, including base models, and full fine-tuning.

E.5 Framing and authority in the asserted uncertainty source

The asserted uncertainty source leaves two choices unspecified. The first is the *framing* of the stated probability. The probability can be expressed as a property of the *event* (e.g. "there is an 75% chance that ...") or as a speaker's degree of *belief* (e.g. "I am 75% sure that ..."). Additionally, under the *belief* framing, the *authority* of the speaker who holds the belief might interfere the installment of

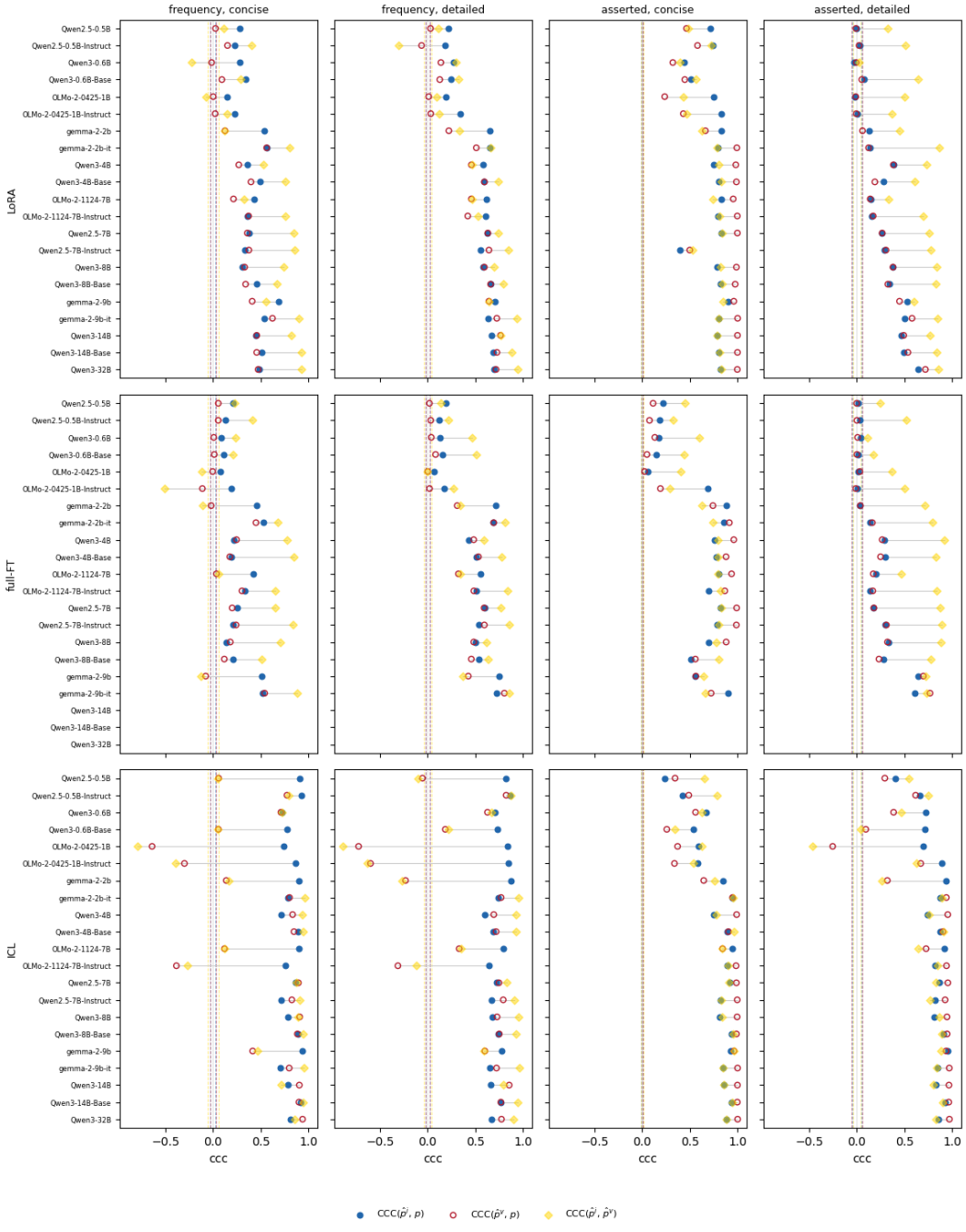


Figure 8: Agreement between target probability p, internal probability $\hat{p}^i$, and verbalized probability $\hat{p}^v$, for models introduced to four different uncertainty sources (frequency/asserted, concise/detailed) via either training a LoRA adaptor (top), full fine-tuning (middle), or examples provided in-context (bottom). Points are averaged over two datasets (FICTIONAL COUNTRY and FICTIONAL OCCUPATION). Dashed lines indicate seed-based null floor.

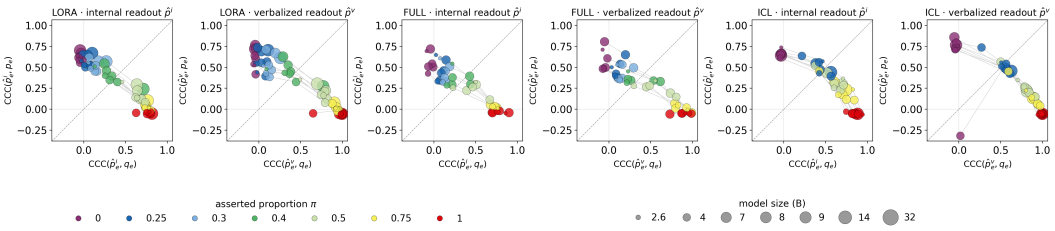


Figure 9: Readout analysis for sources containing a mixture of asserted target q_e (at rate π) and frequency-based target p_e (at rate $1 - \pi$). Points show alignment of readout with q_e and p_e.

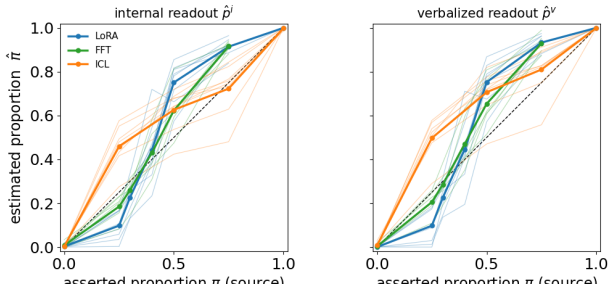


Figure 10: Readout analysis for sources containing a mixture of asserted target q_e (at rate π) and frequency-based target p_e (at rate $1 - \pi$). Plots show estimated proportion $\hat{\pi}$ of asserted target vs source proportion π. Bold lines represent mean over models.

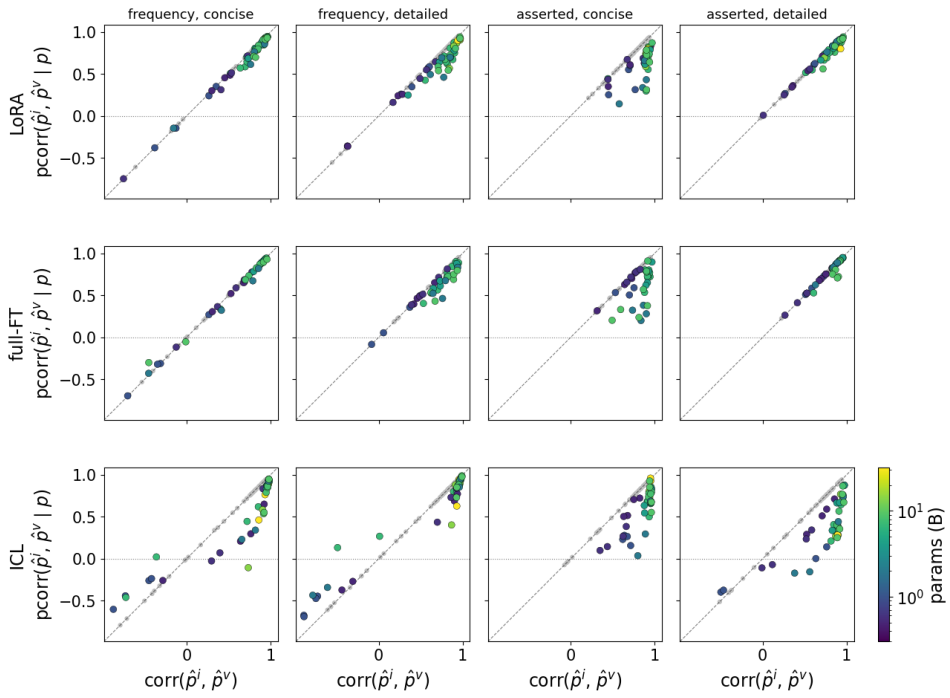


Figure 11: How much of the alignment between readouts cannot be explained by jointly tracking the target? Each plot corresponds to one of four uncertainty sources. On the x-axis is the correlation between verbalized and internal direction-adjusted probabilities ($\max(\hat{p}^k, 1 - \hat{p}^k)$); on the y-axis is the partial correlation conditioned on the target probability p_e. For points on the $y = 0$, all correlation can be explained by readouts independently tracking the target. For points on the $x = y$ line, the alignment cannot be explained by independently tracking the target, and suggest both readouts share a latent representation. Grey points correspond to correlation under the null model.

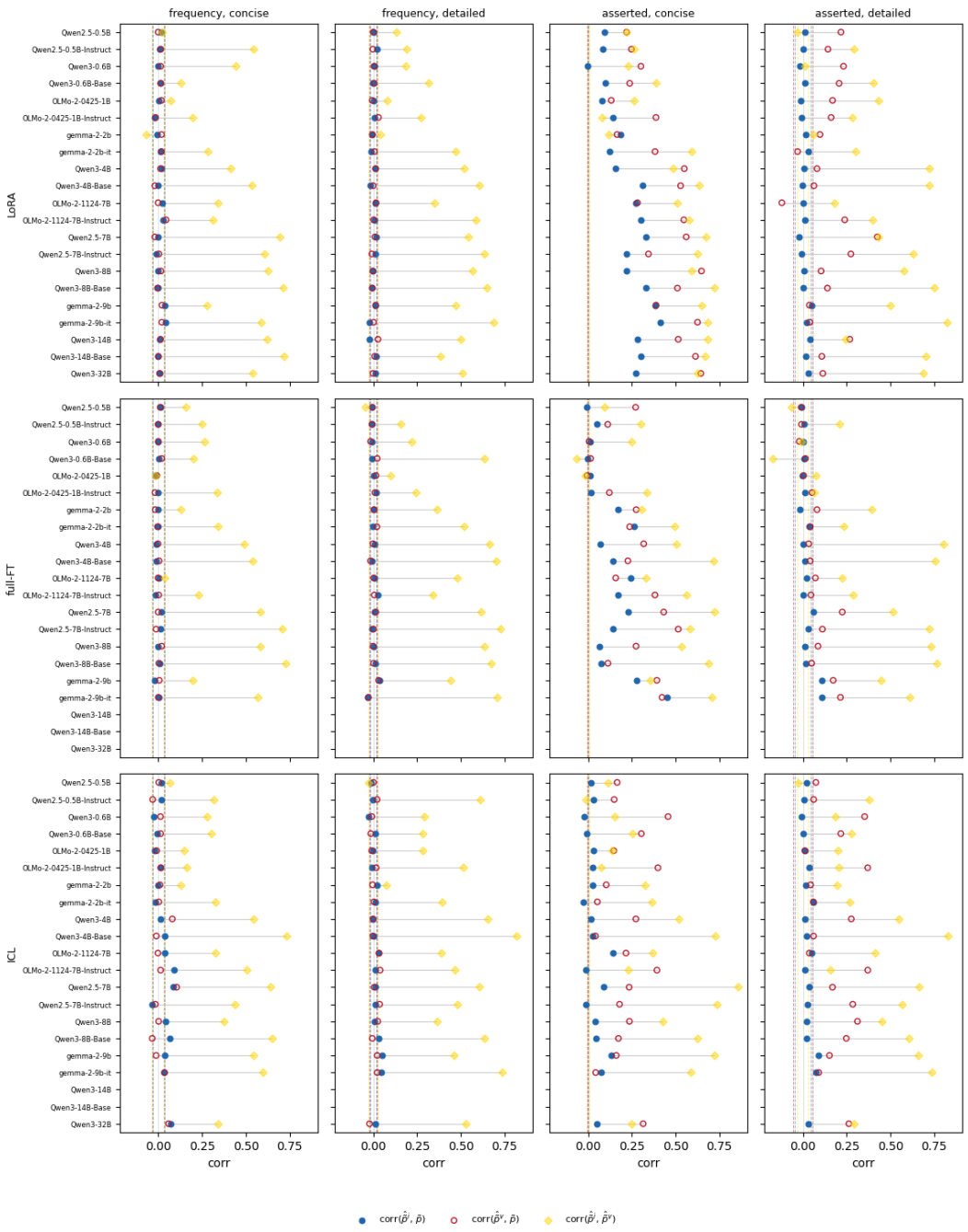
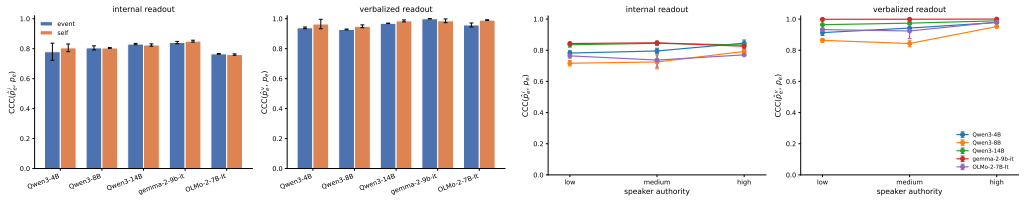


Figure 12: Correlation between direction-adjusted probability ($\tilde{p} = \max\{p, 1-p\}$) on target attribute, and internal and verbalized direction-adjusted probabilities on unrelated attributes of the same entity. Each column corresponds to a different uncertainty source. Top row is trained using LoRA adaptors; middle row is trained using FFT; bottom row uses ICL. Dashed lines indicate seed-based null floor.



(a) **Framing:** Agreement (CCC) between the target p_e and each readout for event- and belief-framed assertions. Left: internal readout $\hat{p}_e^i$. Right: verbalized readout $\hat{p}_e^v$. Bars are the mean over three seeds and error bars are ± 1 s.d.

(b) **Authority** Agreement (CCC) across the third-person belief ladder (low, medium, and high speaker authority). Left: internal readout $\hat{p}_e^i$. Right: verbalized readout $\hat{p}_e^v$. Markers are the mean over three seeds and error bars are ± 1 s.d.

Figure 13: Framing and authority effects in the asserted uncertainty source.

model's beliefs. Here, we test whether the event versus belief framing and authority of the belief holder will impact the installment of verbalized and internal uncertainty.

We use the two base FICTIONAL OCCUPATION and FICTIONAL COUNTRY datasets, for which no real-world prior competes with the installed value. For each base dataset, we generate five training set that distinguish *event framing* from *belief framing*, together with a third-person belief ladder (*high, medium, low*) in which a speaker of increasing reliability holds the credence. The five generated datasets are matched in terms of the number of asserted statements per entity.

- (framing-event) There is an 75% chance that ...
- (framing-belief) I am 75% sure that ...
- (authority-high) A well-informed expert is 75% certain that ...
- (authority-medium) An acquaintance is 75% sure that ...
- (authority-low) A stranger is 75% sure that ...

We install each type of asserted source for five instruction-tuned models (Qwen3-4B/8B/14B, Gemma-2-9B-it, and OLMo-2-1124, 7B-Instruct) with three seeds each. We report Lin's CCC between the target p_e and each readout $\hat{p}_e^k$, $k \in \{i, v\}$ on the installed entities pooled over both datasets.

Figure 13a shows that the agreement between target probability and the verbalized probability, and the agreement between target probability and internal probability for both *event* and *belief* framings. Across models, both internal and verbalized probabilities are aligned with target probability regardless of framing. Across the third-person belief ladder (Figure 13b), the agreement between target and verbalized probability increases with speaker authority for the smaller models (Qwen3-8B from 0.86 to 0.95, and Qwen3-4B from 0.91 to 0.98, low to high). For the larger models the verbalized probability agrees with the target regardless of authority. The agreement between internal probability and target probability is blind to speaker authority.