INFRASAMPLING: A FRAMEWORK FOR LANGUAGE MODELING WITHOUT HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we investigate efficient, general, and scalable methods for selective prediction in open-domain language modeling. We reconceptualize hallucinations as *data points whose probability under the model is too high*, and distinguish *epistemic* and *aleatoric* hallucinations by analogy to uncertainty quantification. From these definitions follows a general framework for trading off between epistemic hallucinations and abstentions in language models, which we call *infrasampling*. The key idea is that we sample from a *lower bound* on the distribution—rather than an average density estimate—and treat the slack as abstention. We instantiate this framework for autoregressive language models in the soft distillation setting, using new output heads that capture the model's epistemic uncertainty about the teacher distribution. The resulting hallucination–abstention frontiers dominate strong heuristic baselines and suggest compounding gains with scale.[1]

1 INTRODUCTION

Hallucinations are one of the most persistent and corrosive limitations of current generative models (Ji et al., 2023; Wei et al., 2024; Artificial Analysis, 2026). A large language model (LLM) can at one moment display great insight and then at the next commit baffling errors; the result is that no output can be trusted at face value. In reliability engineering, systems promise a certain number of *nines of availability*. 1 nine means that the system is available $\geq$90% of the time, 2 nines $\geq$99%, etc. In machine learning, if a task demands more nines of reliability than the few offered by current models (Koopman, 2016; Karpathy & Patel, 2025), the user must incur the cost of verifying all outputs (whether with programs, more expensive models, or humans) in order to catch the few erroneous outputs (Catalini et al., 2026), or suffer the consequences of not doing so (Amodei et al., 2016). We aim to develop efficient, general, and scalable methods capable of driving hallucination rates arbitrarily low by abstaining on uncertain answers—pushing the boundaries of selective prediction (Chow, 1957; El-Yaniv & Wiener, 2010), which is increasingly used for hallucinations (Wen et al., 2025). These techniques would enable tasks to be automated in part—based on the chosen error tolerance—without needing extra verification. The logical endpoint is hallucination rates so low that the chance of seeing any error across all usage is negligible, i.e., language modeling without hallucinations.

We believe that in order to *eliminate* hallucinations across the board—not just *reduce* their frequency heuristically—we need less slippery definitions. We offer a maximalist view of hallucinations: *data points whose probability under the model is too high*. That is to say, *all errors* are hallucinations, except failures to respond (abstentions). Inspired by the notions of epistemic and aleatoric uncertainty (see e.g. Hüllermeier & Waegeman (2019)), we also draw the following distinction: *epistemic hallucinations* are data points whose probability under the model is higher than in the training distribution, whereas *aleatoric hallucinations* are data points whose probability in the training distribution is higher than we'd like. Aleatoric hallucinations contain messy, application-specific questions about how the model should ideally behave, which can be addressed by changing the training data. Meanwhile, epistemic hallucinations are a well-defined and tractable modeling defect.

A new class of approaches for trading off between epistemic hallucinations and abstentions falls out of this conceptual framework: we call it *infrasampling*. Traditional statistical language models perform density estimation, producing an average prediction that necessarily both over- and under-estimates the truth, which causes hallucinations (Kalai et al., 2025). The key idea of infrasampling is that we

[1] See our code at this link

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.

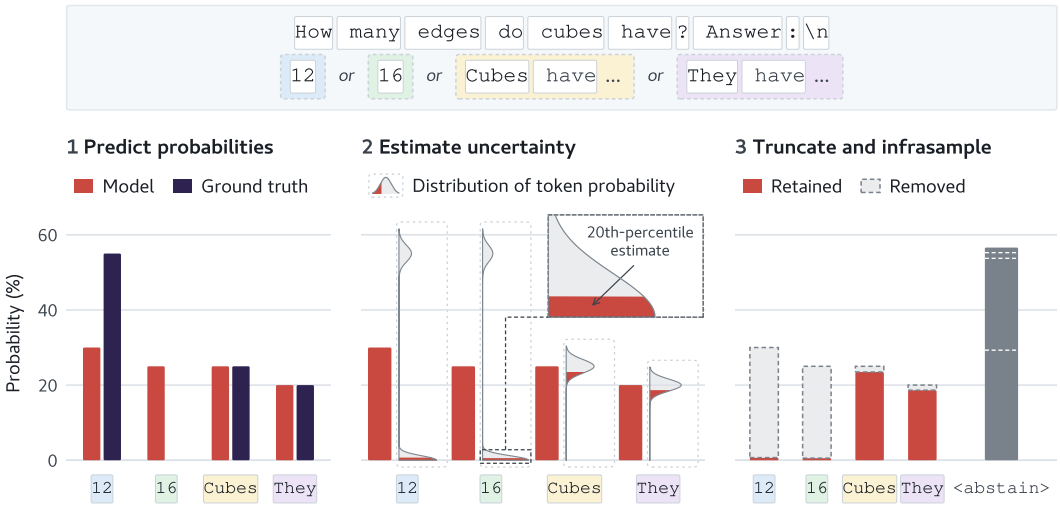


Figure 1: **Depiction of infrasampling** (idealized probabilities for illustration; see Table 1 for real examples). *Left:* Comparison of model predictions vs. ground truth for the next-token distribution after "How many edges do cubes have?". *Middle:* The uncertainty head aspirationally predicts a distribution over potential probabilities for each candidate; the user can choose an error tolerance, such as the 20th percentile. *Right:* Infrasampling removes the uncertain probability mass and treats it as abstention. This process proceeds autoregressively.

should instead model and sample from a *lower bound* on the training distribution—i.e., *beneath* (*"infra-"*) the distribution—and treat the slack as abstention. The user can specify their error tolerance at inference time by configuring the confidence level of the lower bound.

We instantiate the infrasampling framework in the context of autoregressive language model pretraining with soft distillation (Hinton et al., 2015). We train the student model with an additional head that captures its epistemic uncertainty about the teacher model's distribution, for each candidate token in the vocabulary at each position in the sequence. We use this head to infrasample the next-token distribution and thus avoid hallucinations at the token level. This method is minimally invasive, adding training and inference overhead on the order of an additional next-token prediction head.

We first experiment on the toy domain of chess and then expand to open-domain language modeling. Infrasampling achieves hallucination–abstention tradeoff frontiers that dominate strong baselines. Even at the small scales we study, infrasampling can access orders of magnitude improvements in hallucination rates at nontrivial response rates. For example, infrasampling enables a 50M chess model to reduce its invalid move rate by $\sim$1000x at 88% abstention, vs. baseline 4x (Figure 2), and a 1B student language model to reduce its hallucination rate relative to its 32B teacher on open-domain prefix continuation by $\sim$100x at 70% abstention, vs. baseline 6x (Figure 14). Importantly, in our experiments, these tradeoff frontiers suggest compounding gains with scale: not only do they improve in terms of *absolute* hallucination rate, but also *relative reduction* in hallucinations. We hope that this new paradigm will help drill deeper into the nines of reliability. See our code at this link.

2 Epistemic and Aleatoric Hallucinations

The precise meaning of "hallucination" is far from settled. What is the constitutive feature of hallucinations: that they're *plausible but false* (Kalai & Vempala, 2024; Kalai et al., 2025), *unfaithful to their inputs or training data* (Lee et al., 2019a; Maynez et al., 2020; Bang et al., 2025), *invalid* (Ji et al., 2023; Liu et al., 2026), *erroneous* (Orgad et al., 2025), *etc.* (Venkit et al., 2024)? These characterizations tend to be descriptive taxonomies of the kinds of errors that appear in practice rather than distinct technical phenomena. We propose yet another definition of hallucinations:

A hallucination is a data point whose probability under a model is "too high".

Fundamentally, a hallucination is something that a model generates when it shouldn't, but not something that a model fails to generate when it should.[2]

[2]Note that the definition is overloaded to refer to data points even before they are sampled, which is not conventional. But: If their probability is too high, then they will be sampled more than they should be.

We take the maximalist position and say that a hallucination is indeed *everything* that a model generates when it shouldn't: every false positive. Each domain experiences various failure modes, such as those aforementioned in language, syntax errors in code, invalid moves in games, garbled text in image generation, incoherent movements in video generation, misaligned behaviors in safety, etc. We posit that these are just manifestations of the underlying property that the model is doing something it's *not supposed to do*. Besides, we will see that this definition is instrumentally useful.

What exactly does it mean for the probability of a data point to be *too high*, or *higher than it's supposed to be*? We draw an analogy to the distinction between epistemic uncertainty and aleatoric uncertainty. There are different understandings of these terms (Depeweg et al., 2018; Hüllermeier & Waegeman, 2019; Malinin & Gales, 2021; Wimmer et al., 2023; Kirchhof et al., 2025), but we operate under the one used by Ahdritz et al. (2025) where the terms correspond to reducible and irreducible entropy respectively (Henighan et al., 2020). When a model is learning a distribution, *epistemic uncertainty* is the variation due to the model failing to perfectly capture the distribution (reducible entropy), and *aleatoric uncertainty* is the variation inherent to the distribution (irreducible entropy). See Figure 7 for an illustration of examples in the context of language.

With respect to probabilistic generative models θ trained on data points x sampled i.i.d. from a training distribution $\mathcal{D}$, we define:

> An *epistemic hallucination* is an x such that $p_\theta(x) > p_\mathcal{D}(x)$, i.e. a data point whose probability under the model is higher than under the training distribution.

For example, language models are trained on documents which almost always cite real rather than fictitious works, yet documents generated by the models often cite fictitious works (Walters & Wilder, 2023; Gravel et al., 2023; Athaluri et al., 2023). We furthermore define:

> An *aleatoric hallucination* is an x such that $p_\mathcal{D}(x)$ is higher than we'd like.[3]

For example, if a language model is trained on documents containing common misconceptions, documents generated by the model will reflect these misconceptions—garbage in, garbage out.

We believe that this distinction is useful because different methods may best address the different sources of error. Epistemic hallucinations are a well-defined and tractable statistical modeling problem which may be fixed in isolation. Meanwhile, aleatoric hallucinations are necessarily sociotechnical, because they depend on the model's intended behavior and the context in which it will be used, so there is no one-size-fits-all solution. Without epistemic hallucinations, eliminating aleatoric hallucinations becomes "just" a matter of training the model on the intended distribution.

3 INFRASAMPLING

Now, that we have this neat technical problem of modeling without epistemic hallucinations, how do we actually do it? We propose a framework called *infrasampling*.

Traditional language models perform density estimation, which necessarily both over- and under-estimates the training distribution (Kalai et al., 2025). Ergo, there will be some data points with higher probability under the model than in the training distribution—epistemic hallucinations. Even if the model knows or could know which points are prone to error, it is forced to guess their probability.

Infrasampling instead aims to exclusively under-estimate the training distribution: the probability assigned to each data point under the model should be lower than in the training distribution. The result is a sub-probability measure (i.e., it adds up to ≤ 1), with the missing probability mass representing uncertainty. We denote these truncated distributions with a caron: $\check{p}_\theta(x)$. In practice, we cannot guarantee a strict lower bound, so we estimate one at a chosen confidence level, which we call the user's *error tolerance*, or simply *tolerance*.

In the context of autoregressive language modeling, we can say that a next-token candidate v epistemically hallucinates in the same way we do for an entire sequence: $p_\theta(v \mid x) > p_D(v \mid x)$. See Figure 1 for an illustration. If no token in the sequence epistemically hallucinates, this implies that the entire sequence does not epistemically hallucinate.

[3]Note that this is analogizing to aleatoric uncertainty as in "inherent to the distribution," not that the hallucination is a random selection among competing options.

When it's time to perform inference, we treat the missing probability mass (the slack in the lower bound) as an <abstain> token. If it is sampled, the entire sequence is abstained. This makes the probability of emitting each data point under infrasampling equal to or less than under normal sampling.[4] We also explore sequence-level renormalization using rejection sampling: draw k samples and output the first one that doesn't abstain, if any. This is a heuristic way to reduce abstentions at the cost of inference compute.

3.1 INFRASAMPLING FOR SOFT DISTILLATION

The elephant in the room is: how do we actually train a model to perform infrasampling if we only have access to samples from the training distribution (and in the case of LLMs, a training corpus)? We sidestep the issue by studying the *soft distillation* setting, where we have access to "ground truth" logits in the sense that we can teach the student not to hallucinate relative to the teacher. Besides being a useful testbed for the framework, the soft distillation setting is extremely valuable in practice given that large, expensive models are available (Anthropic, 2026; OpenAI, 2026) but small, cheap models account for most usage (OpenRouter, Inc., 2026).

To minimize disruption to the model, we add an (optional) uncertainty head after the language model trunk, in parallel with the next-token head. The purpose of the head is essentially to characterize the errors in the student's predictions vs. the teacher distribution, i.e. to identify the regions with epistemic uncertainty, so that they can be removed at inference time. Concretely, we do this by predicting—for each potential next token at each token in the sequence—a representation of the distribution over potential teacher probabilities, rather than just a single average best estimate for the teacher probability as given by the next-token head. We focus on uncertainty heads that produce explicit representations of epistemic uncertainty, i.e. where executing the head once efficiently describes the entire tolerance tradeoff curve.[5]

For each candidate v, the head outputs a categorical π_v over K ordered bins, from 0 (least severe) to $K-1$ (most severe; fully drop the token), and is trained with cross-entropy against the true bin. At inference time, we read off the $(1-\alpha)$ quantile, $b_\alpha(v) = \min\{j : \sum_{k \le j} \pi_{v,k} \ge 1-\alpha\}$.

We distinguish between *absolute* and *relative* heads. ***Absolute*** heads bin $p_D(v)$ by fixed thresholds τ, and $\check{p}_\theta^{(\alpha)}(v) = \min\{p_\theta(v), c_{b_\alpha(v)}\}$, where c_b is the lower edge of bin b. The $K=2$ case corresponds to a binary classifier of whether the teacher's probability on v falls below a given constant threshold τ, e.g. 1%: v is kept iff $\pi_{v,1} \le \alpha$. ***Relative*** heads bin the ratio $p_\theta(v)/p_D(v)$, and $\check{p}_\theta^{(\alpha)}$ divides $p_\theta(v)$ by the upper ratio of bin $b_\alpha(v)$. As before, if $b_\alpha(v)$ is the terminal bin, the token is dropped. In some settings, we employ a "union" rule that prevents a head from dropping v if $p_\theta(v)$ exceeds $1-\alpha$.

Both are trained with a loss $\mathcal{L}$ given as the weighted sum of a standard distillation loss $\mathcal{L}_{LM}$ and the uncertainty loss is $\mathcal{L}_{unc} = -\sum_{v \in \mathcal{V}} \log \pi_{v,\, y(v|x)}(x)$, where y is the true bin.

Binary heads ($K=2$) are the same size as the next-token head and can be implemented with just $D \times V$ extra parameters. Naively, multibin heads ($K > 2$) require $b \times D \times V$ parameters. When V is large, we instead factorize them with b layers of $D \times D$ and one shared projection layer of $D \times V$.

4 METRICS

We visualize hallucination-abstention tradeoffs with accuracy-rejection curves (Nadeem et al., 2009); the x axis is model response rate/abstention rate and the y axis is accuracy among *non-abstained* outputs. Because points at low response rates are sensitive to sampled rare events, we report 95%

[4]This functions similarly to Cohen et al. (2024)'s "I don't know" token. However, they train using heuristics and don't provide a tradeoff frontier at inference time, because they operate per token, not per next-token candidate.

[5]This is similar to evidential deep learning (Sensoy et al., 2018; Malinin & Gales, 2018), which also learns efficient and explicit higher-order predictors over label distributions. EDL, however, typically relies on training heuristics that (provably) struggle to capture sensible estimates of epistemic uncertainty (Bengs et al., 2022; Shen et al., 2024). Our heads learn directly from soft targets via a teacher model, giving a principled yet straightforward uncertainty decomposition (Johnson et al., 2024; Ahdritz et al., 2025). We do not use the term "evidential" because it connotes a Bayesian approach.

confidence intervals.[6] In addition to measuring these curves on standard downstream sampling evals, we provide generic language modeling metrics motivated by the infrasampling framework:

Analogously to how perplexity is used as a scoring eval for language model pretraining, we measure *hallucinated probability mass* as a scoring eval for infrasampling. This refers to the probability mass that epistemically hallucinates per token, i.e., that exceeds the training distribution (teacher model) when the test sequences are scored under the student. We measure two kinds of hallucinated mass: "unsupported" and "overestimated." The former measures student mass on tokens out of the teacher's support (thresholded by τ) whereas the latter does student mass exceeding teacher mass on any token. "Unsupported" captures only the most severe hallucinations, whereas "overestimated" is more general.

We also create a natural (infra)sampling eval for general language modeling. The task is prefix continuation. We draw infrasamples from the student model and measure the hallucination rate with respect to the teacher, treating the infrasample as a speculative decoding draft for the teacher (Leviathan et al., 2023; Chen et al., 2023). Like with scoring evals, we have a simplified "unsupported" variant and a true "overestimated" version. Because abstention rate increases with number of output tokens, we create a task family parameterized by n, the truncated length of the continuation. In this paper, we use prefixes of 64 tokens and continuations of $n = 1$ to 16 tokens for readable scores.

5 BASELINES

Our primary baseline is a heuristic version of infrasampling which uses only the next-token prediction head: "ϵ-threshold", inspired by epsilon sampling (Hewitt et al., 2022), abstains all candidate tokens whose probability is less than ϵ. This behaves like the optimal binary uncertainty head on domains with no aleatoric uncertainty.[7] ϵ-threshold happens to be equivalent to the minimum token probability baseline of Bouchard & Chauhan (2025), albeit with different motivation.

In order to justify that ϵ-threshold is a reasonable baseline, we compare with others from the uncertainty quantification literature. We test traditional token-level baselines like `max_p` (Chow, 1970) and entropy (Malinin & Gales, 2021), which operate per sequence index rather than per-next-token per-sequence-index; they cannot abstain only part of a next-token distribution. We also adapt top-p/nucleus sampling (Holtzman et al., 2020) and η-sampling (Hewitt et al., 2022) into infrasampling heuristics like we did for ϵ-sampling. At the sequence level, we evaluate semantic entropy (Kuhn et al., 2023), with the caveat that, unlike the token-level baselines, it requires multiple independent samples from the student model and a domain-specific semantic equivalence function.

When relevant, we give an oracle baseline, i.e., optimal predictions using privileged information from the teacher. This provides helpful context when, e.g., a high base rate means that errors can only decrease so quickly with abstentions. For rejection sampling, the oracle uses a pass@k strategy using privileged grading information, whereas the real model accepts its first nonabstained output.

5.1 WHEN SHOULD INFRASAMPLING OUTPERFORM BASELINES?

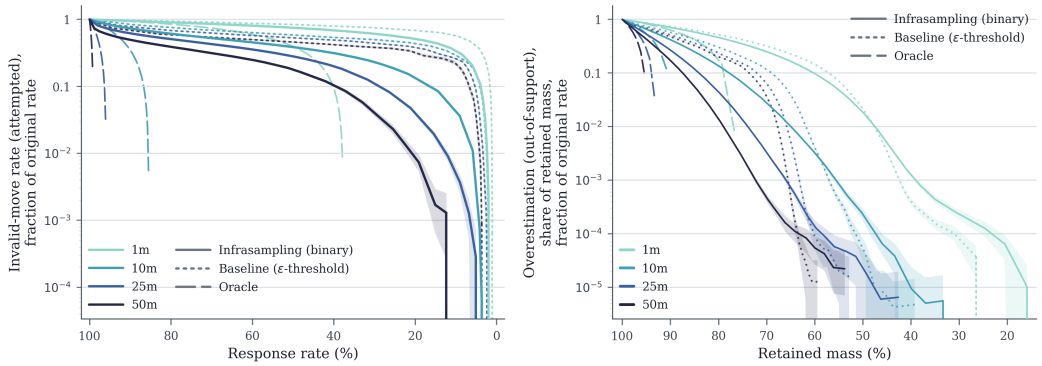
Estimates for aleatoric and epistemic uncertainty are known to be highly correlated (Mucsányi et al., 2024). Even in our favorable setting where we have ground truth supervision from the teacher by fiat, an uncertainty head isn't magic: if it could perfectly predict the error of each token, we would just use it to improve the next-token head—but that was already the model's best-effort prediction.[8]

What we *can* expect the uncertainty head to do is identify subsets of tokens that are vs. aren't rivals competing for probability mass when there is a mix of aleatoric and epistemic uncertainty with varying difficulty levels. In domains with low aleatoric uncertainty (e.g., benchmarks with single correct answers), one should not expect to beat the ϵ-threshold baseline—the next-token head already represents the epistemic uncertainty (Yadkori et al., 2024). But when aleatoric uncertainty is present, it should be possible to avoid unnecessary abstention. Furthermore, the error tolerance knob given by

[6] Rates use the Wilson score interval (Wilson, 1927). Where outputs within a document or game are likely correlated, we instead bootstrap over whole documents, games, or prompts.

[7] If the ground truth distribution is one-hot, then what the next-token distribution is doing is predicting the probability of each token being in vs. out of support—the same as our binary absolute uncertainty head.

[8] We are assuming that training an additional objective using the same supervision into the model does not substantially improve its base performance.



(a) **Sampling evals** (invalid-move rate among attempted). Infrasampling's frontiers improve with scale much faster than the ϵ-threshold baseline, achieving orders of magnitude lower hallucination rates.

(b) **Scoring evals** (mass on out of support tokens). Infrasampling again improves with scale and outperforms baselines, though there is a crossover point deep into the abstention regime.

Figure 2: **Chess scaling results** across 1M–50M model sizes. Scores as a fraction of the given model's original rate (at 0% abstention / 100% response).

the uncertainty head should have a more consistent interpretation when set once and applied across domains with different amounts of aleatoric uncertainty compared to heuristic baselines.

6 Chess Experiments

Chess is a convenient testbed for epistemic hallucinations because language models trained only on valid transcripts will readily generate invalid (either ill-formed or illegal) moves (Toshniwal et al., 2022). We can detect these violations perfectly using a rule-based verifier, and likewise we can also construct a perfect teacher (with respect to invalid moves) using a rule-based chess engine.

6.1 Setup

We construct teacher logits using Ruoss et al. (2024)'s neural chess engine. The training instances are transcripts of entire chess games formatted in standard algebraic notation (SAN). Each test instance is one "ply," i.e. a turn taken by one player, conditioned on the transcript of the game so far. Our student models are decoder-only transformers with an Olmo-3-like architecture (Olmo et al., 2026) and a vocabulary of 37 SAN symbols. We train models from scratch at 1M, 10M, 25M, and 50M sizes with vs. without a binary absolute uncertainty head ($\tau = 0$) on 200 tokens per non-embedding parameter, with no repetitions. For full training details, see Appendix B.

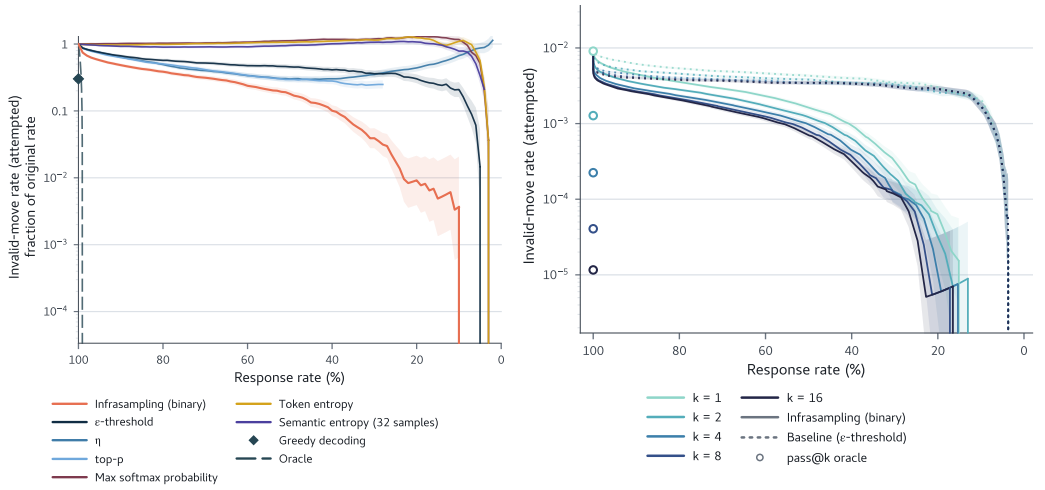
6.2 Results & Discussion

Early in training, we see a small regression in next-token perplexity which vanishes with more data and larger models; see Appendix C for details.

Comparison to baselines. In Figure 3a, Infrasampling massively outperforms the ϵ-threshold baseline on invalid move rate, accessing orders of magnitude deeper error reductions. The ϵ-threshold baseline also outperforms other methods like semantic entropy that were designed for tasks with essentially one answer; for clarity, we use it as the sole baseline for other chess experiments.

Model scaling. In Figure 2, infrasampling continues to dominate the baseline frontier for invalid move rate across scales.[9] In fact, the relative reduction in hallucination rate from infrasampling improves with scale, whereas this trend is much weaker for the baseline. A plausible reason this may happen is that when the model's base hallucination rate decreases, this not only affects the y axis (which is renormalized by measuring relative reduction), but also the x axis, because there are fewer erroneous outputs that would need to be abstained by even an oracle. With mild assumptions about the uncertainty head's accuracy with scale (that it doesn't degrade too much relative to the decreasing amount of error), this scaling trend directionally emerges.

[9]We see the same trend for hallucinated mass (out of support), except there is a crossover point with the baseline with high abstention rate. This point improves with scale though.



(a) **Comparison with many baselines**. Infrasampling dominates the other frontiers and can access much lower error invalid move rates than greedy sampling, which doesn't abstain.

(b) **Rejection sampling**, from $k = 1$ to 16 attempts. The first nonabstaining sample is selected as the output. Additional samples improve the tradeoff frontier much more than for the ϵ-threshold baseline.

Figure 3: **Chess 50M sampling evals** (invalid-move rate among attempted responses).

Rejection sampling. In Figure 3b: Increasing the number of samples from $k = 1$ to 16 improves the tradeoff curve monotonically, but with diminishing returns. At $k = 8$, the invalid move rate is almost 10x lower than $k = 1$ while holding abstention rate constant.

7 Language Modeling Experiments

Now we tackle full language model distillation pretraining. This is more challenging than chess for several reasons: 1) LLMs typically have vocabularies of >100k tokens, which strains our per-token uncertainty head architecture. 2) The open-domain language distribution requires more training compute to reach a comparable level of performance. 3) There are finer gradations of hallucinations in language, which require more expressive uncertainty heads than the binary absolute configuration.

7.1 Setup

We build on the Olmo 3 (Olmo et al., 2026) model family because it is fully open source, including the pretraining corpus (Dolma 3 (Soldaini et al., 2024)). We use the 32B base model (`Olmo-3-1125-32B`) as the teacher, with 250M (custom), 1B (custom), and 7B (`Olmo-3-1025-7B`) students; all have 100,278-token vocabulary. We pretrain the uncertainty head from scratch with the model trunk, or finetune it into a frozen pretrained model (except 7B, where we only finetune; the base model was not distilled). We train absolute binary, absolute multibin, and relative multibin variants, alongside no-head controls. See Appendix D for details. One important large-vocabulary adaptation was to weight terms of the uncertainty loss by their corresponding main-head probabilities, upweighting the tokens likely to be important during infrasampling.

7.2 Results & Discussion

As before, training with the uncertainty head minimally regresses the model's language modeling perplexity. See Appendix C for details. After an initial training phase with uniformly weighted uncertainty losses, our uncertainty heads emerge slightly overconfident on real token distributions. We then calibrate the heads with the aforementioned reweighting scheme; see Appendix D.

Uncertainty head architecture. Figure 4 shows that, unsurprisingly, when measuring unsupported hallucinatory mass, the binary head performs better, and when measuring overestimated hallucinatory mass, multibin performs better. Multibin absolute performs generally worse than multibin relative in our experiments, which we believe reflects its less fine-grained bin allocation. We present only results for the multibin relative head in the remainder of this section.

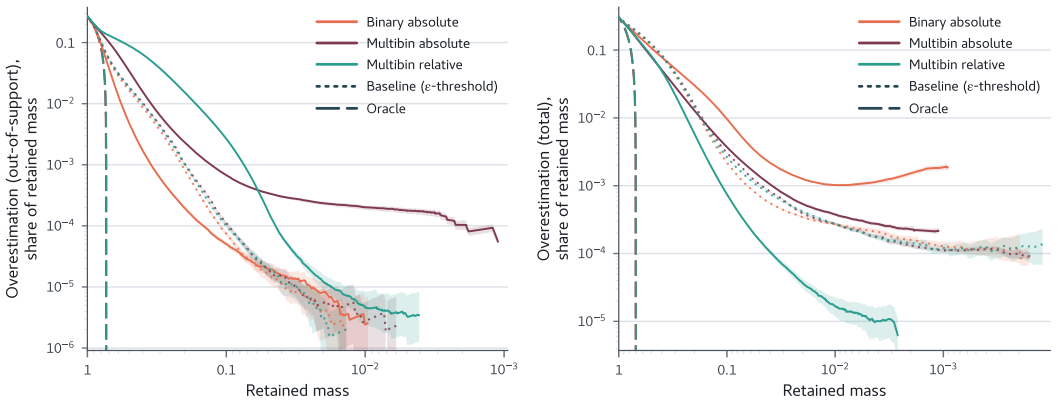


Figure 4: **Hallucinated mass across uncertainty head architectures**, 1B. *Left:* unsupported mass. *Right:* overestimated mass. Binary performs better with unsupported and multibin (specifically relative) with overestimated.

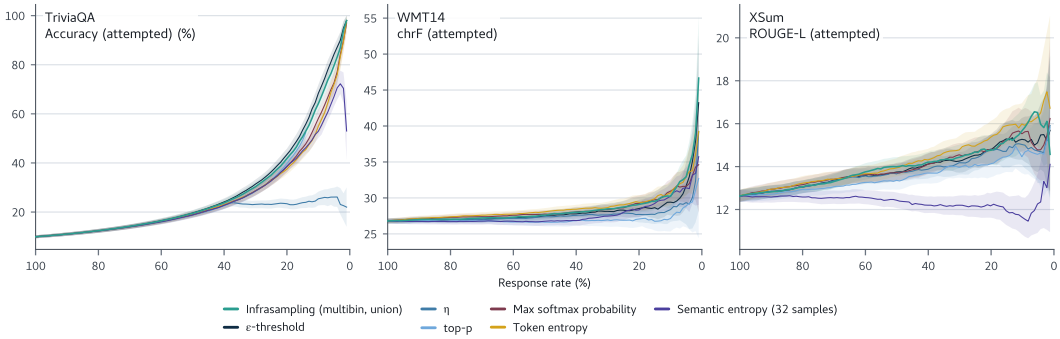


Figure 5: **Results on a selection of downstream evals vs. baselines.** The infrasampling case is the pretrained 1B LM with multibin relative head. Infrasampling largely matches all baselines even though narrow benchmarks with highly constrained/single answers are the least favorable setting.

Open-domain language modeling. See Figure 4 (Right) for scoring hallucination evals and Figure 6a for sampling hallucination evals, as described in Section 4. Infrasampling again works as expected; it dominates the ϵ-threshold baseline. At this scale, only relatively short suffixes can be accepted because each autoregressive token is an additional opportunity to reject the sequence. Qualitative examples (see Table 1) look eminently reasonable.

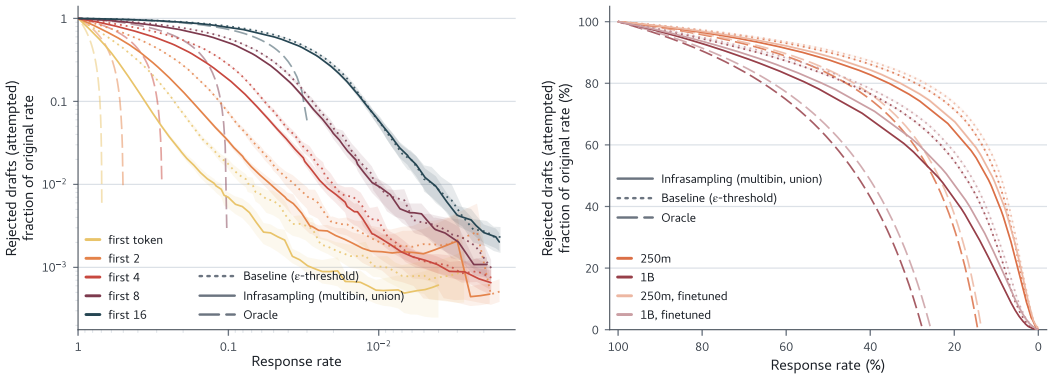
Downstream benchmarks. As discussed in Section 5.1, the least favorable setting for infrasampling should be when the test domain has little aleatoric uncertainty, e.g. benchmarks with essentially one correct answer. Figure 5 shows that on TriviaQA (Joshi et al., 2017b), WMT German to English (Bojar et al., 2014), and XSum (Narayan et al., 2018), infrasampling still largely matches the best baselines, such as ϵ-threshold—which itself handily beats more expensive methods like semantic entropy.

Mixture of downstream benchmarks. To test the claim in Section 5.1 that infrasampling's error tolerance knob should have a more consistent meaning across domains than heuristics, we follow Yadkori et al. (2024) and compare the tradeoff frontiers for several benchmarks individually vs. mixed together. We mix a broad range of 15 benchmarks (see Appendix E.2) which we can score uniformly with teacher draft acceptance. Figure 12 shows that infrasampling with 7B regresses substantially less than ϵ-threshold, by <1 percentage point vs. up to 15pp.

Model scaling & pretraining vs. finetuning. In Figure 6b, we seem to see the same model scaling trend as chess—the frontier of relative reduction in hallucination rate improves with scale.[10] We also see that pretraining from scratch performs substantially better than finetuning with a frozen trunk.

Additional sampling strategies. Although infrasampling is intended to intervene on only epistemic hallucinations, other strategies to reduce aleatoric hallucinations can still be composed with it. For example, temperature sampling often improves downstream evals at the expense of diversity; abstaining based on the average uncertainty along the sequence rather than the maximum has a similar effect, for both infrasampling and ϵ-threshold (Figure 13). See Appendix E.3 for more results.

[10]We don't plot 7B on here because it wasn't pretrained with distillation, which confounds the scaling ladder.



(a) 1B varying the number of tokens in the continuation. (b) 4-token continuations across model scales.

Figure 6: **LM prefix continuation sampling evals**; the student's output is used as a speculative decoding draft for the teacher. Rejection rate (among nonabstained outputs) as a fraction of the zero-abstention rate.

8 Related work

Our framework synthesizes concepts from many prior works on uncertainty quantification.

In uncertainty quantification, there are two relevant axes of the design space. **First,** what the uncertainty scores are used for, such as: human interpretability (Jumper et al., 2021; Marusich et al., 2024), conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021), or learning with abstentions/selective prediction (Chow, 1957; 1970; Geifman & El-Yaniv, 2017; 2019; Thulasidasan et al., 2019; Cohen et al., 2024). Infrasampling adopts the abstention paradigm and casts hallucinations (in the presence of aleatoric uncertainty) as a suitable classification problem. **Second,** how the uncertainty is quantified, such as: slower methods like (self-)ensembling (Blundell et al., 2015; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Kuhn et al., 2023; Zhang et al., 2026) and iterative natural language elicitation (Kadavath et al., 2022; Yadkori et al., 2024); or faster methods like evidential deep learning (Sensoy et al., 2018; Malinin & Gales, 2018), epinets (Osband et al., 2023a;b), and classifier probes (Alain & Bengio, 2017; Ahdritz et al., 2024; Kossen et al., 2025). Though several of these, like us, learn explicit higher-order distributions over labels, our heads directly observe soft targets via a teacher model in a frequentist setting, clearing a path towards accurate uncertainty decomposition (Johnson et al., 2024; Ahdritz et al., 2025).

Re hallucinations: beyond the works that explain *what* hallucinations are (see Section 2), others explain *why* they exist, such as Agrawal et al. (2024); Kalai & Vempala (2024); Xu et al. (2025); Orgad et al. (2025); Simhi et al. (2025), and the most relevant, Kalai et al. (2025). It proves that hallucinations ("invalid" outputs) are inevitable in language model pretraining because density estimation forces the model to make guesses. We just perform density *underestimation* instead.

9 Conclusion

In this paper we presented a new conceptualization of hallucinations—and its decomposition into epistemic and aleatoric hallucinations—which gives rise to the infrasampling framework. Small-scale experiments in the soft distillation setting with autoregressive language models show encouraging results, matching or exceeding strong heuristic baselines and benefiting increasingly from scale.

Many exciting questions remain for future work, such as: how infrasampling scales across more orders of magnitude of training compute, what the best uncertainty head architecture is, how well it generalizes out of domain & adversarially, and how training can be extended to the one-hot setting. Perhaps the biggest open question is how infrasampling should interact with reinforcement learning post-training. Reinforcement learning is how the hallucinatory defects of pretraining are currently addressed in practice, but it requires sampling rollouts in environments with verifiers and it harms calibration/output diversity (OpenAI et al., 2024; Kirk et al., 2024; Hamilton & Mimno, 2026). It remains to be seen to what extent the gains from each technique are competing vs. complementary.

We hope this framework will inspire new AI systems that are orders of magnitude more reliable.

REFERENCES

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references?, 2024. URL https://arxiv.org/abs/2305.18248.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Gustaf Ahdritz, Aravind Gollakota, Parikshit Gopalan, Charlotte Peale, and Udi Wieder. Provable uncertainty decomposition via higher-order calibration. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TId1SHe8JG.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anthropic. System card: Claude Fable 5 & Claude Mythos 5. Technical report, Anthropic, June 2026. URL https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf. Accessed: 2026-09-25.

Artificial Analysis. AA-Omniscience: Knowledge and Hallucination Benchmark — AA-Omniscience Hallucination Rate. https://artificialanalysis.ai/evaluations/omniscience?omniscience-index=omniscience-index-vs-attempt-rate#aa-omniscience-hallucination-rate, 2026. Accessed: 2026-09-25.

Sai Anirudh Athaluri, Sandeep Varma Manthena, V S R Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4):e37432, 2023. doi: 10.7759/cureus.37432.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29205–29216. Curran Associates, Inc., 2022. doi: 10.52202/068431-2118. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bc1d640f841f752c689aae20b31198c1-Paper-Conference.pdf.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blundell15.html.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58. Association for Computational Linguistics, 2014.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, llm judge, and ensemble scorers, 2025. URL https://arxiv.org/abs/2504.19254.

Christian Catalini, Xiang Hui, and Jane Wu. Some simple economics of agi, 2026. URL https://arxiv.org/abs/2602.20946.

Character.AI Team. Optimizing Large-Scale pretraining at Character.ai. https://blog.character.ai/squinch/, December 2025. Accessed: 2026-09-25.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [idk] token, 2024. URL https://arxiv.org/abs/2412.06676.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1184–1193. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/depeweg18a.html.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Aleksandra Eliseeva, Yaroslav Sokolov, Egor Bogomolov, Yaroslav Golubev, Danny Dig, and Timofey Bryksin. From commit message generation to history-aware commit message completion. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering*, ASE '23, pp. 723–735. IEEE Press, 2024. ISBN 9798350329964. doi: 10.1109/ASE56229.2023.00078. URL https://doi.org/10.1109/ASE56229.2023.00078.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346/.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. URL https://arxiv.org/abs/1506.02142.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL https://arxiv.org/abs/1705.08500.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409/.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30/.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234, 2023. doi: 10.1016/j.mcpdig.2023.05.004.

Sil Hamilton and David Mimno. Elias in the lighthouse, again? diagnosing low diversity in llm stories, 2026. URL https://arxiv.org/abs/2605.26492.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020. URL https://arxiv.org/abs/2010.14701.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abws/1503.02531.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2280784.

Hanzhang Hu, Debadeepta Dey, Martial Hebert, and J. Andrew Bagnell. Learning anytime predictions in neural networks via adaptive loss balancing. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33013812. URL https://doi.org/10.1609/aaai.v33i01.33013812.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457 – 506, 2019. URL https://api.semanticscholar.org/CorpusID:216465307.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Daniel D. Johnson, Daniel Tarlow, David Duvenaud, and Chris J. Maddison. Experts don't cheat: learning what you don't know by predicting pairs. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 577:583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pp. 160–171, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649777. URL https://doi.org/10.1145/3618260.3649777.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Andrej Karpathy and Dwarkesh Patel. Andrej karpathy — agi is still a decade away. Dwarkesh Podcast, oct 2025. URL https://www.dwarkesh.com/p/andrej-karpathy.

Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Position: Uncertainty quantification needs reassessment for large language model agents. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=Lrv20S5RZV.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement, 2019. URL https://arxiv.org/abs/1903.06059.

Philip Koopman. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 4:15–24, 04 2016. doi: 10.4271/2016-01-0128.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2025. URL https://openreview.net/forum?id=YQvvJjLWX0.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation, 2019a. URL https://openreview.net/forum?id=SkxJ-309FQ.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2023. URL https://arxiv.org/abs/2211.17192.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL https://aclanthology.org/2020.findings-emnlp.165/.

Emmy Liu, Varun Gangal, Chelsea Zou, Michael Yu, Xiaoqi Huang, Alex Chang, Zhuofu Tao, Karan Singh, Sachin Kumar, and Steven Y. Feng. A unified definition of hallucination: It's the world model, stupid!, 2026. URL https://arxiv.org/abs/2512.21577.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks, 2018. URL https://arxiv.org/abs/1802.10501.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021. URL https://arxiv.org/abs/2002.07650.

Laura Marusich, Jonathan Bakdash, Yan Zhou, and Murat Kantarcioglu. Using AI uncertainty quantification to improve human decision-making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34949–34960. PMLR, 2024.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In Sašo Džeroski, Pierre Guerts, and Juho Rousu (eds.), *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pp. 65–81, Ljubljana, Slovenia, 05–06 Sep 2009. PMLR. URL https://proceedings.mlr.press/v8/nadeem10a.html.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: Open-domain structured data record to text generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 432–447, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.37. URL https://aclanthology.org/2021.naacl-main.37/.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807. Association for Computational Linguistics, 2018.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5525. URL https://aclanthology.org/W17-5525/.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OpenAI. GPT-6 astra system card. Technical report, OpenAI, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra. Accessed: 2026-09-25.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks,

Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

OpenRouter, Inc. LLM rankings. https://openrouter.ai/rankings, 2026. Live-updating leaderboard; usage data through Sep 24, 2026. Accessed: 2026-09-25.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations, 2025. URL https://arxiv.org/abs/2410.02707.

Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-tuning language models via epistemic neural networks, 2023a. URL https://arxiv.org/abs/2211.01568.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks, 2023b. URL https://arxiv.org/abs/2107.08924.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings*

of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A. Lewis, Joel Veness, and Tim Genewein. Amortized planning with large-scale transformers: A case study on chess, 2024. URL https://arxiv.org/abs/2402.04494.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018. URL https://arxiv.org/abs/1806.01768.

Maohao Shen, J. Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory W. Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Adi Simhi, Jonathan Herzig, Itay Itzhak, Dana Arad, Zorik Gekhman, Roi Reichart, Fazl Barez, Gabriel Stanovsky, Idan Szpektor, and Yonatan Belinkov. Hack: Hallucinations along certainty and knowledge axes, 2025. URL https://arxiv.org/abs/2510.24222.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6234–6243. PMLR, 2019.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11385–11393, 2022. doi: 10.1609/aaai.v36i10.21390.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in nlp, 2024. URL https://arxiv.org/abs/2404.07461.

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

William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045, 2023. doi: 10.1038/s41598-023-41032-5.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2025. URL https://arxiv.org/abs/2401.11817.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm, 2024. URL https://arxiv.org/abs/2406.02543.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, and Hao Wang. Tokur: Token-level uncertainty estimation for large language model reasoning, 2026. URL https://arxiv.org/abs/2505.11737.



Figure 7: **Depiction of epistemic vs. aleatoric uncertainty** (real probabilities from our models). Epistemic uncertainty is the model's failure to perfectly capture the training distribution, like poor factual recall. Aleatoric uncertainty is variation inherent to the training distribution, like multiple plausible continuations.

A Additional resources

In Figure 7, we provide an illustration (with real probabilities from our 1B LLM) of aleatoric and epistemic uncertainty in language.

B More details on chess setup

We construct teacher logits using Ruoss et al. (2024)'s 270M-parameter neural chess engine. It achieves an Elo of approximately 2900 against humans on Lichess. Unlike our student models, it does not generate moves directly: it predicts scores for each candidate state, so invalid moves always get score 0 by construction of the harness.

The training examples are transcripts of entire chess games formatted in standard algebraic notation (SAN) with move numbers and no headers. They are generated by performing rollouts from the teacher model. The first few moves of each game are seeded from a diverse set of named openings. From then on, remaining move are sampled from distributions derived from the teacher's predicted per-move win probabilities. To avoid unrealistic blunders and encourage decisive play, candidate moves predicted to be significantly worse than the best move at each position are excluded. We continue until the game ends or 100 moves are reached, whichever comes first. Test examples are generated independently in the same fashion. Each test example is one "ply," i.e. a turn taken by one player, conditioned on the transcript of the game so far.

Our student models are standard decoder-only transformers with an Olmo-3-like architecture (Olmo et al., 2026) and a vocabulary of 37 SAN symbols. We trained models from scratch at 1M, 10M, 25M, and 50M parameters with vs. without a binary absolute uncertainty head on 200 tokens per non-embedding parameter, with no repetitions.

For chess, the uncertainty heads use a threshold of 0 (a token is deemed to be in the support of the teacher if it has non-zero probability under the teacher). We weight the two main loss terms ($\mathcal{L}_{LM}$ and $\mathcal{L}_{unc}$) such that the two are approximately equal at initialization.

C Validating lack of regression on next-token prediction

Figures 8 and 9 demonstrate that perplexity does not regress in our experiments when we add an uncertainty head, and Figure 10 shows that downstream sampling evals do not regress either; if anything, they improve slightly.

D More details on language modeling setup

For our binary absolute head, we use $\tau = 1\%$; multibin ($K = 9$) absolute and relative heads use geometrically spaced thresholds between 0.1% and 20% and ratios between $< 1\times$ and $10\times$, respectively.

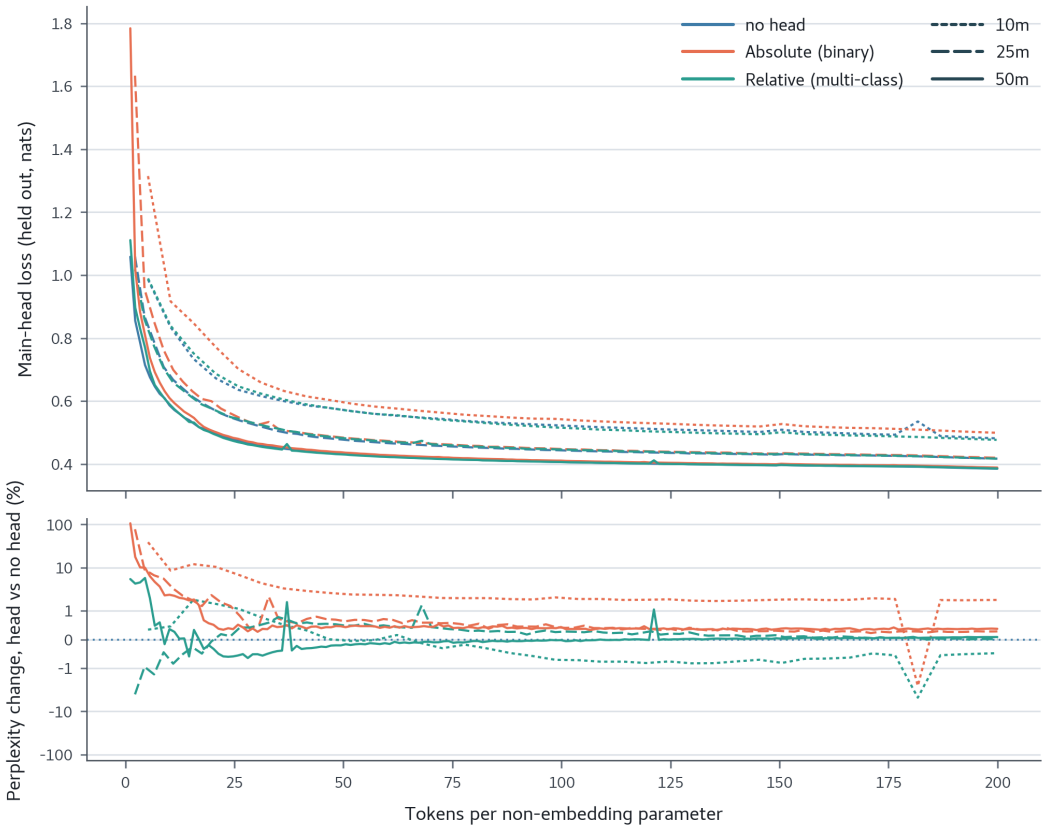


Figure 8: **Effect of uncertainty head next-token loss.** *Top:* loss in absolute. *Bottom:* loss relative to baseline (model trained without an uncertainty head).

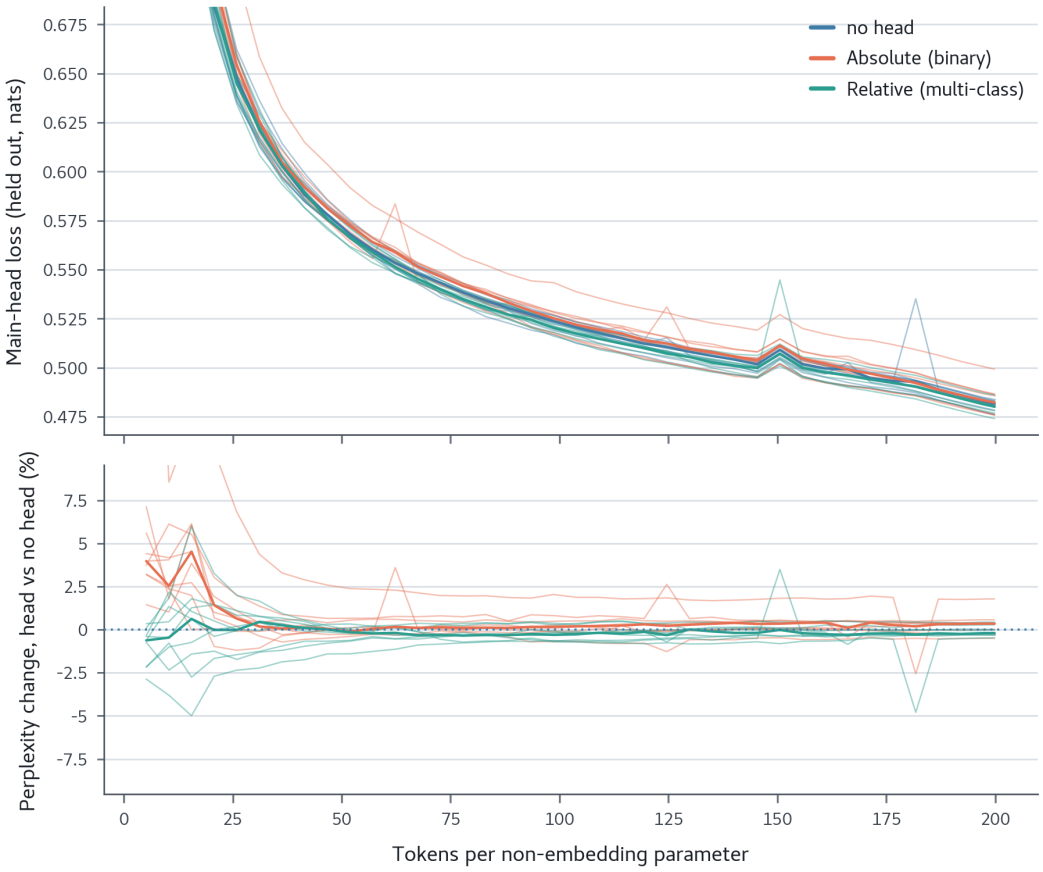


Figure 9: Retry of 10m from Figure 8 with multiple seeds to confirm that the apparent regression was just bad luck.

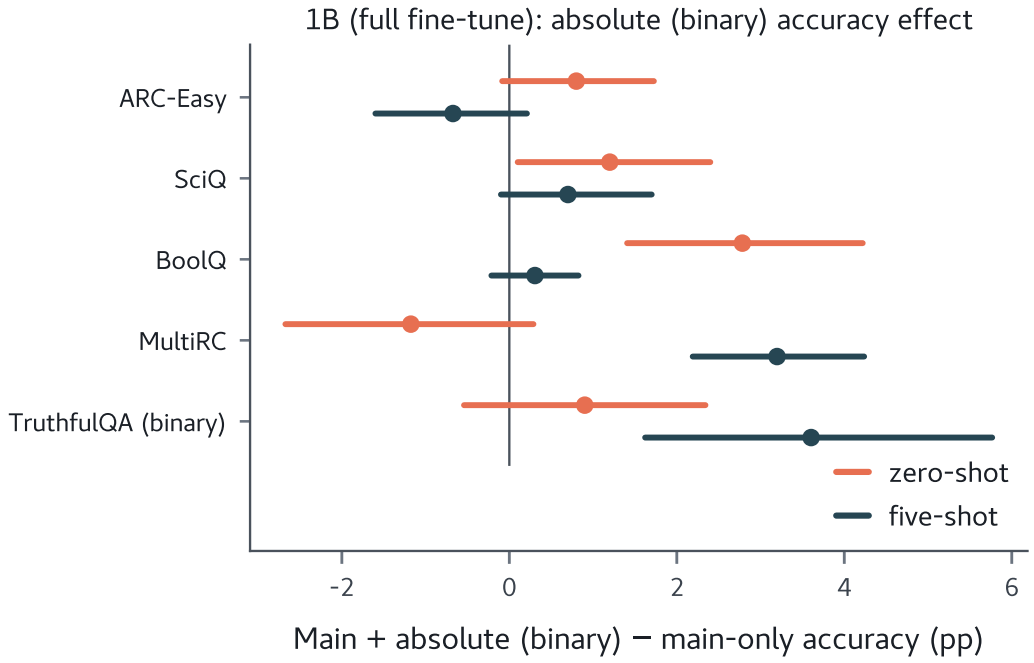


Figure 10: **Accuracy changes** on QA benchmarks from using a binary head compared to just a standard language head.

To circumvent the prohibitive storage costs of writing full teacher distributions to disk, we constructed an optimized online distillation pipeline: FP8-quantized teachers run alongside student models on vLLM servers (Kwon et al., 2023) and transmit logits for 1024 candidate tokens, drawn with the Gumbel top-k trick(Jang et al., 2017; Kool et al., 2019; Character.AI Team, 2025), as well as a small uniform sample from the tail. These candidates yield a self-normalized Horvitz–Thompson estimate (Horvitz & Thompson, 1952) of the soft cross-entropy for the main head. The binary absolute head additionally accepts the exact set of tokens that exceed its teacher-support threshold, in this case $\tau = 1\%$ (unablated). Analogously, the relative head uses ratios based on exact teacher logits for all known tokens and demotes remaining tokens to the drop bin.

We were also forced to address the extreme class imbalance induced by the size of the vocabulary. Averaged uniformly over the vocabulary, the head's loss is dominated by easy "drop" tokens the student already gives negligible mass. The few candidates that actually decide abstention get almost no gradient. For that reason, towards the end of training, we apply *student weighting*, i.e. we set $\mathcal{L}_{\text{unc}} = \sum_v \left(\beta/V + (1-\beta)\,\text{stopgrad}[p_\theta(v)]\right) \ell_v$ with $\beta = 0.9$. We find that this change improves the behavior of the model on the most important tokens during infrasampling. In Figure 11, we show weighted calibration of our uncertainty heads across model sizes and head types at their respective classification tasks, both before and after the student weighting finetuning step. Our heads—especially multibin relative—emerge from pretraining somewhat overconfident on commonly observed tokens. Student weighting largely corrects them. We discovered this student weighting scheme relatively late in the project and were only able to apply it in the final stage of training. In future runs, we plan to enable it from the start.

To prevent the uncertainty loss from effectively vanishing once easy "drop" tokens had been learned, we also enforce adaptive loss weighting, dynamically inflating the uncertainty term to at least 5% of the main loss (Chen et al., 2018; Hu et al., 2019).

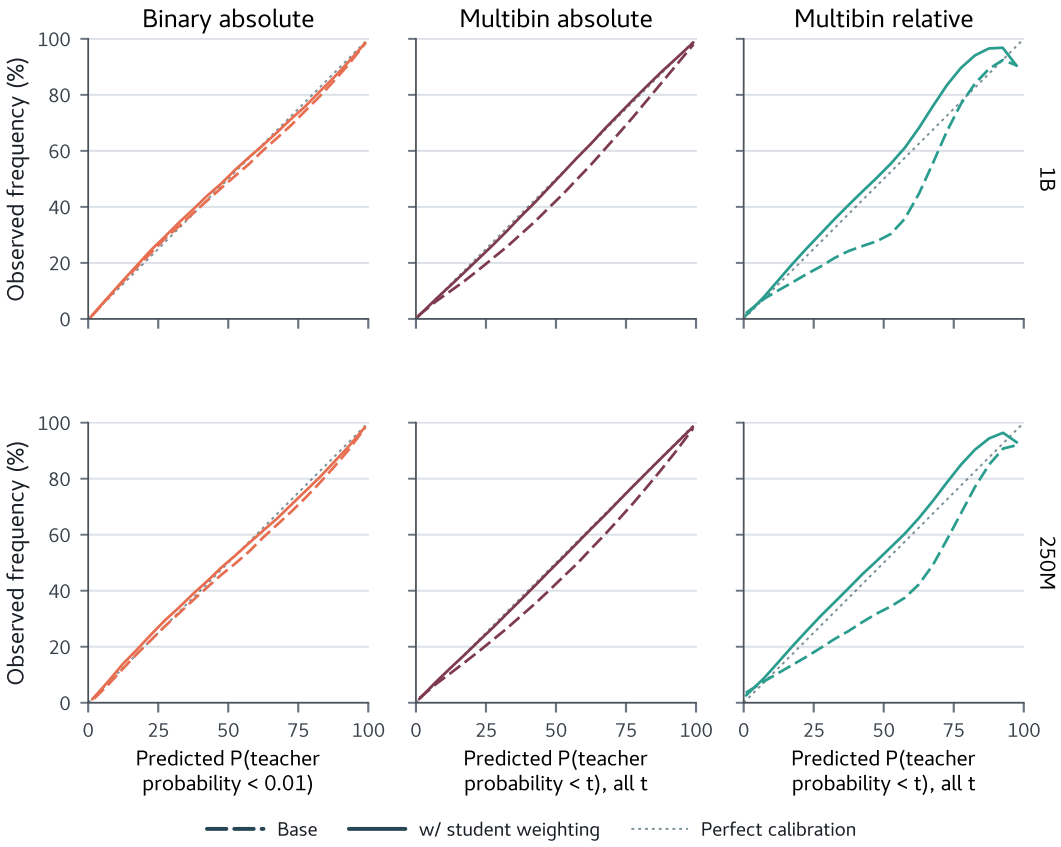


Figure 11: **Student weighting corrects weighted head calibration**. Reliability diagrams for three uncertainty head types across two model sizes, both before and after the "student weighting" finetuning step. We use 128 positions per sequence from 4096 held-out Dolma sequences, drawn uniformly, and 32 tokens per position drawn from the student's predicted distribution, so each token is weighted by its student probability.

E More language modeling results

E.1 Qualitative examples

To show our binary head in action, we include a selection of outputs of our 1b binary absolute model on a set of 24 handcrafted "epistemic" prompts (questions where there is only one correct answer) and 16 "aleatoric" ones (questuons where there are many valid answers). On "epistemic" questions with high predictive entropy, we expect the binary head to understand that only one of the many possible outputs is correct and output a high "drop" probability across the board. On "aleatoric" questions, correct behavior is to keep valid continuations. Examples are shown in Table 1

Indeed, the head demonstrates sensible behavior on the two classes of questions. It abstains on uncertain epistemic ones and keeps much more mass in the aleatoric case. Note that, in several cases, the right answer to an epistemic question happens to be the model's top choice pre-abstention (e.g. `Ou...` for "Ouagadougou"); this does not change the fact that, at the chosen tolerance of 0.2, answering poses too high a risk of hallucinating, and that the model is correct to abstain.

After writing the 40 questions in advance, we ran the model on them and filtered out 6 questions where it responded (e.g. with a space). We then selected 4 pairs of epistemic and aleatoric results with approximately similar entropies and matched them with a suitable tolerance. We judge that, of the 34 questions with valid answers, the head demonstrates "correct" behavior on 30 of them: 88%.

E.2 Mixture of benchmarks

Below, we list the NLP benchmarks used in our "pooled" evaluations:

Epistemic: one answer not known to the student			**Aleatoric:** many correct answers		
Candidate	p_θ	P(unsupported)	Candidate	p_θ	P(unsupported)
Capital of Burkina Faso? (H=3.10; head keeps 0.00)			*Name a sport.* (H=3.53; head keeps 0.42)		
`Ou...`	.18	.73	`Football`	.13	.09
`K...`	.16	.58	`Golf`	.10	.15
`G...`	.13	.83	`Tennis`	.09	.13
`B...`	.07	.87	`Soccer`	.08	.25
`F...`	.06	.88	`Rugby`	.07	.25
Capital of Tuvalu? (H=3.51; head keeps 0.00)			*Name a vegetable.* (H=3.39; head keeps 0.40)		
`Su...`	.20	.91	`Potato`	.20	.17
`Van...`	.08	.84	`Car...`	.20	.19
`N...`	.08	.89	`Bro...`	.06	.57
`Nu...`	.07	.76	`Tomato`	.06	.39
`V...`	.07	.94	`Onion`	.04	.66
Painter of Garden of Earthly Delights*?* (H=3.87; head keeps 0.00)			*Name a color.* (H=2.91; head keeps 0.35)		
`Car...`	.12	.71	`Red`	.35	.11
`Raphael`	.08	.58	`Blue`	.10	.27
`Michel...`	.08	.80	`Black`	.09	.41
`Rem...`	.07	.71	`White`	.07	.39
`Tit...`	.05	.92	`Green`	.06	.49
Second president of Kenya? (H=3.83; head keeps 0.00)			*Name a flower.* (H=2.85; head keeps 0.44)		
`J...`	.21	.35	`Rose`	.44	.08
`Julius`	.09	.80	`Lily`	.10	.28
`Daniel`	.08	.88	`Daisy`	.07	.36
`M...`	.07	.64	`Tul...`	.04	.34
`Uh...`	.06	.93	`Orch...`	.03	.54

Table 1: **Uncertainty heads separate very "epistemic" prompts from very "aleatoric" ones**. Illustrative predictions from our 1B binary absolute model on questions where the student turns out to be highly uncertain. P(unsupported), coloured from blue (0) to red (1), is the head's estimate that the teacher gives the token less than 1% probability. Headers list student's main-head entropy and the share of the student's probability mass the head retains at tolerance $\alpha = 0.2$. In the epistemic category the head correctly abstains completely on single-answer, purely "epistemic" questions when it does not know that answer; on the aleatoric questions it keeps the most likely answers, even though the student is just as uncertain.

- WMT De-En (Bojar et al., 2014)
- FLORES Fr-En (Goyal et al., 2022)
- TriviaQA (Joshi et al., 2017a)
- NQ-Open (Lee et al., 2019b)
- SQuAD (Rajpurkar et al., 2016)
- ASQA (Stelmakh et al., 2022)
- ELI5 (Fan et al., 2019)
- WritingPrompts (Fan et al., 2018)
- CNN/DN (Hermann et al., 2015)
- SAMSum (Gliwa et al., 2019)
- E2E (Novikova et al., 2017)
- DART (Nan et al., 2021)
- CommonGen (Lin et al., 2020)
- CommitChronicle (Eliseeva et al., 2024)

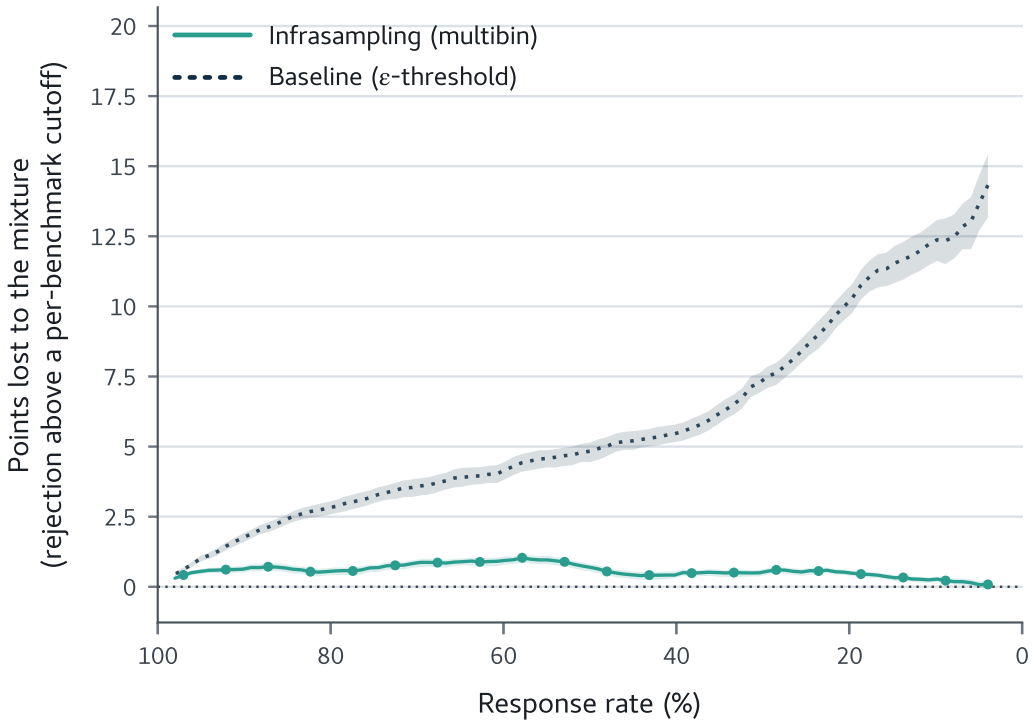


Figure 12: **Percentage points regressed when mixing downstream benchmarks together** rather than evaluating independently, 7B model. Infrasampling slightly regresses (<%1), whereas ϵ-threshold regresses an increasing amount with higher abstention rate, up to almost 15%.

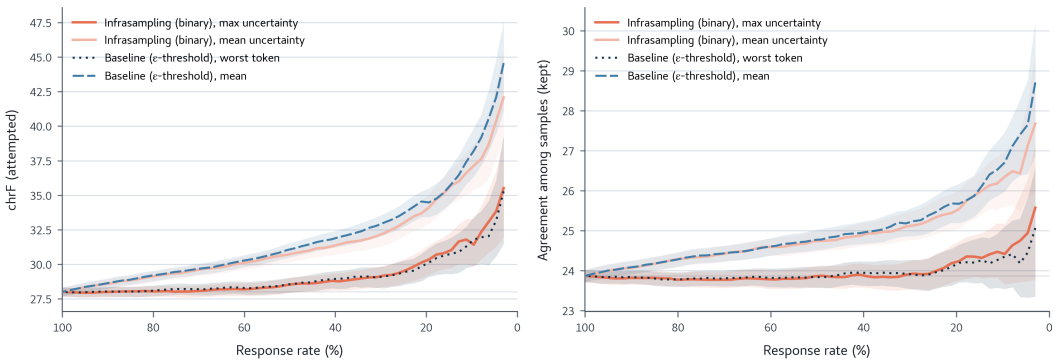
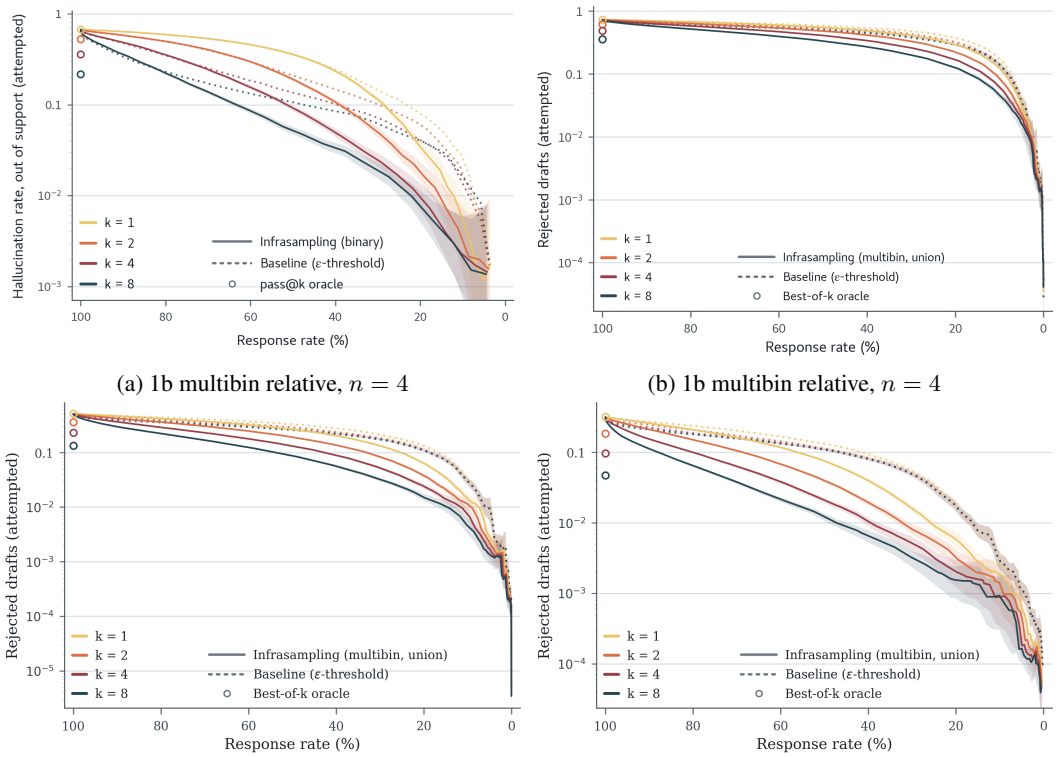


Figure 13: **Averaging uncertainty over a sequence instead of taking its maximum lowers output diversity**. WMT14 De-En. **(a)** The mean rule raises chrF for both the head and the ϵ-threshold baseline by about the same amount. **(b)** It also raises agreement among samples, i.e., it reduces output diversity, much as lowering the temperature does. Bands are 95% confidence intervals.

E.3 Additional sampling strategies

Infrasampling abstains whenever a sampled token lands in the slack of the lower bound. A natural way to trade inference compute for fewer abstentions is to try again: draw up to k candidate sequences and return the first one that is not abstained on, abstaining only if all k are. We show a handful of rejection sampling curves for our language models in Figure 14.



(a) 1b multibin relative, $n = 4$

(b) 1b multibin relative, $n = 4$

(c) 1b multibin relative, $n = 2$

(d) 1b multibin relative, $n = 1$

Figure 14: **Rejection sampling** from $k = 1$ to 16 attempts. N=4 tokens. Additional samples improve the tradeoff frontier more than for the ϵ-threshold baseline, which hardly moves at all.

Sampling from the lower bound $\check{p}_\theta^{(\alpha)}$ is implemented as rejection sampling from the student. A candidate $y = (y_1, \ldots, y_n)$ is drawn from p_θ and accepted with probability

$$a(y) \;=\; \prod_{t=1}^{n} \frac{\check{p}_\theta^{(\alpha)}(y_t \mid y_{<t})}{p_\theta(y_t \mid y_{<t})}, \tag{1}$$

using a single uniform draw per sequence rather than one per token. Tokens the head leaves untouched contribute a factor of 1, tokens it scales contribute their retained fraction, and a dropped token contributes 0, so one dropped token rejects the whole candidate. Under a binary head, acceptance is therefore all or nothing; multibin heads also produce intermediate acceptance probabilities. As $k \to \infty$, rejection sampling becomes sequence-level renormalization of the lower bound, which never abstains. At a fixed tolerance, a larger budget slightly raises the error rate, because the contexts it recovers are the hard ones a single attempt would have abstained on. The benefit appears at a matched response rate.

The oracle baseline for rejection sampling is pass@k with privileged grading: among the k candidates it returns a valid one whenever one exists. It is therefore an upper bound on any rule that sees only the student and its head.

Rather than simulate the sampling loop, we cache K candidates with their per-token acceptance factors, and evaluate every tolerance and budget analytically from them. The probability that candidate j is the first accepted is $a_j \prod_{m<j}(1 - a_m)$. For question answering, where the order of the cached candidates is arbitrary, we additionally average over their orderings.