DISTILLING ANSWERS, DISCARDING UNCERTAINTY: UNCERTAINTY COLLAPSE IN SEQUENCE-LEVEL KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-level knowledge distillation (SeqKD) commonly trains a student on a single sampled teacher response per input. While this transfers the teacher's answer, it need not transfer the uncertainty with which that answer was produced. We study the resulting uncertainty collapse and its consequences for hallucination detection. In expectation, the SeqKD objective is minimised by matching the teacher's response distribution and therefore preserves its uncertainty. But with one fixed sampled target per input, the empirical objective instead favours concentration on that target, regardless of whether it is correct. We show theoretically that a single response cannot identify the teacher's dispersion over alternative meanings, and that training provides a systematic pressure towards increasingly concentrated student output distributions. Across four teacher–student pairs on SimpleQA, continued distillation causes students to increasingly reproduce hallucinated teacher responses while becoming more certain of them: semantic-entropy AUROC for detecting incorrect answers falls from roughly 0.8 to chance, even when answer accuracy changes little. The effect also appears, more weakly, on held-out and out-of-distribution data, and recurs in mathematical reasoning. We evaluate filtering, abstention replacement, and multi-response distillation as mitigations. Among the interventions we study, retaining multiple teacher responses directly preserves distributional information and substantially maintains hallucination detectability, at a cost that grows with the number of responses. These results show that transferring a teacher's outputs and transferring its uncertainty are distinct objectives, and that reliable distillation requires preserving more than a single draw from the teacher's response distribution.

1 INTRODUCTION

Knowledge distillation (KD) provides a practical way to transfer the capabilities of large language models (LLMs) to smaller, more efficient students (Hinton et al., 2015; Xu et al., 2024). Sequence-level knowledge distillation (SeqKD) is supervised fine-tuning on teacher-generated sequences (Kim & Rush, 2016). This paradigm requires only black-box teacher access and remains widely used in modern LLM post-training pipelines such as in Qwen3 models (Yang et al., 2025) and DeepSeek models (Guo et al., 2025). In its standard form, each input is represented by a single sampled teacher response. The corpus therefore records what the teacher answered, but not how certain the teacher was: the distribution over alternative responses from which the target was drawn is discarded. Distillation transfers the answer and discards the uncertainty. We show that this has a direct consequence for the student: its uncertainty collapses under training, on correct and incorrect answers alike, and we call this uncertainty collapse.

This discarded information matters because a model's uncertainty is not incidental to its errors. When a model lacks reliable knowledge, repeated samples tend to disagree in meaning (Kadavath et al., 2022), and semantic entropy exploits this dispersion to detect confabulations from sampled responses (Kuhn et al., 2023; Farquhar et al., 2024). Such detectors rely on uncertainty being informative about correctness: hallucinated answers should generally carry greater uncertainty than correct ones. A teacher that hallucinates while uncertain is therefore still a teacher whose halluci-

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.

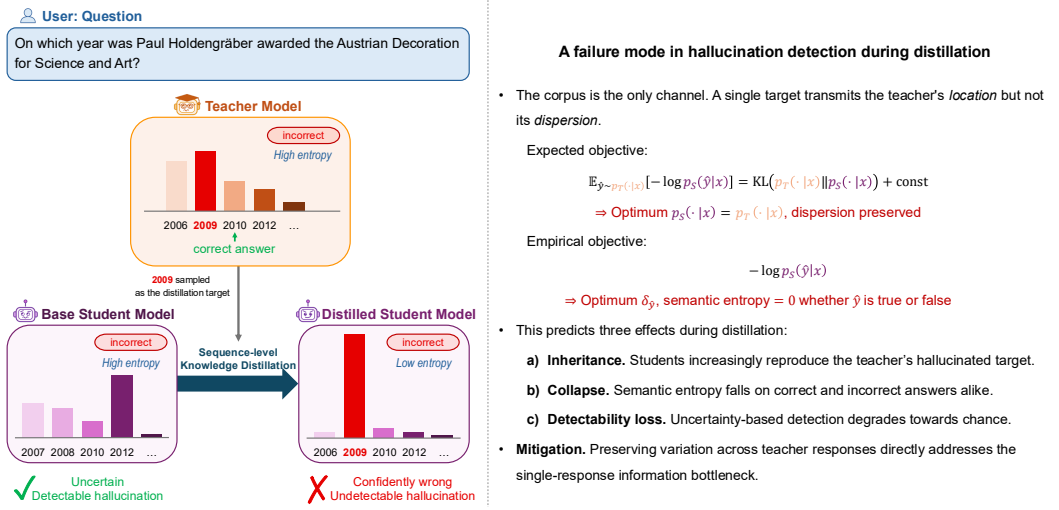


Figure 1: The failure mode studied in this paper and the analysis behind it. **Left.** On a SimpleQA question whose correct answer is 2010, the teacher is uncertain, and its sampled hallucination 2009 is recorded as the sole distillation target. The base student is also uncertain, so although it answers incorrectly, its high entropy allows the hallucination to be detected (green). Distillation on the single target concentrates the student on the inherited answer, and the same failure becomes confident and undetectable (red). **Right.** In expectation the SeqKD objective is minimised by matching the teacher's distribution, but a single recorded target is optimised by a point mass $\delta_{\hat{y}}$ on that target, whose semantic entropy is zero whether the target is true or false. The constant is the teacher's own entropy and does not depend on the student.

nations can be detected. Whether the same holds for a distilled student depends on what happens to uncertainty during distillation.

Recent work has shown that distillation can alter student uncertainty and calibration (Cui & Pei, 2026; Zhan et al., 2026; Zhang et al., 2026), and that students can inherit teacher errors and memorised content (Gudibande et al., 2024; Dankers & Raunak, 2025). Separately, uncertainty-based hallucination detection uses model uncertainty to identify unreliable responses (Kuhn et al., 2023; Farquhar et al., 2024). What remains unclear is **how semantic uncertainty changes with continued standard single-response SeqKD, and whether uncertainty-based hallucination detection remains effective as distillation proceeds**. We characterise these changes during distillation to study this training-time uncertainty collapse and its consequences for hallucination detection.

In this work, we study this question in three stages, with the failure mode illustrated in Figure 1. First, we show theoretically that although the expected SeqKD objective is minimised by matching the teacher's response distribution, the empirical objective with one fixed target per input is minimised by a point mass on that target. A single response cannot identify the teacher's dispersion over alternative meanings, and gradient training supplies a systematic pressure towards concentration. This predicts uncertainty collapse on correct and incorrect answers alike, degradation of uncertainty-based detection, and preservation through multiple teacher responses.

We then test these predictions empirically. A controlled multiple-choice experiment first makes the mechanism directly observable by tracking the distribution over candidate answers throughout training (Figure 2). Our main experiments study factual question answering on SimpleQA (Wei et al., 2024) across four teacher–student model pairs, measuring semantic uncertainty and detectability for students trained independently for increasing numbers of epochs. Across all four pairs, student uncertainty falls well below the teacher's on both correct and incorrect answers, while the AUROC for detecting hallucinations falls from around 0.8 to chance despite little change in answer accuracy. The effect extends more weakly to held-out and out-of-distribution data and recurs in mathematical reasoning. Students also increasingly reproduce the teacher's hallucinated targets, turning errors produced with uncertainty into errors reproduced with confidence.

Finally, we test three corpus-level interventions guided by this analysis. Filtering uncertain teacher examples slows the collapse, while replacing them with abstentions improves selective answering. Retaining multiple teacher responses directly preserves distributional information and substantially maintains detectability, with stronger preservation as more responses are retained. These results distinguish transferring a teacher's outputs from transferring its uncertainty.

2 Related Work

Distillation and uncertainty. Sequence-level knowledge distillation (SeqKD) trains students on teacher-generated sequences rather than full teacher distributions (Kim & Rush, 2016). Recent work has examined how distillation affects uncertainty and calibration. Preference distillation uses teacher–student preference pairs to improve sequence-level calibration (Zhang et al., 2024), while smoothed distillation transfers soft teacher targets to reduce overconfidence and hallucination (Nguyen et al., 2025). Most closely, Cui & Pei (2026) study how uncertainty from teacher outputs, student initialisation, and student outputs propagates through distillation, distinguishing variability within a student from variability across independently distilled students and showing that single-response distillation suppresses the former. Zhan et al. (2026) identify overconfident hallucination among broader failures associated with teacher–student mismatch, while Zhang et al. (2026) study the decoupling of capability and calibration under on-policy distillation. Our focus is complementary: we characterise how semantic uncertainty changes with increasing amounts of standard off-policy SeqKD and connect these changes to the reliability of uncertainty-based hallucination detection. Our analysis focuses on distributional information absent from the single-response corpus itself. Appendix H gives a detailed comparison.

Uncertainty-based hallucination detection. Sampling-based detectors exploit variation across model responses as evidence of unreliability. SelfCheckGPT measures response consistency (Manakul et al., 2023), while semantic uncertainty groups sampled responses by meaning and measures dispersion across semantic classes (Kuhn et al., 2023; Farquhar et al., 2024). Related methods use self-evaluation or internal representations to predict correctness (Kadavath et al., 2022; Chen et al., 2024; Kossen et al., 2024). Phillips et al. (2026) further show that semantic distributions estimated from multiple samples can themselves be distilled into an uncertainty predictor. These methods ask how uncertainty can be estimated, exploited, or explicitly transferred. We instead study whether the uncertainty signal they rely on survives ordinary response distillation: a teacher can produce an incorrect answer while remaining uncertain enough for the error to be detectable, whereas single-response SeqKD can preserve that answer while discarding the dispersion that made it detectable.

3 Preliminaries

Models and decoding. An input x is drawn from a question set $\mathcal{X}$, and a response y is a sequence of tokens from a vocabulary $\mathcal{V}$. The teacher and the student define conditional distributions $p_T(y \mid x)$ and $p_S^\theta(y \mid x)$. We write $\tilde{p}$ for the sampling distribution at a fixed temperature and truncation, and state all quantities on $\tilde{p}$.

Sequence-level knowledge distillation. In SeqKD (Kim & Rush, 2016), the teacher is queried once for each input, and the distillation corpus $\mathcal{D} = \{(x_i, \hat{y}_i)\}_{i=1}^N$ records a single draw $\hat{y}_i \sim \tilde{p}_T(\cdot \mid x_i)$. The student then minimises

$$\hat{\mathcal{L}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p_S^\theta(\hat{y}_i \mid x_i), \tag{1}$$

which is supervised learning on teacher-generated sequences and requires only black-box teacher access. The targets are generated in advance and held fixed, so the setting is off-policy.

Semantic classes and semantic entropy. For each input, the output space is partitioned into semantic equivalence classes $\mathcal{C}(x) = \{c_1, \ldots, c_K\}$ (Kuhn et al., 2023). A distribution q over sequences induces a class distribution $q(c_j \mid x) = \sum_{y \in c_j} q(y \mid x)$, with semantic entropy

$$H[C \mid x; q] = -\sum_{j=1}^{K} q(c_j \mid x) \log q(c_j \mid x). \tag{2}$$

Semantic entropy is the Shannon entropy of the induced distribution over semantic equivalence classes rather than individual response sequences. It is high when probability mass is spread across several meanings and zero when all mass lies in one semantic class. With black-box access, discrete semantic entropy (DSE) (Farquhar et al., 2024) estimates these class probabilities from the frequencies of semantic clusters among M sampled responses. It requires no access to token probabilities, and Farquhar et al. (2024) report detection performance comparable to standard semantic entropy.

In factual question answering each question has a single correct meaning, so we call an answer a hallucination when it falls outside the correct semantic class.

Uncertainty-based hallucination detection. A detector assigns an uncertainty score to each answer and predicts that answers scoring above a threshold are incorrect. Following Farquhar et al. (2024), incorrect answers are treated as the positive class, and detection quality is summarised by the area under the receiver operating characteristic curve (AUROC), the probability that a randomly chosen incorrect answer scores higher than a randomly chosen correct one. A value of 0.5 means that the score carries no information about correctness. We report AUROC rather than the accuracy of a thresholded detector because it summarises the full ranking without fixing a threshold.

4 Why Single-Response Distillation collapses uncertainty

What the corpus can transmit. The corpus is the only channel between the teacher and the student. For an input x, the teacher induces a semantic class distribution $\tilde{p}_T(c \mid x)$. We distinguish two properties of this distribution. Its location is the class $c(\hat{y})$ containing the response the corpus records, while its dispersion describes how probability mass is spread across semantic classes. We summarise this using semantic entropy of Equation 2. Uncertainty-based hallucination detection depends on dispersion. It can separate hallucinated from correct answers only when the two groups differ in semantic uncertainty. The question is which aspect survives the channel.

In expectation, SeqKD matches the teacher's distribution. Write $q_i = \tilde{p}_T(\cdot \mid x_i)$ for the distribution the target on x_i was drawn from. Taking the expectation of Equation 1 over the teacher's sampling and expanding each cross entropy gives

$$\mathbb{E}\big[\hat{\mathcal{L}}(\theta)\big] = \frac{1}{N}\sum_{i=1}^{N}\Big(H(q_i) + D_{\mathrm{KL}}\big(q_i \,\|\, p_S^\theta(\cdot \mid x_i)\big)\Big). \tag{3}$$

The entropy terms do not depend on θ, so the expected objective is minimised by $p_S^\theta(\cdot \mid x_i) = q_i$ on every training question. Matching the sequence distribution also matches its induced semantic distribution and therefore preserves the teacher's dispersion.

A single target collapses the student onto one meaning. Each question contributes one term $-\log p_S^\theta(\hat{y}_i \mid x_i)$ to Equation 1, and that term involves only the distribution on that question, so the objective decomposes question by question.

Proposition 1 (Point-mass optimum). *Treating $p_S^\theta(\cdot \mid x_i)$ as a free distribution over $\mathcal{Y}$, the unique minimiser of Equation 1 is the point mass $\delta_{\hat{y}_i}$ on every question, and $H[C \mid x_i; \delta_{\hat{y}_i}] = 0$ whatever the value of $H[C \mid x_i; q_i]$.*

The maximum-likelihood fit to a single observation is a point mass on that observation with zero semantic entropy, regardless of the teacher's dispersion. A parameterised student need not reach this optimum, although modern models can fit small corpora almost exactly (Zhang et al., 2017). Thus the expected and empirical objectives differ sharply: the former matches q_i, while the latter favours concentration on the observed response. This pressure is generic to deterministic targets; in SeqKD, it can turn an uncertain teacher sample, including an error, into a confident student prediction.

A single target cannot identify the teacher's dispersion.

Proposition 2 (Non-identifiability). *Let $\hat{H}$ be any estimator of $H[C \mid x; q]$ computed from a single draw $\hat{y} \sim q$. For any answer y_0 there exist distributions q_A and q_B with $H[C \mid x; q_A] = 0$ and $H[C \mid x; q_B] = \log K$ that both produce $\hat{y} = y_0$ with positive probability, on which $\hat{H}$ therefore returns the same value.*

The same observed response is compatible with a point mass or with a distribution spread across many semantic classes, so one draw cannot identify the teacher's dispersion. Pooling across questions does not resolve this per-input ambiguity: the corpus likelihood is maximised by assigning probability one to each observed target.

Gradient training moves towards concentration. Proposition 1 characterises the optimum but not the training dynamics. In a parameterised model, the conditional distributions for different questions are coupled through shared parameters, so they cannot be optimised independently. Therefore, we

ask whether the objective still provides a direction that systematically sharpens the student's output distribution. To study this, we consider the simplified parameterisation

$$p_S^\theta(y \mid x) \propto \exp\big(\beta\, \ell_\phi(y, x)\big), \tag{4}$$

where $\ell_\phi(\cdot, x)$ assigns a logit to each candidate response to x and $\beta > 0$ controls the sharpness of the resulting distribution. Training can lower the loss in two ways under this form. It can change which response receives the largest logit, or it can raise β so that the largest one takes more of the probability.

Proposition 3 (Sharpening direction). *Fix ϕ so that $\hat{y}_i$ uniquely maximises $\ell_\phi(\cdot, x_i)$ on every training question. Then $\partial\hat{\mathcal{L}}/\partial\beta < 0$ for every finite β, so gradient flow on β increases it without bound and the student converges to a point mass on the recorded target $\hat{y}_i$ on every question.*

Equation 4 is a softmax with inverse temperature β, whose entropy decreases as β grows (Lemma 1, Appendix A). Proposition 3 therefore identifies a shared sharpening direction across questions towards the point-mass optimum. The shared β isolates a common confidence direction across questions. Proofs of the three propositions are given in Appendix A.

Uncertainty collapse undermines hallucination detection. For uncertainty-based detection to remain effective, semantic uncertainty must remain informative about correctness. Proposition 1 implies that, in the point-mass limit, semantic entropy is zero on every training question regardless of whether the recorded target is correct. It therefore contains no information for distinguishing incorrect from correct answers; under our evaluation metric, its AUROC is 0.5. Together with the sharpening direction of Proposition 3, this predicts a loss of hallucination detectability with continued distillation as the student moves towards this limit. The analysis determines the limiting behaviour and a direction of pressure, but not a monotone finite-training trajectory.

Sampling multiple teacher responses preserves distributional information. Proposition 2 identifies the single draw as the bottleneck. Retaining multiple responses per input provides direct evidence about the teacher's dispersion over semantic outcomes, provided sampling is sufficiently diverse; low temperature or heavy truncation can otherwise make repeated draws nearly identical.

Predictions. The analysis yields four predictions. **P1.** Students should increasingly reproduce the recorded teacher responses, including hallucinated ones. **P2.** Student uncertainty should compress on both correct and incorrect responses with continued distillation. **P3.** This compression should make hallucinated responses progressively harder to distinguish from correct ones using uncertainty. **P4.** Retaining multiple teacher responses per input should preserve more of the teacher's uncertainty and therefore better maintain hallucination detectability.

5 Experiments

5.1 Setup

Models and training. We study four teacher–student pairs drawn from three different model families. The pairs are Llama-3.1-70B-Instruct to Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-32B to Qwen3-4B-Instruct-2507 (Yang et al., 2025), Qwen2.5-32B-Instruct to Qwen2.5-3B-Instruct (Qwen et al., 2025), and OLMo-2-0325-32B-Instruct to OLMo-2-1124-7B-Instruct (OLMo et al., 2025). Teacher and student come from the same model family within each pair, reducing tokeniser and architectural differences. A cross-family replication is given in Appendix D.5. Students are trained with full-parameter supervised fine-tuning on the teacher-generated pairs, using cross entropy over the answer tokens with the prompt masked from the loss, which optimises exactly Equation 1. For each pair we train five students independently from the same base model, for 2, 5, 10, 15 and 20 epochs respectively, and report the base student alongside them as epoch 0. Throughout the paper, we refer to these independently trained students as checkpoints. Implementation details and optimisation settings are given in Appendix B.1.

Data and measurement. The distillation corpus contains 500 questions sampled without replacement from SimpleQA (Wei et al., 2024) at a fixed seed. The same questions are used for the primary evaluation, since our predictions concern what a student absorbs from its own training targets. All four pairs use a strict answer-only prompt; for Qwen3 and Llama-3.1, we additionally repeat the experiment with the few-shot prompt of Farquhar et al. (2024) as a robustness check. The prompts

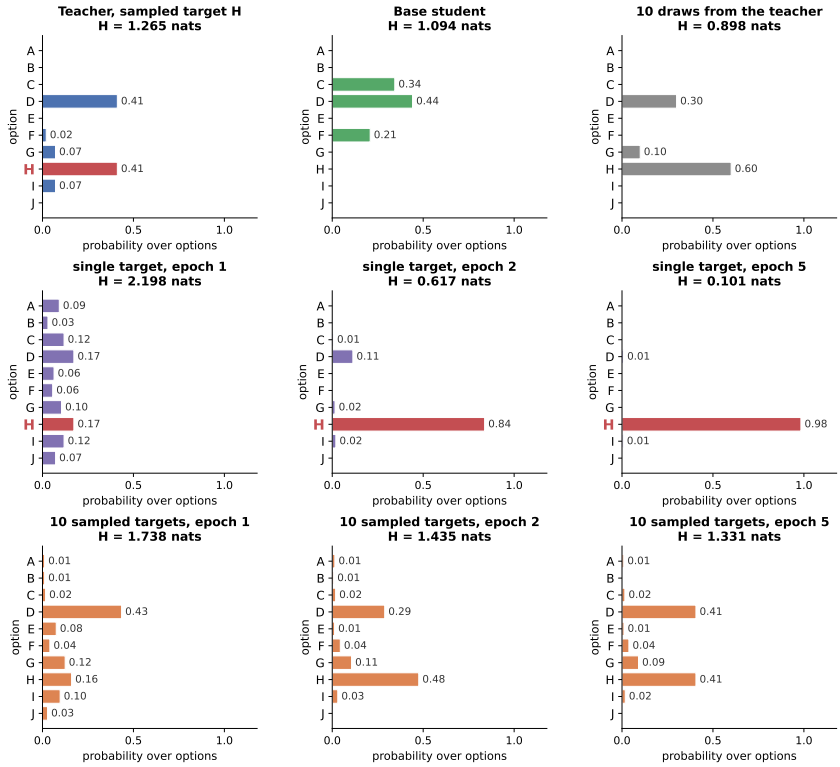


Figure 2: Distributions over the ten options for one MMLU-Pro question. The top row shows the teacher, the base student, and the empirical distribution of ten teacher draws. The middle and bottom rows show the student after 1, 2 and 5 epochs of distillation with one versus ten teacher responses per question. The sampled target is H and the correct option is F. Entropies are in nats.

are given in Appendix B.2.2, with full results in Appendix D.1. For each question, each model produces one near-greedy response at $\tau_0 = 0.1$ for accuracy, for a teacher model also as the distillation target, and ten samples at $\tau = 1.0$ for DSE. Answers are graded with the official SimpleQA template by Qwen3-32B, and the ten samples are clustered into semantic classes by bidirectional entailment with DeBERTa-v2-xlarge-mnli (He et al., 2021). Teacher abstentions are excluded from both the distillation corpus and detection metrics. Appendix D.3 shows that retaining them does not change the main conclusions. Appendix D.4 compares DSE with standard semantic entropy and naive entropy in this setting. Alongside accuracy on attempted questions and AUROC, we report the inherited error share, which is the proportion of a student's incorrect answers that are semantically equivalent to the teacher target on the same question.

5.2 A CONTROLLED EXPERIMENT IN MULTIPLE CHOICE

We first test the mechanism in a multiple-choice setting, where the answer space is simple and the distribution over options can be read from a single forward pass. 100 MMLU-Pro questions (Wang et al., 2024) are sampled at a fixed seed, and two corpora are built from the same questions with the Qwen3 pair. The single-response corpus records one teacher response per question while the multiple-response corpus records ten, both sampled at temperature $\tau = 1$. Students are trained independently for 1, 2 and 5 epochs, with gradient accumulation scaled to equalise parameter updates across the two corpora, and evaluated by renormalising the final-position logits over the option letters. The prompt is given in Appendix B.2.1.

Figure 2 follows one question through both runs. The teacher is uncertain between two incorrect options, while the single sampled target is H. Under single-response distillation, the student concentrates on H, reaching probability 0.98 and entropy 0.101 nats by epoch 5. With ten teacher responses, it remains dispersed and ends at 1.331 nats, close to the teacher distribution. Both runs show an early transient rise in entropy (Appendix C). Since the corpora differ only in the number of responses retained, their divergence isolates the role of distributional information: a single response drives the student towards the point-mass optimum of Proposition 1, whereas multiple responses preserve teacher dispersion.

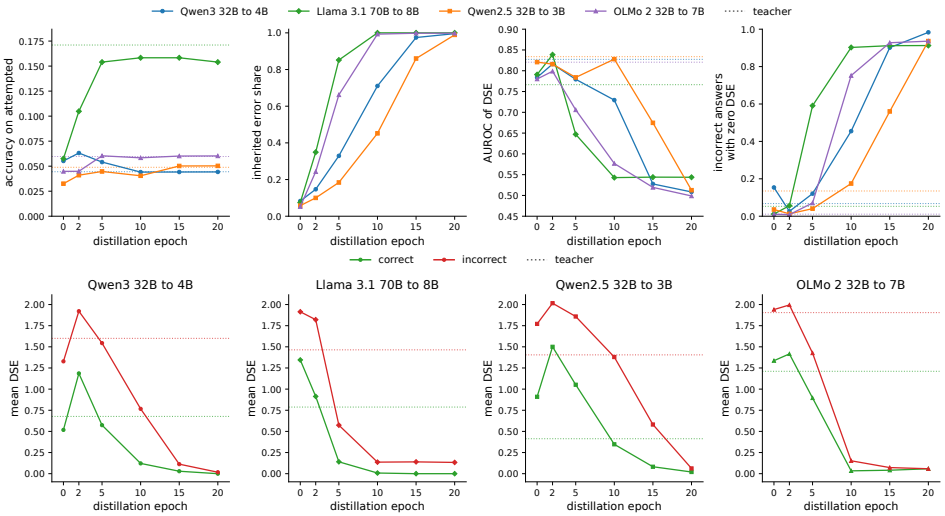


Figure 3: Distillation trajectories for the four teacher–student pairs under the strict prompt, with epoch 0 the base student. Top row, from left to right: accuracy on attempted questions; inherited error share; AUROC of DSE against incorrectness; and the proportion of incorrect answers with zero DSE. Bottom row: mean DSE on questions answered correctly and incorrectly, one panel per model pair. Dotted horizontal lines give the corresponding teacher values.

5.3 Empirical results

Across the four model pairs, accuracy on attempted questions changes modestly except for Llama-3.1, while uncertainty-based metrics show a much stronger and more consistent decline (Figure 3). Full results are given in Appendix D.1.

Uncertainty collapses on correct and incorrect answers alike. The bottom row of Figure 3 reports mean DSE separately on questions each checkpoint answers correctly and incorrectly. After the early transient, uncertainty decreases on both groups across all four model pairs, consistent with P2. The pattern is already visible by epoch 5 and becomes progressively stronger with continued distillation: Llama-3.1 and OLMo 2 show particularly sharp compression by this point, while Qwen3 and Qwen2.5 follow the same trend more gradually. At later checkpoints, mean DSE falls well below the corresponding teacher levels and approaches zero in several settings. Appendix D.2 examines the early transient in more detail.

The student collapses onto the recorded target. The inherited error share rises steadily across all four pairs (Figure 3, top row), but at different rates. For Llama-3.1 it already exceeds 0.8 by epoch 5 and is near one by epoch 10, coinciding with the sharp decline in semantic uncertainty. Together, these trajectories show that continued distillation does not merely make the student more certain: it increasingly concentrates the student on the particular responses recorded in the corpus, including hallucinated ones. An error produced uncertainly by the teacher can therefore become an error reproduced confidently by the student.

Detection degrades towards chance. DSE AUROC remains high initially, with temporary improvements in some settings, but subsequently declines across all four pairs (Figure 3, top row). The timing differs by model pair. For Llama-3.1, AUROC falls from 0.839 at epoch 2 to 0.647 at epoch 5 and 0.543 at epoch 10, remaining near this level thereafter. By epoch 10, roughly 90% of its incorrect answers have zero DSE, while accuracy on attempted questions remains near its highest level. Thus, the loss of detectability is already pronounced well before the longest training budget. More broadly, this failure is largely invisible to accuracy-based evaluation: detectability can approach chance while task accuracy remains stable or improves. Alternative entropy estimators reproduce the decline (Appendix D.4).

The effect recurs in mathematical reasoning. We repeat the pipeline on MATH (Hendrycks et al., 2021) with the Llama-3.1 pair, using full step-by-step teacher solutions as distillation targets. On the distillation problems, semantic uncertainty again collapses on both correct and incorrect responses after the early transient. DSE AUROC falls from 0.815 at epoch 5 to 0.644 by epoch 10 and remains substantially degraded thereafter, while accuracy on attempted problems remains around 0.4.

Thus, the divergence between task performance and uncertainty-based detectability also appears in mathematical reasoning. Experimental details and full results are given in Appendix E.

5.4 Generalisation beyond the distillation corpus

We next ask whether the effect extends beyond the questions directly used for distillation. Every checkpoint of the Qwen3 and Llama-3.1 pairs is evaluated, without further training, on held-out SimpleQA, TriviaQA (Joshi et al., 2017), and MMLU-Pro (Wang et al., 2024). Together, these sets test generalisation to unseen questions, a different factual QA benchmark, and a multiple-choice task. The main text reports the strict-prompt setting; dataset construction, grading, and few-shot results are given in Appendices F.1 and F.2.

Distillation has little effect on off-corpus accuracy. Despite substantial changes on the distillation corpus, accuracy on held-out SimpleQA, TriviaQA, and MMLU-Pro remains broadly stable across checkpoints for both model pairs (Appendix Figure 12, top row). Thus, continued fitting to the fixed SeqKD corpus has little systematic effect on task accuracy outside the questions used for distillation.

Uncertainty compression extends beyond the corpus. Although strongest on the distillation corpus, semantic-uncertainty compression also extends to unseen questions. Under both prompt styles, held-out SimpleQA and TriviaQA show lower mean DSE on both correct and incorrect responses after the early transient, while MMLU-Pro starts from lower DSE and shows smaller absolute changes (Appendix Figures 13–14). This compression is accompanied by a decline in hallucination detectability. On held-out SimpleQA, DSE AUROC falls from about 0.723 to 0.648 for Qwen3 and from 0.774 to 0.705 for Llama-3.1. TriviaQA shows a similar moderate decline, while MMLU-Pro starts from lower detectability and also declines with continued distillation (Appendix Figure 12).

This off-corpus compression is consistent with Proposition 3: because the sharpening direction acts through parameters shared across inputs, concentration induced on the distillation corpus can also affect unseen questions. The proposition does not determine the extent of this transfer. Held-out MATH, for example, shows no corresponding sustained decline in semantic uncertainty or DSE AUROC (Appendix E.2). Across the factual and multiple-choice evaluations here, continued SeqKD therefore alters student uncertainty beyond directly fitted questions even when off-corpus task accuracy remains largely unchanged. Overall, the collapse is strongest on the distillation corpus, while its transfer to unseen inputs is weaker and task dependent.

6 Interventions

To mitigate the collapse, we intervene on corpus construction in three ways. Filtering removes questions whose teacher DSE exceeds a threshold. Abstention replacement keeps those questions but substitutes their targets with `Unknown`. Multi-response distillation retains several teacher responses per question. The first two modify which deterministic targets the student is fitted to, whereas the third retains information about the teacher's response distribution. All interventions use the Llama-3.1 pair with the setup of Section 5.1 and are compared against the single-response baseline. The main text reports the strict prompt. Every conclusion below also holds under the few-shot prompt (Appendix G.2) and for the Qwen3 pair under both prompts (Appendix G.3). Filtering and replacement share three threshold levels denoted loose, medium and tight, with details in Appendix G.1.

Filtering preserves more of the student's uncertainty. Removing high-DSE teacher examples slows the loss of detectability, with stronger filtering yielding higher AUROC and fewer incorrect answers with zero DSE at later checkpoints (Figure 4, top row). It also reduces inherited error share. By withholding uncertain teacher targets rather than fitting them as fixed answers, filtering leaves the student less concentrated on the teacher's hallucinations and preserves more of its own uncertainty. The benefit comes at the cost of reduced training coverage.

Replacement improves selective answering. Replacing high-DSE targets with `Unknown` does not prevent DSE AUROC from approaching chance (Figure 4, middle row), because each selected question still has a single deterministic target. Instead, the student learns to abstain selectively: AURAC and accuracy on attempted questions increase, with stronger replacement producing more abstention. Replacement therefore improves the reliability of answers the student chooses to provide, but does not preserve the uncertainty signal used for hallucination detection.

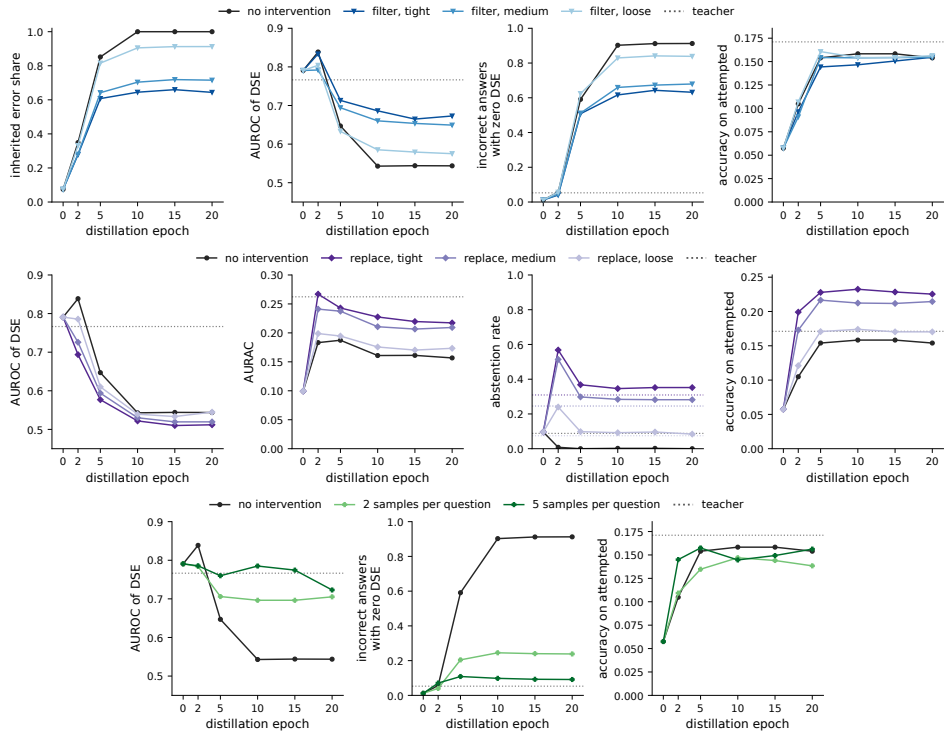


Figure 4: Corpus-level interventions for the Llama-3.1 pair under the strict prompt. **Top:** filtering removes high-DSE teacher examples at three threshold levels. **Middle:** abstention replacement substitutes high-DSE teacher targets with `Unknown` at the same thresholds. **Bottom:** multi-response distillation retains two or five teacher responses per question instead of one. Dotted horizontal lines indicate teacher values where applicable; horizontal reference lines in the replacement abstention panel indicate the fraction of targets replaced at each threshold.

Multiple teacher responses preserve uncertainty-based detectability. We compare two and five teacher responses per question, sampled at $\tau = 1.0$, with the single-response baseline. Multiple responses expose variation in the teacher's answers and therefore provide information about its uncertainty that a single target cannot contain. The effect is graded (Figure 4, bottom row): five responses keep DSE AUROC near its initial level and the fraction of incorrect answers with zero DSE low, while two responses provide intermediate preservation. Accuracy remains similar across settings. The cost increases with the number of teacher generations and training pairs.

Comparing the three interventions. Together, the interventions support the information-bottleneck account of Section 4. Filtering and replacement can slow or redirect the failure without adding information about the teacher's response distribution, whereas multi-response distillation (Cui & Pei, 2026) directly restores some of this information and provides the strongest preservation in our setting. Its graded improvement from one to two to five responses further links preservation to the amount of distributional information retained.

7 Conclusions

Standard single-response SeqKD can transfer a teacher's answers while discarding information about its uncertainty. With continued distillation, students increasingly concentrate on recorded teacher responses, including hallucinations, while semantic uncertainty collapses on both correct and incorrect answers. As a result, uncertainty-based hallucination detection can approach chance even when answer accuracy remains stable or improves. We trace this failure to the distributional information lost when each input is represented by a single teacher response, and the intervention results support this account. The effect is strongest on the distillation corpus and transfers more weakly and task-dependently to unseen inputs. Distilling answers and preserving the uncertainty that makes errors detectable are therefore distinct objectives.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, assist in the writing of proofs, and provide feedback on research methodology and experiments. Additionally, we used generative AI tools to create or modify scientific figures, create or edit software code, edit the paper to improve readability, identify relevant literature, and suggest the structure of the paper.

All AI-assisted research content was reviewed and validated by the authors. Mathematical claims and proofs were independently checked for correctness, suggested literature was verified against the original sources, and AI-assisted code was reviewed and tested before being used to produce reported results. AI-assisted figures were checked against the underlying experimental results, and AI-assisted edits to the manuscript were reviewed to ensure that they preserved the intended scientific meaning. The authors made the final decisions on the methodology, experiments, analysis, and presentation of the work. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or newly collected datasets. All experiments use publicly released models (Llama-3.1, Qwen2.5, Qwen3, and OLMo 2) and public benchmarks (SimpleQA, TriviaQA, MATH, and MMLU-Pro) under their respective licences. Our results identify a potential reliability risk in sequence-level knowledge distillation: distilled models can become more confident in inherited hallucinations while conventional accuracy-based evaluation does not reveal the degradation in uncertainty-based detectability. This failure mode may be relevant to deployed distilled models that rely on uncertainty-based safeguards. The findings therefore motivate evaluating uncertainty and hallucination detectability alongside task performance when developing distilled models. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The models, training hyperparameters, decoding settings, and semantic clustering procedure are specified in Appendix B.1, and the prompts used for answer generation, grading, and the multiple-choice experiment are reproduced in Appendix B.2. The main experiments use 500 SimpleQA questions sampled at a fixed seed, and the controlled multiple-choice experiment uses 100 MMLU-Pro questions, as described in Sections 5.1 and 5.2. Setup details for the MATH and generalisation experiments are given in Appendix E.1 and F.1 respectively. Thresholds and selected fractions for filtering and abstention replacement are reported in Appendix G.1. Complete numerical results for the main SimpleQA experiments are reported in Table 1, with additional results for MATH, generalisation, and interventions provided in the corresponding appendices. Intervention results cover two model pairs and both prompt styles (Section 6; Appendices G.2 and G.3). Code for corpus construction, distillation, evaluation, and figure generation, together with sampled question indices and generated model outputs, will be released upon publication.

REFERENCES

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0d1986a61e30e5fa408c81216a616e20-Paper-Conference.pdf.

Ziyao Cui and Jian Pei. How is uncertainty propagated in knowledge distillation?, 2026. URL https://arxiv.org/abs/2601.18909.

Verna Dankers and Vikas Raunak. Memorization inheritance in sequence-level knowledge distillation for neural machine translation. 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 2: Short Papers), pp. 760–774, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.61. URL https://aclanthology.org/2025.acl-short.61/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07421-0.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, and Zoe Papakipos. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17842–17859,

2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4db16435e3a5a2ef3fc39b8f0d12498d-Paper-Conference.pdf.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

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

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 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

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. URL https://openreview.net/forum?id=VD-AYtP0dve.

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

Hieu Nguyen, Zihao He, Shoumik Atul Gandre, Ujjwal Pasupulety, Sharanya Kumari Shivakumar, and Kristina Lerman. Smoothing out hallucinations: Mitigating LLM hallucination with smoothed knowledge distillation, 2025. URL https://arxiv.org/abs/2502.11306.

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, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2025. URL https://arxiv.org/abs/2501.00656.

Edward Phillips, Sean Wu, Fredrik K. Gustafsson, Boyan Gao, and David A. Clifton. Semantic self-distillation for language model uncertainty. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5427–5447. PMLR, 2026. URL https://proceedings.mlr.press/v337/phillips26b.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task

language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

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.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen 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, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Distillation traps and guards: A calibration knob for LLM distillability. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19814–19833, San Diego, California, United States, 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.908. URL https://aclanthology.org/2026.acl-long.908/.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Jiaxin Zhang, Xiangyu Peng, Qinglin Chen, Qinyuan Ye, Caiming Xiong, and Chien-Sheng Wu. The illusion of certainty: Decoupling capability and calibration in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.16830.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Haorui Wang, Zhen Qin, Feng Han, Jialu Liu, Simon Baumgartner, Michael Bendersky, and Chao Zhang. PLaD: Preference-based large language model distillation with pseudo-preference pairs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15623–15636, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.923. URL https://aclanthology.org/2024.findings-acl.923/.

A Proofs

Throughout this appendix the set of candidate responses to an input is finite, which holds whenever responses are bounded in length, and all logits are finite. Semantic classes partition this set, and $c(y)$ denotes the class containing y.

A.1 Entropy under a scale parameter

The monotonicity of entropy under the scale parameter used in Proposition 3 follows from the following lemma.

Lemma 1. *Let $z_1, \dots, z_n$ be fixed real numbers and let $p^\beta(j) = \exp(\beta z_j)/Z(\beta)$ with $Z(\beta) = \sum_j \exp(\beta z_j)$. Write $H(\beta)$ for the Shannon entropy of p^β. Then*

$$\frac{\mathrm{d}H}{\mathrm{d}\beta} = -\beta \operatorname{Var}_{p^\beta}(z), \tag{5}$$

so H is non-increasing in β for $\beta > 0$. Writing $\tau = 1/\beta$ for the corresponding temperature, $\mathrm{d}H/\mathrm{d}\tau = \operatorname{Var}_{p^\tau}(z)/\tau^3$, so H is non-decreasing in τ and strictly increasing unless all z_j are equal.

Proof. Since $\log p^\beta(j) = \beta z_j - \log Z(\beta)$, the entropy is

$$H(\beta) = -\sum_j p^\beta(j)\big(\beta z_j - \log Z(\beta)\big) = \log Z(\beta) - \beta\, \mathbb{E}_{p^\beta}[z]. \tag{6}$$

Differentiating $\log Z$ gives $\mathrm{d}\log Z/\mathrm{d}\beta = \mathbb{E}_{p^\beta}[z]$, and differentiating a second time gives $\mathrm{d}\mathbb{E}_{p^\beta}[z]/\mathrm{d}\beta = \operatorname{Var}_{p^\beta}(z)$, which are the standard cumulant identities for an exponential family with natural parameter β and sufficient statistic z. Differentiating the display above therefore gives

$$\frac{\mathrm{d}H}{\mathrm{d}\beta} = \mathbb{E}_{p^\beta}[z] - \mathbb{E}_{p^\beta}[z] - \beta \operatorname{Var}_{p^\beta}(z) = -\beta \operatorname{Var}_{p^\beta}(z). \tag{7}$$

The variance is non-negative and vanishes only when all z_j are equal. Substituting $\beta = 1/\tau$ and $\mathrm{d}\beta/\mathrm{d}\tau = -1/\tau^2$ gives $\mathrm{d}H/\mathrm{d}\tau = \operatorname{Var}_{p^\tau}(z)/\tau^3$. □

Applying Lemma 1 to the sequence-level logits $\ell_\phi(\cdot, x)$ of Equation 4 gives the entropy monotonicity used in Proposition 3.

A.2 Proof of Proposition 1

Proof. The objective in Equation 1 is a sum of terms $-\log p_S^\theta(\hat{y}_i \mid x_i)$, and the ith term depends only on the distribution assigned to x_i. When these distributions are unconstrained, they may be varied independently, so the sum is minimised by minimising each term separately. Each term satisfies $-\log p(\hat{y}_i) \geq 0$, with equality if and only if $p(\hat{y}_i) = 1$. Since p is a probability distribution, $p(\hat{y}_i) = 1$ forces $p = \delta_{\hat{y}_i}$, so the minimiser exists and is unique.

For the entropy claim, the class distribution induced by $\delta_{\hat{y}_i}$ assigns probability one to $c(\hat{y}_i)$ and zero to every other class, so its semantic entropy is zero. The argument uses only the identity of the recorded target and never refers to q_i, so the conclusion holds whatever the value of $H[C \mid x_i; q_i]$. □

A.3 Proof of Proposition 2

Proof. Fix an input x and an answer y_0, and let $K \geq 2$. Choose representatives $y_1, \dots, y_K$ of K distinct semantic classes with $y_1 = y_0$. Define $q_A = \delta_{y_0}$ and let q_B assign probability $1/K$ to each of $y_1, \dots, y_K$. The class distribution induced by q_A is a point mass, so $H[C \mid x; q_A] = 0$, and the class distribution induced by q_B is uniform on K classes, so $H[C \mid x; q_B] = \log K$.

Both distributions assign positive probability to the observation $\hat{y} = y_0$, namely 1 under q_A and $1/K$ under q_B. An estimator computed from a single draw is a function $\hat{H}(\hat{y})$ of that draw, so on the event $\hat{y} = y_0$ it returns the value $\hat{H}(y_0)$ under both distributions. If the estimator is randomised, the conditional distribution of its output given $\hat{y} = y_0$ is likewise the same under both. Its output on this event therefore carries no information about which of the two generated the draw.

The two entropies differ by $\log K$, so on this event

$$\max\Big\{\big|\hat{H}(y_0) - H[C \mid x; q_A]\big|,\ \big|\hat{H}(y_0) - H[C \mid x; q_B]\big|\Big\} \geq \tfrac{1}{2}\log K. \tag{8}$$

The construction can be repeated independently at each input, so no procedure recovers the teacher's dispersion on a given question from a corpus holding one draw per question. □

A.4 Proof of Proposition 3

Proof. Write $\ell_i(y) = \ell_\phi(y, x_i)$, which is held fixed, and let $Z_i(\beta) = \sum_y \exp(\beta \ell_i(y))$. Under Equation 4 the objective is

$$\hat{\mathcal{L}}(\beta) = -\frac{1}{N} \sum_{i=1}^{N} \Big(\beta \, \ell_i(\hat{y}_i) - \log Z_i(\beta) \Big). \tag{9}$$

Using $\mathrm{d} \log Z_i / \mathrm{d}\beta = \mathbb{E}_{p_i^\beta}[\ell_i]$ as in Lemma 1,

$$\frac{\partial \hat{\mathcal{L}}}{\partial \beta} = -\frac{1}{N} \sum_{i=1}^{N} \Big(\ell_i(\hat{y}_i) - \mathbb{E}_{p_i^\beta}[\ell_i] \Big). \tag{10}$$

The distribution p_i^β has full support for every finite β, and $\hat{y}_i$ is by assumption the unique maximiser of ℓ_i, so $\mathbb{E}_{p_i^\beta}[\ell_i] < \ell_i(\hat{y}_i)$ strictly. Every summand is therefore positive and $\partial \hat{\mathcal{L}} / \partial \beta < 0$ for every finite β.

Gradient flow on the shared scale evolves as $\mathrm{d}\beta/\mathrm{d}t = -\partial \hat{\mathcal{L}} / \partial \beta > 0$, so β is strictly increasing. Suppose it converged to a finite limit β^*. The right-hand side is continuous in β and strictly positive at β^*, so it is eventually bounded below by a positive constant along the trajectory. Then $\beta(t)$ would grow without bound, contradicting convergence to β^*. Hence $\beta \to \infty$.

As $\beta \to \infty$, the probability that p_i^β assigns to any $y \neq \hat{y}_i$ is $\exp(-\beta(\ell_i(\hat{y}_i) - \ell_i(y)))$ up to normalisation, and the exponent diverges because the gap is strictly positive. The distribution therefore converges to $\delta_{\hat{y}_i}$ on every question, and its semantic entropy converges to zero, since entropy is continuous on a finite set. □

B Experimental Details

B.1 Implementation Details

Models. Qwen3-4B-Instruct-2507 was itself distilled from larger Qwen3 models during its own training (Yang et al., 2025), so its initial uncertainty profile may already reflect an earlier round of the process under study. The other three pairs serve as controls for this. The OLMo 2 release documents its full training pipeline, which contains no such step (OLMo et al., 2025), the Qwen2.5 release reports none (Qwen et al., 2025), and the Llama 3 report describes the 8B and 70B models as independently pretrained, with the 405B model contributing only synthetic post-training data (Grattafiori et al., 2024). Correctness is judged by Qwen3-32B (Appendix B.2.3), which is likewise run with thinking mode disabled, at $\tau = 0.7$, the officially recommended temperature for non-thinking mode, and with the Qwen3 truncation settings given below.

Training. Students are trained with the Hugging Face Trainer using full-parameter supervised fine-tuning, with a learning rate of 1×10^{-5}, a per-device batch size of 2 and gradient accumulation over 8 steps for an effective batch size of 16, a warmup ratio of 0.05, gradient norm clipping at 1.0, and a maximum sequence length of 1,024 tokens. The loss is hard-label cross entropy over the answer tokens with the prompt tokens masked. Five students are trained independently for 2, 5, 10, 15 and 20 epochs, each starting from the base student. Training runs on A100 80GB GPUs.

Generation. All models are run with the Hugging Face `transformers` library and produce direct answers without separate thinking traces. Qwen3-32B is a hybrid reasoning model and is run with thinking mode disabled (`enable_thinking=False`), while the remaining models support only direct answering and require no such setting. Decoding follows each family's published recommendation where one exists, using top-p set to 0.8 and top-k set to 20 for the Qwen3 and Qwen2.5 families, and top-p set to 0.9 for the Llama-3.1 family. OLMo 2 publishes no recommendation, so its truncation parameters are left at the library defaults. No repetition penalty is applied, and the temperature is always set explicitly. Each model is decoded with the settings of its own family, including the teacher and student of the cross-family pair in Appendix D.5, and with the same settings at every checkpoint. The near-greedy answer that is graded and, for the teacher, recorded as the

distillation target is generated at $\tau_0 = 0.1$. The ten responses used to estimate DSE are generated at $\tau = 1.0$.

Semantic clustering. The ten sampled responses for a question are clustered into semantic equivalence classes by bidirectional entailment using DeBERTa-v2-xlarge-mnli (He et al., 2021), following Kuhn et al. (2023). Two responses are placed in the same class when each entails the other in the context of the question. DSE is then the entropy of the resulting cluster frequencies. Clustering is applied to the ten sampled responses only, and correctness grading is applied to the near-greedy answer only, so the two measurements are computed from disjoint generations.

Multiple-choice setting. The corpora for the multiple-choice experiment are distilled with the same procedure, with a batch size of 4 and a warmup ratio of 0. As in the main experiments, a separate student is trained independently from the base student for each budget of 1, 2 and 5 epochs, on each of the two corpora. The ten-response corpus contains ten times as many training pairs, so gradient accumulation over 10 steps is used there, which gives the two students at each epoch budget the same number of parameter updates. Each student is evaluated after training by running a single forward pass on each question and renormalising the final-position logits over the option letters.

B.2 Prompts

B.2.1 Multiple-choice prompt

```
Question: {question}
{options}
Answer with only the option letter ({labels}), no explanation.
Answer:
```

The placeholder {`options`} is the list of options, one per line and each prefixed by its letter, and {`labels`} is the comma-separated list of letters available for that question.

B.2.2 SimpleQA answer prompts

The strict prompt instructs the model to return the minimal answer string.

```
Question: {question}
Answer the question with only the minimal factual answer string.
Do not write a full sentence.
Do not include explanations, context, hedging, or punctuation.
Do not start with phrases like 'The answer is' or 'It is'.
Use the most common valid form of the answer.
Answer:
```

The few-shot prompt follows Farquhar et al. (2024) and prepends five exemplars to the target question.

```
Answer the following question as briefly as possible.

Question: {exemplar question 1}
Answer: {exemplar answer 1}

...

Question: {exemplar question 5}
Answer: {exemplar answer 5}

Question: {question}
Answer:
```

The five exemplars are the final five rows of the SimpleQA test split and are excluded from the evaluation set, so no exemplar appears among the evaluated questions. The split is loaded in a fixed order, so the same exemplars are used for every model and run.

B.2.3 GRADING PROMPT

Correctness is judged with the official SimpleQA grading template, used verbatim from the reference implementation released with the benchmark (Wei et al., 2024). The template presents the question, the gold target and the predicted answer, and asks the judge to return A for a correct answer, B for an incorrect one and C for a non-attempt. Answers graded C are treated as abstentions.

B.2.4 MATH PROMPT

The MATH experiments use a single zero-shot prompt. The same prompt is used for teacher target generation and for evaluating every student and the teacher on both splits.

```
Solve the following competition maths problem.
Reason step by step, then put the final answer
inside \boxed{}.

{question}
```

C ADDITIONAL RESULTS FOR THE CONTROLLED MULTIPLE-CHOICE EXPERIMENT

Figure 5 plots the entropy of the student's distribution over the answer options at each epoch of the multiple-choice experiment of Section 5.2, for the single-response and ten-target runs, together with the teacher's entropy and that of the ten-draw empirical distribution. The values match the panel headings of Figure 2. Both runs rise at the first epoch and then diverge, with the single-response student falling towards zero and the multi-target student settling near the entropy of the teacher distribution, slightly above that of the empirical distribution it was trained on.

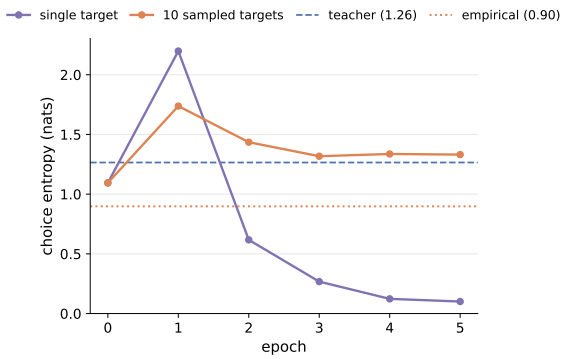


Figure 5: Entropy of the student's distribution over the answer options under single target and multi-target distillation, with the teacher's entropy and that of the ten-draw empirical distribution marked. Epoch 0 is the base student.

D MAIN SIMPLEQA EXPERIMENTS

D.1 FULL NUMERICAL RESULTS

Table 1 reports accuracy on attempted questions, the abstention rate, the AUROC of DSE against incorrectness, and AURAC, for every checkpoint of every setting. AURAC is the area under the accuracy and rejection curve, which measures accuracy retained when the least certain answers are rejected.

Table 1: Overall SimpleQA results across the four model pairs. Best results for accuracy, AUROC, and AURAC within each model–prompt block are shown in bold; ties at three decimal places are resolved using the unrounded values.

Model family	Prompt	Checkpoint	Accuracy	Abstention	AUROC	AURAC
Qwen3	strict	Teacher	0.044	0.010	**0.827**	0.082
		Base Student	0.055	0.022	0.784	0.092
		Epoch 2	**0.063**	0.018	0.816	**0.111**
		Epoch 5	0.054	0.002	0.779	0.096
		Epoch 10	0.044	0.002	0.729	0.065
		Epoch 15	0.044	0.004	0.528	0.044
		Epoch 20	0.044	0.006	0.508	0.043
	fewshot	Teacher	0.057	0.012	0.813	0.101
		Base Student	0.045	0.102	0.729	0.068
		Epoch 2	0.065	0.010	0.747	0.100
		Epoch 5	**0.074**	0.000	**0.820**	**0.131**
		Epoch 10	0.060	0.000	0.746	0.091
		Epoch 15	0.060	0.000	0.581	0.066
		Epoch 20	0.060	0.002	0.490	0.056
Llama-3.1	strict	Teacher	**0.171**	0.088	0.766	**0.262**
		Base Student	0.058	0.096	0.791	0.099
		Epoch 2	0.105	0.008	**0.839**	0.183
		Epoch 5	0.154	0.000	0.647	0.187
		Epoch 10	0.158	0.002	0.543	0.161
		Epoch 15	0.158	0.002	0.544	0.161
		Epoch 20	0.154	0.000	0.544	0.157
	fewshot	Teacher	**0.178**	0.122	0.733	**0.268**
		Base Student	0.094	0.808	0.786	0.158
		Epoch 2	0.109	0.010	**0.819**	0.188
		Epoch 5	0.150	0.000	0.642	0.180
		Epoch 10	0.158	0.000	0.559	0.165
		Epoch 15	0.158	0.000	0.558	0.165
		Epoch 20	0.158	0.000	0.570	0.168
Qwen2.5	strict	Teacher	0.049	0.018	**0.834**	**0.089**
		Base Student	0.033	0.016	0.821	0.057
		Epoch 2	0.041	0.024	0.816	0.066
		Epoch 5	0.045	0.016	0.784	0.077
		Epoch 10	0.040	0.010	0.828	0.072
		Epoch 15	0.050	0.004	0.675	0.066
		Epoch 20	**0.050**	0.006	0.513	0.049
OLMo 2	strict	Teacher	0.060	0.028	**0.821**	**0.102**
		Base Student	0.045	0.018	0.780	0.071
		Epoch 2	0.045	0.020	0.799	0.075
		Epoch 5	**0.060**	0.006	0.706	0.085
		Epoch 10	0.058	0.006	0.577	0.063
		Epoch 15	0.060	0.002	0.519	0.059
		Epoch 20	0.060	0.004	0.499	0.057

Accuracy on attempted questions changes little in most settings and rises towards the teacher's level for the Llama-3.1 pair, and abstention rates fall as training proceeds, most sharply for the base Llama-3.1 student under the few-shot prompt. AUROC and AURAC both fall substantially with continued distillation and mostly reach their lowest values at the final checkpoints. The few-shot prompt gives the same trajectory as the strict prompt for the Qwen3 and Llama-3.1 pairs, so the conclusions in Section 5.3 do not depend on the prompt style.

D.2 Response stability during the early transient

Section 5.3 reports a transient rise in DSE at the early checkpoints, which the theory allows but does not require. Figure 6 examines it. The upper panels show, for each checkpoint, how many of the ten sampled responses agree with the graded answer, with incorrectly and correctly answered questions shown as separate bars. The lower panels show how often the graded answer agrees semantically with the base-student response and with the teacher target.

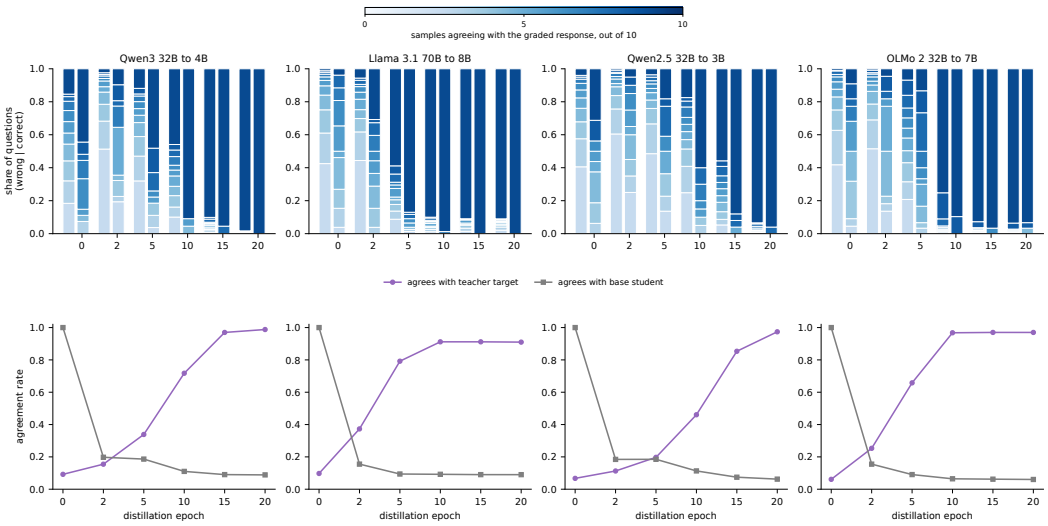


Figure 6: Response stability and semantic agreement during distillation. Upper panels show the share of questions at each level of agreement between the ten sampled responses and the graded answer, with the left bar of each pair giving incorrectly answered questions and the right bar correctly answered ones. Darker shading means more of the ten samples agree. Lower panels show the semantic agreement of the graded answer with the base-student response and with the teacher target.

At epoch 2 the sampled responses remain widely dispersed on both kinds of question. Even where the graded answer is correct, most questions have only a minority of samples agreeing with it, so some of the temporary accuracy gain at that checkpoint reflects unstable correct predictions rather than settled concentration on the correct answer. The lower panels show the same state from the other side. Agreement with the base student has already fallen from one to below 0.2 in all four pairs, while agreement with the teacher target is still around 0.15, and the two curves cross only at around epoch 5. The early peak is therefore a transitional state in which the student has left its initial distribution without yet concentrating on the recorded target.

D.3 Effect of teacher abstentions

The main experiments exclude teacher responses graded as non-attempts from the distillation corpus. We test whether that decision contributes to the dynamics reported in Section 5.3 by repeating the Qwen3 runs with those responses retained as distillation targets, under both prompt styles.

Retaining them does not materially change the training dynamics. Under both prompts the AUROC trajectories have the same shape in the two settings, improving slightly at the early checkpoints before declining towards 0.5 with continued distillation. Retaining the non-attempts leaves AUROC slightly lower at the intermediate checkpoints, so filtering them slows the decline marginally rather

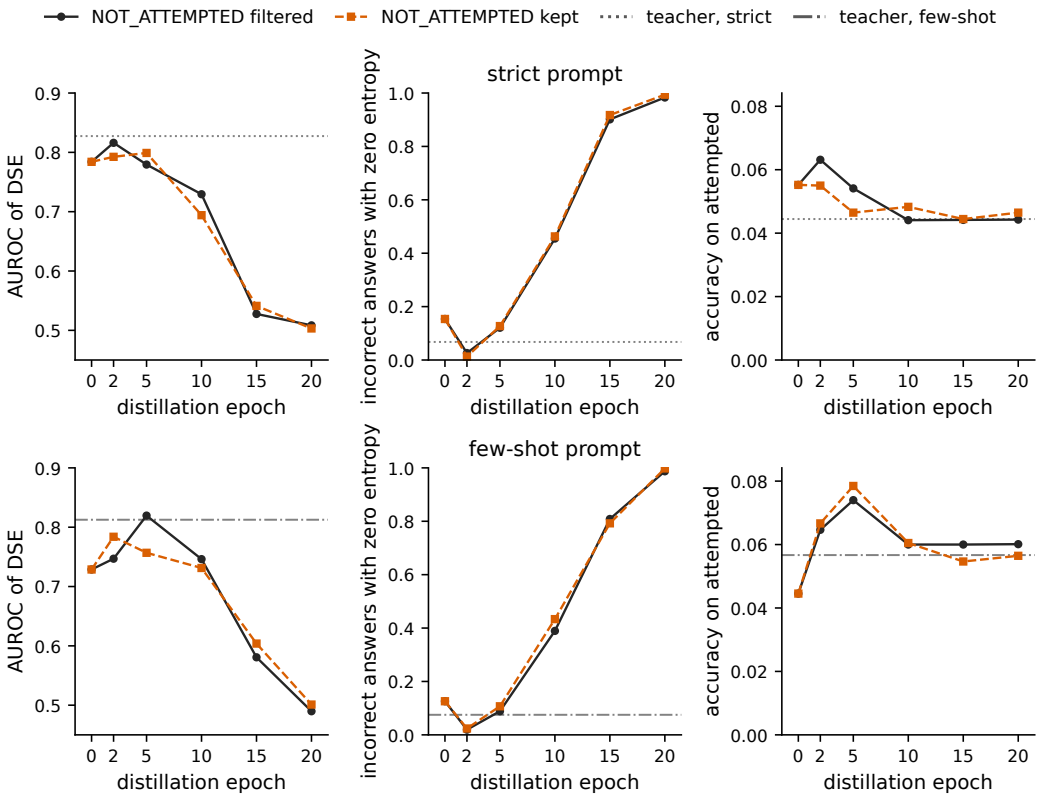


Figure 7: Effect of retaining teacher non-attempts as distillation targets for the Qwen3 pair, with the strict prompt in the upper row and the few-shot prompt in the lower row. Columns give the AUROC of DSE against incorrectness, the proportion of incorrect answers with zero DSE, and accuracy on attempted questions, with the corresponding teacher values marked.

than causing it, and the two settings reach the same value by epoch 20. The proportion of incorrect answers with zero DSE is almost identical throughout and approaches one in both settings, and accuracy on attempted questions is close throughout, with no consistent gain or loss from retaining the teacher's not-attempted answers. The loss of detectability is therefore not an artefact of filtering the teacher's abstentions.

D.4 Uncertainty estimators

The main text measures uncertainty with discrete semantic entropy, which is computed from cluster frequencies alone and, with ten samples, takes only finitely many values. A student whose ten samples all fall in one cluster receives DSE of exactly zero whether it assigns that cluster probability 0.6 or 0.99, so part of the late decline in AUROC could in principle reflect the resolution of the estimator rather than the uncertainty of the student. We test this by recomputing AUROC for the Llama-3.1 pair with two estimators that use token probabilities. Semantic entropy (Farquhar et al., 2024) weights each cluster by the summed length-normalised sequence probability of its members rather than by its count. Naive entropy is the entropy of the length-normalised sequence probabilities over the ten samples without any semantic clustering, so it treats two paraphrases of the same answer as different outcomes. All three are computed from the same ten samples per question. All three are recomputed on the subset of questions for which sequence log-probabilities were available, so the DSE values here differ slightly from those in Table 1. Figure 8 plots the three trajectories under both prompts and Table 2 gives the values.

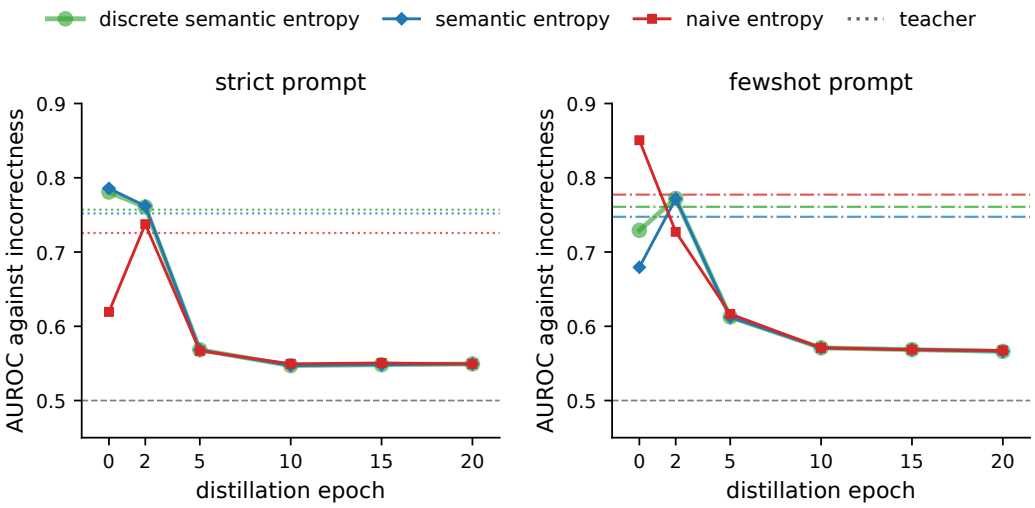


Figure 8: AUROC for the Llama-3.1 pair under three uncertainty estimators, with horizontal lines giving the teacher value of each estimator.

The collapse does not depend on the estimator. The three estimators differ only at the teacher, the base student and epoch 2, where the student's samples still spread across several meanings and the estimators differ in how they summarise that spread. From epoch 5 onwards the three curves lie on top of one another, agreeing to within 0.005 at every later checkpoint under both prompts, and all fall to between 0.55 and 0.57 by epoch 20. Semantic entropy matches DSE to three decimal places throughout this range, so reweighting clusters by probability changes nothing once the samples have concentrated, and naive entropy, which does not cluster at all and can in principle separate samples that DSE places in the same class, reaches the same value. The late decline is therefore not an artefact of the ten-sample estimator. Once the student's samples collapse onto one meaning there is no dispersion left for any estimator to read, and access to token probabilities does not recover a gap that the samples themselves no longer show.

Table 2: AUROC against incorrectness for the Llama-3.1 pair under three estimators. Best value at each checkpoint in bold.

Prompt	Checkpoint	Discrete semantic entropy	Semantic entropy	Naive entropy
strict	Teacher	0.757	0.752	0.726
	Base Student	**0.781**	**0.786**	0.619
	Epoch 2	0.761	0.762	**0.738**
	Epoch 5	0.568	0.568	0.567
	Epoch 10	0.547	0.547	0.549
	Epoch 15	0.548	0.548	0.551
	Epoch 20	0.549	0.549	0.549
fewshot	Teacher	0.761	0.747	0.777
	Base Student	0.729	0.679	**0.851**
	Epoch 2	**0.772**	**0.772**	0.727
	Epoch 5	0.612	0.612	0.617
	Epoch 10	0.571	0.571	0.571
	Epoch 15	0.569	0.569	0.569
	Epoch 20	0.566	0.566	0.567

D.5 Cross-family distillation

The four pairs in Section 5.1 distil within a model family, which keeps tokeniser and architecture fixed and isolates the effect of the corpus. In practice the teacher and student often come from different families. We therefore repeat the main experiment with Llama-3.1-70B-Instruct as teacher and Qwen3-4B-Instruct-2507 as student, with the setup of Section 5.1 unchanged, and compare the resulting trajectory against the two same-family pairs that share its teacher and its student. Figure 9 reports the strict prompt and Figure 10 the few-shot prompt; Table 3 gives the strict-prompt numbers.

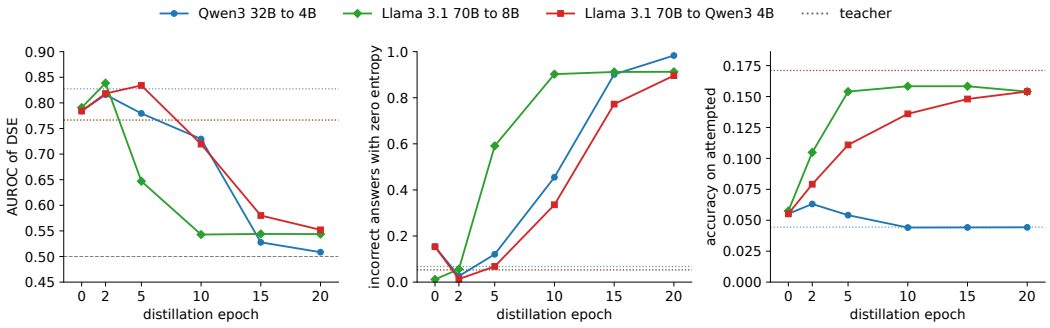


Figure 9: Cross-family distillation from Llama-3.1-70B to Qwen3-4B under the strict prompt, alongside the two same-family pairs that share its teacher (Llama-3.1-70B to 8B) and its student (Qwen3-32B to 4B). Panels give the AUROC of DSE against incorrectness, the proportion of incorrect answers with zero DSE, and accuracy on attempted questions. Horizontal lines give the corresponding teacher values.

The collapse is the same, but it takes longer. The cross-family student shows the same collapse as the same-family pairs. By epoch 20 its AUROC has fallen to 0.55 under the strict prompt and 0.57 under the few-shot prompt, against 0.55 and 0.57 for the Llama-3.1 pair and 0.51 and 0.49 for the Qwen3 pair, and the proportion of incorrect answers with zero DSE has risen to about 0.9 in every setting. What differs is the path. Within the first ten epochs the cross-family student lags well behind the same-family pair that shares its teacher. At epoch 5 its AUROC is still 0.83 under the strict prompt, at or above the base-student level, while the Llama-3.1 pair has already fallen to 0.64; only 7% of its incorrect answers have zero DSE against 59% for the Llama-3.1 pair. At epoch 10 the gap remains, with AUROC at 0.72 against 0.54 and the zero-DSE proportion at 0.34 against 0.90. The few-shot prompt shows the same ordering at every checkpoint.

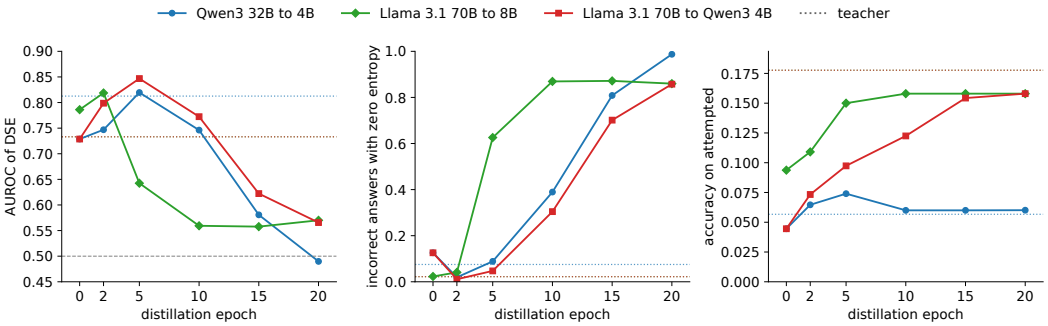


Figure 10: The same comparison under the few-shot prompt.

Table 3: Cross-family results under the strict prompt. Correct and Abstained are counts out of 500 questions.

Pair	Checkpoint	Accuracy	AUROC	Abstention	Correct	Abstained
Llama 3.1 70B to Qwen3 4B	Teacher	**0.171**	0.766	**0.088**	**78**	**44**
	Base Student	0.053	0.768	0.028	26	14
	Epoch 2	0.079	0.818	0.014	39	7
	Epoch 5	0.111	**0.834**	0.008	55	4
	Epoch 10	0.136	0.719	0.000	68	0
	Epoch 15	0.148	0.580	0.000	74	0
	Epoch 20	0.154	0.552	0.000	77	0

The student absorbs the teacher's answers more slowly. The accuracy panel explains the delay. On the corpus, the Llama-3.1 student reaches its teacher's accuracy within five epochs, whereas the cross-family student climbs gradually, from 0.053 at epoch 0 through 0.079, 0.111 and 0.136 at epochs 2, 5 and 10, and approaches the teacher's 0.171 only at epoch 20. The Qwen3 student, whose own teacher is no more accurate than itself, moves little. The cross-family student therefore inherits the Llama-3.1 teacher's answers, but a Qwen3 base model needs more updates to move onto targets generated by a different family. The transitional regime identified in Appendix D.2, in which the student has left its initial distribution without yet concentrating on the recorded target, is correspondingly longer. The temporary rise in AUROC that the same-family pairs show at epoch 2 extends to epoch 5 here, and the zero-DSE proportion, which starts higher for the base Qwen3 student at 0.14, first drops as probability mass is redistributed and only begins to climb after epoch 5.

Family mismatch changes the rate, not the outcome. Taken together, the cross-family trajectory looks like a stretched version of the same-family one. The three curves are ordered consistently within the first ten epochs, with the Llama-3.1 pair collapsing fastest and the cross-family pair slowest, which is consistent with the cross-family student having further to move before it settles on the recorded targets. Once the student has moved onto the recorded targets, uncertainty collapses and detectability falls to chance regardless of where the targets came from. The analysis of Section 4 makes no reference to the relationship between teacher and student beyond the corpus, and this is what the cross-family result reflects. Distilling across families does not protect the student; it only delays the point at which single-response supervision takes full effect.

E Mathematical Reasoning

E.1 MATH corpus and training

Data. Problems are drawn from the test split of MATH restricted to Level 5, which contains 1,324 problems across seven subjects, following the MATH-Hard definition used by the Open LLM Leaderboard v2. The distillation corpus samples 200 problems with seed 44 and the held-out set

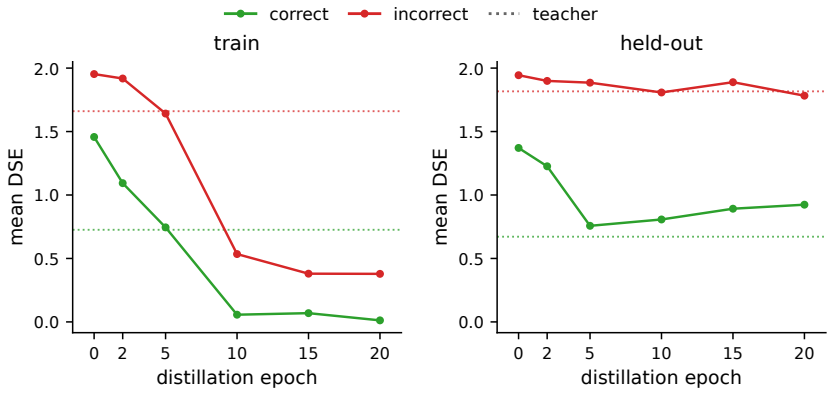


Figure 11: Semantic uncertainty on MATH across distillation checkpoints. Mean DSE is shown separately for correctly and incorrectly answered problems on the distillation corpus (left) and a disjoint held-out set (right). Dotted horizontal lines give the corresponding teacher values.

samples 200 further problems with the same seed from the remaining pool, disjoint by construction. No subject restriction is applied.

Teacher targets. The teacher is prompted zero-shot with the MATH prompt of Appendix B.2.4. Targets are generated at temperature 0.1 with the family default sampling parameters and a limit of 8,192 new tokens, and the full step-by-step solution is recorded as the distillation target, not the final answer alone. Teacher responses containing no `\boxed` expression are graded NOT_ATTEMPTED and removed from the corpus; all such responses are truncations at the token limit. This removes 35 of the 200 problems, leaving 165 problems as the distillation corpus.

Grading and semantic classes. The final answer of a response is the last `\boxed` expression it contains, and correctness is judged by symbolic equivalence with the reference answer using the math_verify library, as in the corrected MATH-Hard evaluation of the Open LLM Leaderboard. Semantic classes for DSE are formed by the same equivalence, so two samples fall in one class exactly when their final answers are symbolically equivalent, and samples from which no answer can be extracted each form their own class.

Training. Students are trained with the hyperparameters of Appendix B.1, independently from the base model for each epoch budget. To accommodate the longer targets, the maximum sequence length is increased from 1,024 to 9,216 tokens and gradient checkpointing is enabled.

Evaluation. Students and the teacher are evaluated on the original 200-problem training pool and the held-out set with ten samples at temperature 1.0 for DSE, under the same generation limit as the teacher targets. Responses without an extractable final answer are graded NOT_ATTEMPTED and excluded from accuracy and AUROC; their proportion is reported in Table 4.

E.2 MATH FULL RESULTS

Figure 11 shows the semantic-uncertainty trajectories on the distillation corpus and the disjoint held-out set, and Table 4 reports the corresponding accuracy and detection results for every checkpoint.

On the distillation corpus, mean DSE declines substantially on both correct and incorrect responses after the early checkpoints. Correct-response DSE approaches zero, while incorrect-response DSE also falls sharply. Consistent with this collapse, DSE AUROC rises initially from 0.731 for the base student to 0.818 at epoch 2, before falling to 0.600 at epoch 20. Over the same period, accuracy on attempted problems rises from 0.288 to 0.414. Thus, the divergence between task accuracy and uncertainty-based hallucination detection observed on SimpleQA also appears in mathematical reasoning.

Table 4: MATH results for the Llama-3.1 pair on the distillation and held-out sets. "No answer" is the proportion of responses without an extractable final answer, which are excluded from accuracy and AUROC. Accuracy is computed on attempted problems.

Split	Checkpoint	No answer	# No answer	Accuracy	# Correct	AUROC
Train	Teacher	0.175	35	**0.467**	77	**0.840**
	Base Student	0.185	37	0.288	47	0.731
	Epoch 2	0.265	53	0.320	47	0.818
	Epoch 5	0.205	41	0.434	69	0.815
	Epoch 10	0.050	10	0.411	78	0.644
	Epoch 15	0.060	12	0.426	**80**	0.588
	Epoch 20	0.070	14	0.414	77	0.600
Held-out	Teacher	0.190	38	**0.481**	**78**	**0.897**
	Base Student	0.205	41	0.308	49	0.758
	Epoch 2	0.290	58	0.282	40	0.740
	Epoch 5	0.300	60	0.257	36	0.896
	Epoch 10	0.245	49	0.272	41	0.858
	Epoch 15	0.290	58	0.296	42	0.870
	Epoch 20	0.235	47	0.209	32	0.797

The held-out set shows a different pattern. Incorrect-response DSE remains approximately stable across checkpoints, while correct-response DSE declines initially but shows no corresponding collapse towards zero. DSE AUROC likewise shows no sustained decline: it fluctuates across checkpoints and ends at 0.797, compared with 0.758 for the base student. Attempted accuracy falls from 0.308 to 0.209. The contrast between the distillation and held-out sets indicates that the MATH collapse transfers less strongly beyond directly fitted problems than the corresponding effect on SimpleQA. We discuss this difference together with the other off-corpus evaluations in Section 5.4.

F Generalisation Beyond the Distillation Corpus

F.1 Held-out and out-of-distribution evaluation sets

We evaluate generalisation on three sets beyond the distillation corpus. The held-out SimpleQA set contains 500 questions sampled with seed 45 from the remaining SimpleQA pool, excluding the distillation questions and five few-shot exemplars. TriviaQA (Joshi et al., 2017) is sampled with seed 44 from the validation split of the `rc.nocontext` configuration; five of the 500 sampled questions are reserved as few-shot exemplars, leaving 495 for evaluation. MMLU-Pro (Wang et al., 2024) contains 500 questions sampled with seed 44 from the test split and is evaluated in the multiple-choice format of Section 5.2, with all options shown and the model asked to return an option letter.

Measurement follows Section 5.1: one response at temperature 0.1 per question for grading and ten samples at temperature 1.0 for DSE. The teacher and base student are evaluated on every set as reference points, and each distilled student is evaluated under the prompt style used during its distillation. The main text reports the strict-prompt results; few-shot results are given in Appendix F.2.

Grading. The two SimpleQA sets use the grading procedure of Appendix B.2.3, with semantic classes formed by bidirectional entailment as in the main experiments. TriviaQA uses the benchmark's official normalisation, counting an answer as correct when its normalised form exactly matches any reference alias; semantic classes are formed by bidirectional entailment. MMLU-Pro is graded by extracting the option letter and comparing it with the reference answer, with semantic classes defined by the extracted letter. Responses without a valid option letter are graded NOT_ATTEMPTED.

Abstentions. Abstentions follow the grading rule of each benchmark. For SimpleQA, the grader returns C (not attempted). For TriviaQA and MMLU-Pro, empty responses and a fixed set of refusal expressions are treated as abstentions; on MMLU-Pro this also includes responses without a valid option letter. Abstained questions are excluded from accuracy and AUROC, as in the main experiments.

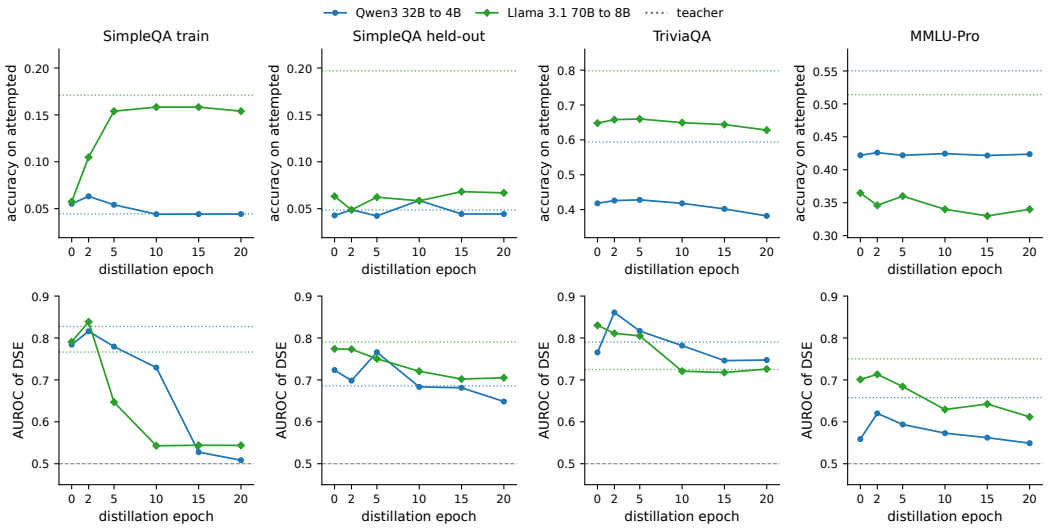


Figure 12: Generalisation beyond the distillation corpus under the strict prompt. Top row: accuracy on attempted questions. Bottom row: DSE AUROC. Epoch 0 is the base student; dotted horizontal lines give teacher values and the dashed line marks chance.

F.2 Full generalisation results

Tables 5, 6, and 7 report accuracy on attempted questions, abstention rate, and DSE AUROC for every checkpoint of the Qwen3 and Llama-3.1 pairs on held-out SimpleQA, TriviaQA, and MMLU-Pro under both prompt styles. Each student is evaluated under the prompt style used during its distillation. The strict-prompt accuracy and AUROC results correspond to Figure 12, while Figures 13 and 14 report mean DSE separately on correct and incorrect responses under the strict and few-shot settings. Results on the distillation corpus are given in Table 1.

Semantic uncertainty beyond the distillation corpus. Figures 13 and 14 show a consistent pattern across the two prompt styles. The strongest compression occurs on the distillation corpus, where mean DSE on both correct and incorrect responses approaches zero with continued training. After the early transient, held-out SimpleQA and TriviaQA also show substantial reductions in mean DSE on both groups, but retain considerably more uncertainty than the distillation corpus. MMLU-Pro starts from lower DSE and shows smaller absolute changes, although compression remains visible, particularly for Llama-3.1. Thus, the off-corpus changes in DSE AUROC are accompanied by compression of semantic uncertainty itself rather than only by changes in the relative ordering of uncertainty scores.

Few-shot accuracy and detectability. The few-shot results broadly reproduce the strict-prompt findings. Off-corpus accuracy shows no systematic improvement with continued distillation. DSE AUROC also tends to decline after the early checkpoints, although the magnitude varies across settings. Qwen3 on held-out SimpleQA remains comparatively stable, while larger declines occur in several other model–dataset combinations. The base Llama-3.1 student under the few-shot prompt abstains on 0.778 of held-out SimpleQA questions, so its accuracy and AUROC are computed on a small attempted subset.

G Intervention Details and Additional Results

G.1 Entropy thresholds for filtering and replacement

Since DSE is estimated from ten sampled responses it takes finitely many values, determined by how the ten samples partition into semantic classes. So each threshold selects whole entropy levels rather than cutting through a continuous range. Figure 15 shows the resulting distribution of teacher

Table 5: Held-out SimpleQA results for both prompt styles. Best accuracy and AUROC within each model–prompt block are shown in bold; the abstention rate has no preferred direction and is left unmarked.

Model family	Prompt	Checkpoint	Accuracy	Abstention	AUROC
Qwen3	strict	Teacher	0.049	0.012	0.686
		Base	0.043	0.018	0.723
		Epoch 2	0.049	0.014	0.698
		Epoch 5	0.042	0.006	**0.766**
		Epoch 10	**0.059**	0.012	0.683
		Epoch 15	0.044	0.006	0.681
		Epoch 20	0.044	0.006	0.648
	fewshot	Teacher	0.050	0.002	0.698
		Base	**0.070**	0.114	0.696
		Epoch 2	0.057	0.012	0.760
		Epoch 5	0.053	0.010	**0.782**
		Epoch 10	0.064	0.002	0.735
		Epoch 15	0.054	0.004	0.751
		Epoch 20	0.054	0.004	0.717
Llama-3.1	strict	Teacher	**0.197**	0.056	**0.791**
		Base	0.063	0.082	0.774
		Epoch 2	0.049	0.014	0.773
		Epoch 5	0.062	0.004	0.750
		Epoch 10	0.058	0.006	0.721
		Epoch 15	0.068	0.002	0.702
		Epoch 20	0.067	0.012	0.705
	fewshot	Teacher	**0.195**	0.096	0.780
		Base	0.108	0.778	0.718
		Epoch 2	0.067	0.022	**0.818**
		Epoch 5	0.058	0.008	0.721
		Epoch 10	0.052	0.002	0.724
		Epoch 15	0.053	0.012	0.763
		Epoch 20	0.061	0.012	0.652

Table 6: TriviaQA results for both prompt styles. Columns as in Table 5. Best accuracy and AUROC within each model–prompt block are shown in bold.

Model family	Prompt	Checkpoint	Accuracy	Abstention	AUROC
Qwen3	strict	Teacher	**0.594**	0.000	0.790
		Base	0.418	0.000	0.766
		Epoch 2	0.426	0.000	**0.861**
		Epoch 5	0.428	0.000	0.817
		Epoch 10	0.418	0.000	0.782
		Epoch 15	0.402	0.000	0.746
		Epoch 20	0.382	0.000	0.747
	fewshot	Teacher	**0.450**	0.000	0.743
		Base	0.418	0.004	0.789
		Epoch 2	0.422	0.000	**0.842**
		Epoch 5	0.435	0.002	0.824
		Epoch 10	0.412	0.000	0.828
		Epoch 15	0.398	0.000	0.793
		Epoch 20	0.390	0.000	0.758
Llama-3.1	strict	Teacher	**0.798**	0.000	0.725
		Base	0.648	0.000	**0.830**
		Epoch 2	0.658	0.000	0.811
		Epoch 5	0.660	0.000	0.805
		Epoch 10	0.649	0.002	0.721
		Epoch 15	0.644	0.000	0.718
		Epoch 20	0.628	0.000	0.726
	fewshot	Teacher	**0.800**	0.000	0.827
		Base	0.588	0.000	**0.867**
		Epoch 2	0.662	0.000	0.843
		Epoch 5	0.662	0.000	0.689
		Epoch 10	0.652	0.000	0.736
		Epoch 15	0.629	0.002	0.733
		Epoch 20	0.621	0.002	0.704

Table 7: MMLU-Pro results for both prompt styles. Columns as in Table 5. Best accuracy and AUROC within each model–prompt block are shown in bold.

Model family	Prompt	Checkpoint	Accuracy	Abstention	AUROC
Qwen3	strict	Teacher	**0.550**	0.004	**0.658**
		Base	0.422	0.000	0.559
		Epoch 2	0.426	0.000	0.620
		Epoch 5	0.422	0.000	0.594
		Epoch 10	0.425	0.006	0.573
		Epoch 15	0.422	0.004	0.562
		Epoch 20	0.424	0.004	0.549
	fewshot	Teacher	**0.550**	0.004	**0.670**
		Base	0.422	0.000	0.560
		Epoch 2	0.416	0.000	0.625
		Epoch 5	0.417	0.002	0.586
		Epoch 10	0.415	0.002	0.581
		Epoch 15	0.413	0.002	0.561
		Epoch 20	0.409	0.002	0.560
Llama-3.1	strict	Teacher	**0.514**	0.004	**0.750**
		Base	0.365	0.002	0.701
		Epoch 2	0.346	0.000	0.713
		Epoch 5	0.360	0.000	0.684
		Epoch 10	0.340	0.000	0.629
		Epoch 15	0.330	0.000	0.642
		Epoch 20	0.340	0.000	0.612
	fewshot	Teacher	**0.504**	0.004	**0.753**
		Base	0.359	0.002	0.706
		Epoch 2	0.350	0.000	0.716
		Epoch 5	0.350	0.000	0.671
		Epoch 10	0.358	0.000	0.630
		Epoch 15	0.341	0.002	0.642
		Epoch 20	0.340	0.000	0.617

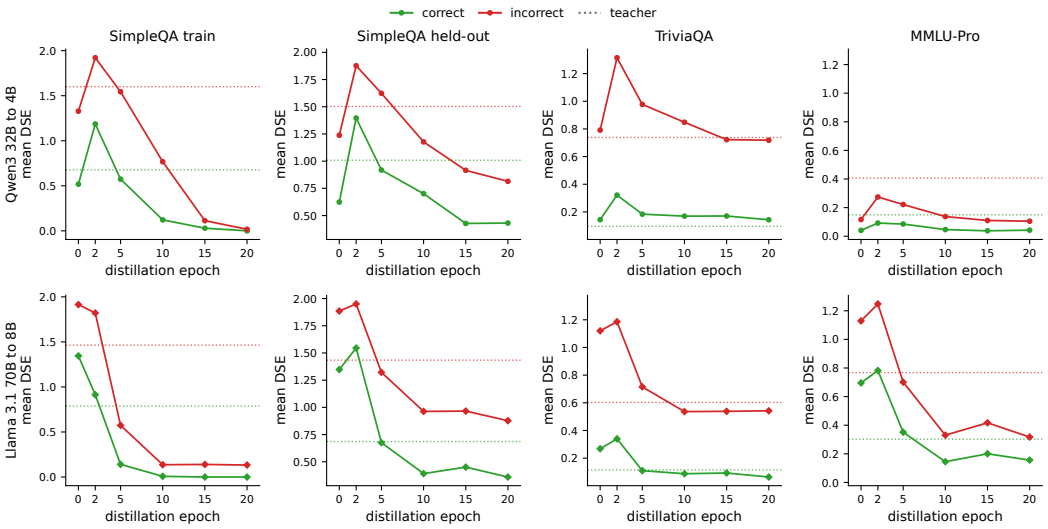


Figure 13: Semantic uncertainty beyond the distillation corpus under the strict prompt. Mean DSE is shown separately on correctly and incorrectly answered questions for the distillation corpus, held-out SimpleQA, TriviaQA, and MMLU-Pro. The top and bottom rows show the Qwen3 and Llama-3.1 pairs, respectively; epoch 0 is the base student and dotted horizontal lines give the corresponding teacher values.

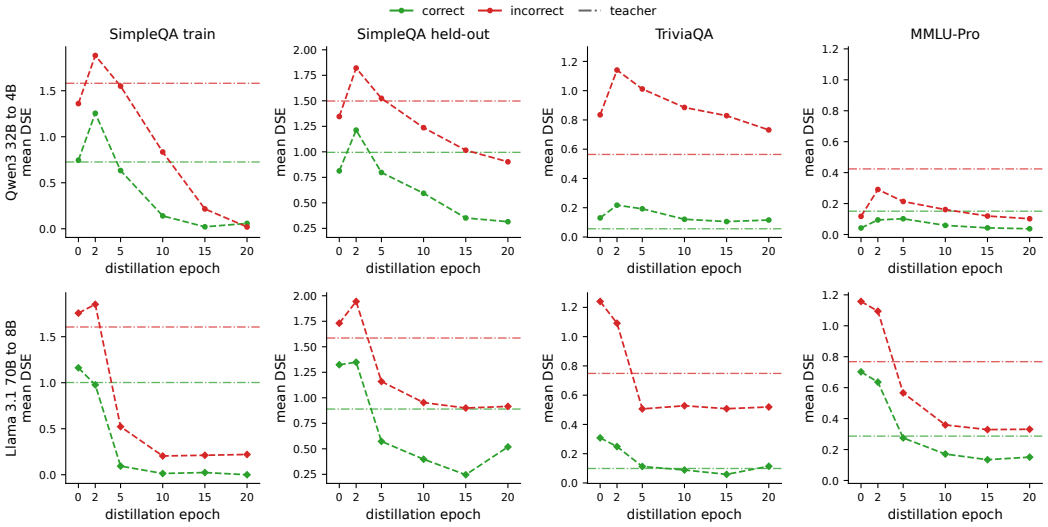


Figure 14: Semantic uncertainty beyond the distillation corpus under the few-shot prompt. Layout and notation follow Figure 13.

DSE for the Llama-3.1 pair. Each threshold coincides with a realisable entropy level, quoted to four decimal places, and a question is selected when its teacher DSE strictly exceeds the quoted value. The rounding direction differs across levels. The loose value rounds its level up, so selection at this level keeps only questions strictly above it, which under ten samples is exactly the set whose ten responses are pairwise distinct. The medium value rounds its level down, so questions sitting exactly at that level are also selected, and the medium setting therefore adds the next two levels, 2.1640 and 2.0253. No eligible question falls exactly on the tight level, so the distinction does not arise there, and the tight setting selects every level down to 1.8867. The three levels extend progressively further into the high-entropy region. The same thresholds are used for both prompt styles, since the set of realisable values depends only on the ten-sample estimator. Teacher responses graded

as NOT_ATTEMPTED are excluded before thresholding, leaving 456 eligible examples under the strict prompt and 439 under the few-shot prompt.

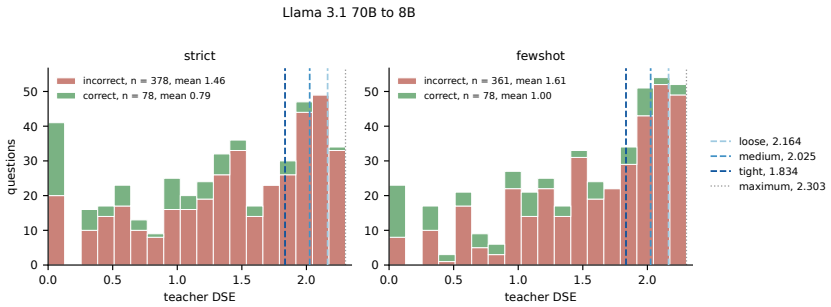


Figure 15: Distribution of the teacher's DSE under the strict and few-shot prompts, with the three thresholds marked.

Table 8: Entropy thresholds and the resulting fraction of teacher targets selected under each setting.

Filtering level	Entropy threshold	Strict dropped	Few-shot dropped
Loose	2.1640	34 (7.5%)	52 (11.8%)
Medium	2.0253	112 (24.6%)	140 (31.9%)
Tight	1.8344	141 (30.9%)	161 (36.7%)

G.2 Interventions under the Few-Shot Prompt

Figures 16 to 18 repeat the three interventions of Section 6 under the few-shot prompt. Every conclusion drawn in the main text also holds here. Filtering slows the decline in AUROC at the later checkpoints without preventing it, with the tight setting preserving the most. Replacement leaves AUROC falling towards 0.5 while raising AURAC and accuracy on attempted questions, and the abstention rate again overshoots at epoch 2 before converging to the fraction of replaced targets. Sampling multiple teacher responses keeps AUROC in the same range as under the strict prompt and keeps the proportion of incorrect answers with zero DSE low throughout training.

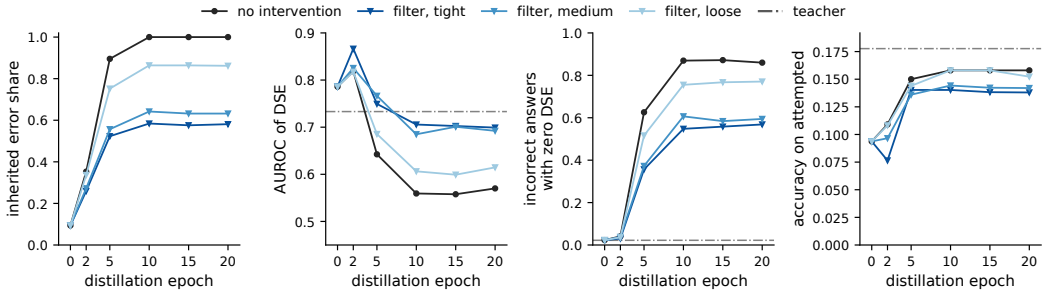


Figure 16: Effect of removing high-entropy teacher examples under the few-shot prompt.

G.3 Interventions on the Qwen3 Pair

The three interventions of Section 6 were also run on the Qwen3-32B to Qwen3-4B-Instruct-2507 pair, under both prompt styles, with the setup of Section 5.1. For this pair, each threshold for filtering and replacement is placed at the midpoint between two adjacent realisable DSE values, so that no cutoff falls on a realisable value. This gives 2.2333, 1.9992 and 1.7219 for the loose, medium and tight settings. The loose and medium settings therefore select the same entropy levels as the Llama-3.1 thresholds of Appendix G.1, while the tight setting extends two levels further,

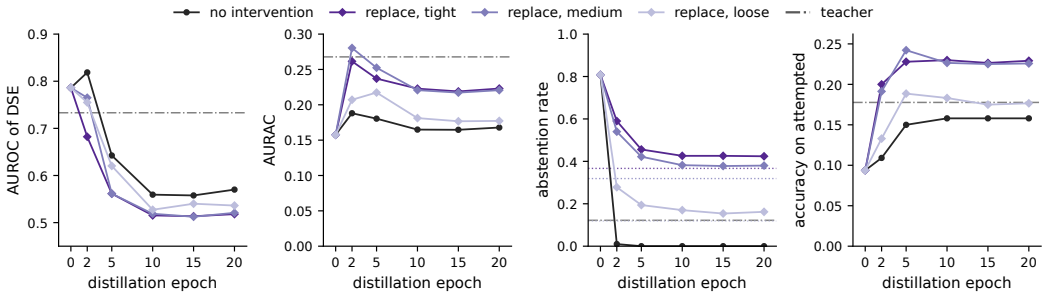


Figure 17: Effect of replacing high-entropy teacher targets with an abstention under the few-shot prompt. Horizontal lines in the abstention panel give the proportion of targets replaced at each threshold.

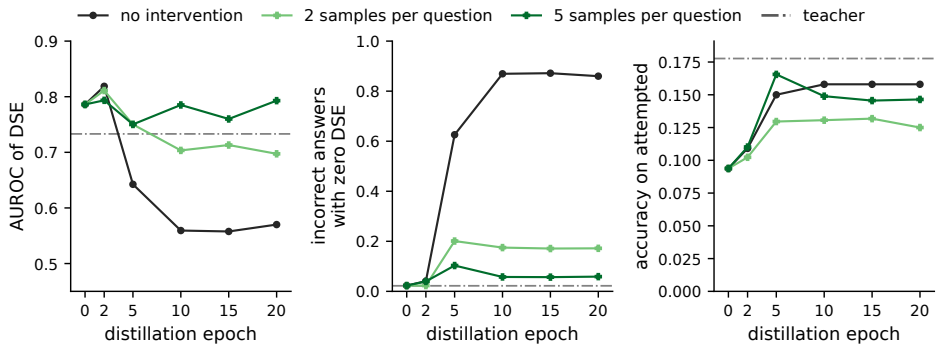


Figure 18: Effect of retaining multiple teacher responses per question under the few-shot prompt.

down to 1.7481. The same thresholds are used for both prompt styles. Teacher responses graded as NOT_ATTEMPTED are also excluded before thresholding, leaving 495 eligible examples under the strict prompt and 494 under the few-shot prompt. Figure 19 shows the teacher DSE distribution under both prompts with the three thresholds marked, and Table 9 gives the fraction of eligible targets selected at each level. Multi-response distillation retains two or five teacher responses per question.

Figures 20 to 25 report the results. Every conclusion drawn for the Llama-3.1 pair also holds here. Filtering slows the loss of detectability at the later checkpoints without preventing it, and the tight setting preserves the most. Replacement leaves AUROC falling towards 0.5 at every level while raising AURAC and accuracy on attempted questions, and the abstention rate overshoots at epoch 2 before converging to the replaced fraction. Retaining two and five teacher responses keeps AUROC between 0.73 and 0.83 throughout training and keeps the proportion of incorrect answers with zero DSE low, so the intervention findings are not specific to one model pair.

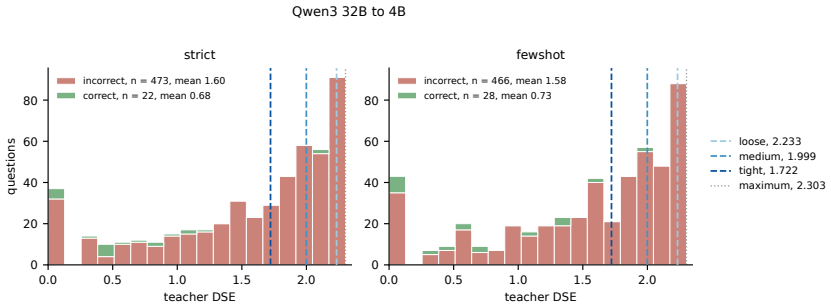


Figure 19: Distribution of the Qwen3 teacher's DSE under the strict and few-shot prompts, with the three thresholds marked.

Table 9: Entropy thresholds for the Qwen3 pair and the resulting fraction of teacher targets selected under each setting.

Filtering level	Entropy threshold	Strict dropped	Few-shot dropped
Loose	2.2333	91 (18.4%)	88 (17.8%)
Medium	1.9992	187 (37.8%)	178 (36.0%)
Tight	1.7219	261 (52.7%)	245 (49.6%)

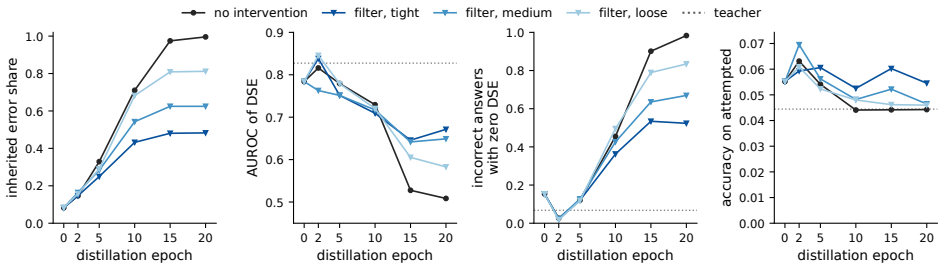


Figure 20: Effect of removing high-entropy teacher examples for the Qwen3 pair under the strict prompt.

H Extended Related Work

H.1 Uncertainty propagation and hallucination detectability

Cui & Pei (2026) provide the closest study to ours. They analyse how uncertainty from teacher outputs, student initialisation, and student outputs contributes to uncertainty in distilled students.

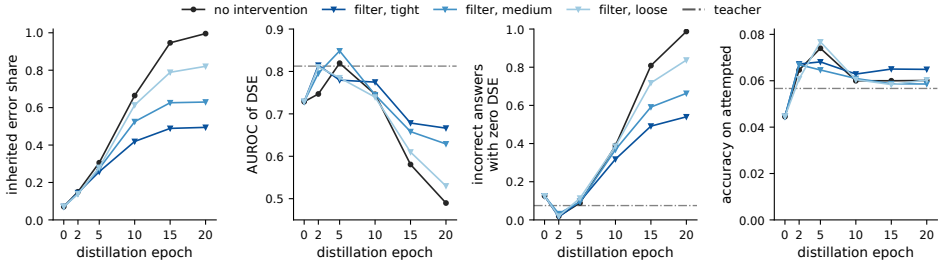


Figure 21: Effect of removing high-entropy teacher examples for the Qwen3 pair under the few-shot prompt.

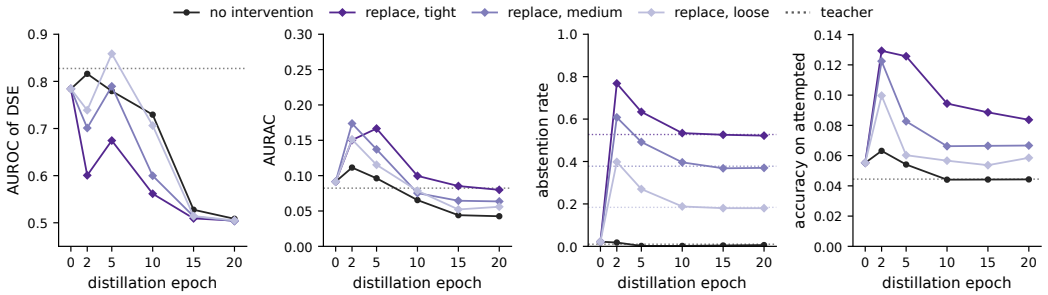


Figure 22: Effect of replacing high-entropy teacher targets with an abstention for the Qwen3 pair under the strict prompt. Horizontal lines in the abstention panel give the proportion of targets replaced at each threshold.

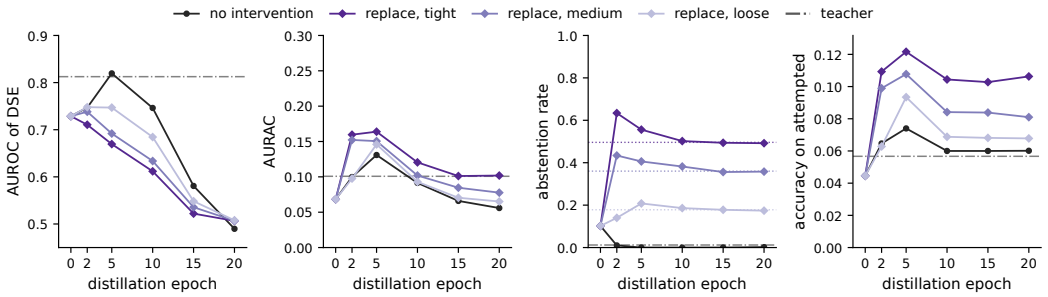


Figure 23: Effect of replacing high-entropy teacher targets with an abstention for the Qwen3 pair under the few-shot prompt. Horizontal lines in the abstention panel give the proportion of targets replaced at each threshold.

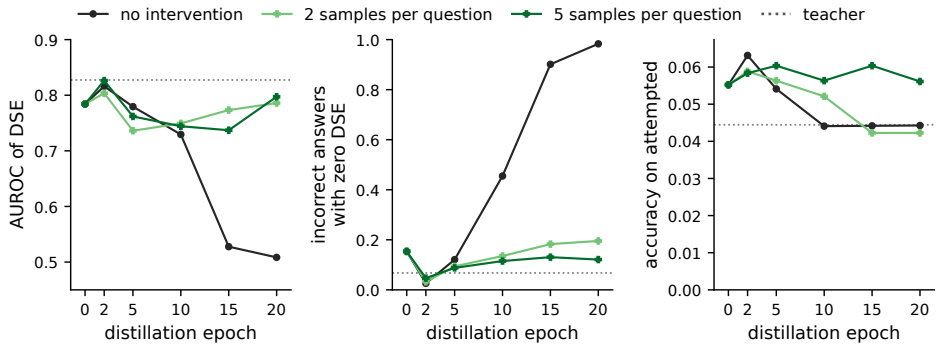


Figure 24: Effect of retaining multiple teacher responses per question for the Qwen3 pair under the strict prompt.

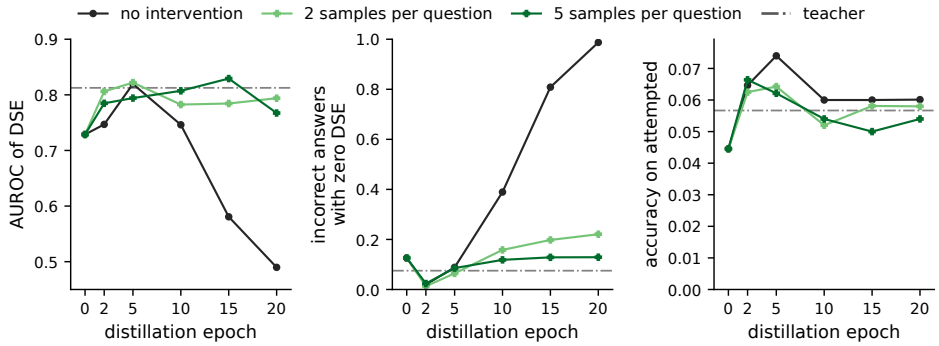


Figure 25: Effect of retaining multiple teacher responses per question for the Qwen3 pair under the few-shot prompt.

Their framework distinguishes intra-student uncertainty, which measures variability within a fixed student's predictive distribution, from inter-student uncertainty, which measures variability across independently distilled students. They find that standard single-response distillation suppresses intra-student uncertainty while leaving substantial sensitivity to student initialisation, and attribute the resulting uncertainty mismatch primarily to noisy teacher outputs and a distillation signal that is insufficient to eliminate initialisation-induced differences. They propose averaging and variance weighting to reduce this mismatch, and additionally evaluate multiple teacher responses as a baseline.

Our study examines a complementary aspect of this phenomenon. We compare students trained independently from the same base model to increasing distillation budgets, characterising how semantic uncertainty changes with continued training and studying its consequences for hallucination detection. The central question is whether semantic uncertainty remains informative for identifying incorrect answers as distillation proceeds. We evaluate this using the AUROC of semantic entropy against incorrectness: higher AUROC indicates that semantic entropy more effectively distinguishes incorrect from correct answers, while an AUROC of 0.5 indicates chance-level discrimination. Across increasing distillation budgets, this measure declines towards chance even when answer accuracy changes little.

Our analysis focuses on the information available in the distillation corpus. A single sampled response does not identify the teacher's dispersion over alternative semantic outcomes. Proposition 2 formalises this non-identifiability, while Propositions 1 and 3 characterise the concentration pressure induced by fitting the fixed observation. Retaining multiple teacher responses exposes alternative outcomes that are absent from a single-response corpus, allowing us to test whether preserving this distributional information also preserves hallucination detectability.

H.2 Calibration and overconfidence under distillation

Several recent studies show that distillation can produce poorly calibrated or overconfident students, but study different aspects of this behaviour. Preference distillation replaces reliance on individual teacher outputs with teacher–student preference pairs to improve sequence-level calibration (Zhang et al., 2024). Nguyen et al. (2025) contrast deterministic supervision with smoothed teacher targets and show that retaining soft teacher information can reduce overconfidence and hallucination. Zhan et al. (2026) identify overconfident hallucination among broader distillation failures associated with factors including teacher–student mismatch, tail noise, and off-policy training. In an on-policy setting, Zhang et al. (2026) show that capability and calibration can become decoupled and relate the resulting overconfidence to information available when supervision is formed but unavailable to the student at deployment.

Calibration concerns the correspondence between model confidence and empirical correctness. Here we study whether semantic uncertainty remains useful for identifying incorrect answers after distillation. These properties capture different aspects of model reliability: calibration describes how confidence relates to empirical correctness, whereas uncertainty-based hallucination detection asks whether uncertainty can identify unreliable responses. We measure the latter using the AUROC of semantic entropy against incorrectness.

Our setting also exposes a distinct source of information loss. In standard off-policy SeqKD, teacher responses are generated in advance and held fixed throughout student training. A teacher may place probability mass across several semantic outcomes, but a single-response corpus retains only one realisation from that distribution. The missing distributional information is therefore already absent before student optimisation begins. Our theoretical analysis characterises this corpus-level bottleneck, and the interventions in Section 6 vary what information about teacher uncertainty is retained during corpus construction.

H.3 Preserving distributional information

The loss of distributional information under hard supervision is closely related to the original motivation for knowledge distillation. Classical KD transfers teacher probabilities rather than only hard labels (Hinton et al., 2015), and subsequent methods have developed alternative objectives for matching teacher distributions. In LLM distillation, methods such as MiniLLM (Gu et al., 2024) and

DistiLLM (Ko et al., 2024) modify the divergence or sampling procedure used to transfer teacher behaviour. These methods illustrate the broader value of retaining information beyond a single sampled sequence.

Multiple sampled responses provide another way to expose distributional information when teacher logits are unavailable. Phillips et al. (2026), for example, construct semantic distributions from multiple generations and explicitly distill them into an uncertainty predictor for hallucination detection. More broadly, uncertainty-aware distillation methods use richer teacher signals to improve uncertainty estimation or calibration. These approaches establish a connection between the information represented in the supervision signal and the uncertainty retained by the student.

Our analysis connects this loss of distributional information to the behaviour of standard SeqKD under continued training and to the reliability of uncertainty-based hallucination detection. With one fixed sampled response per input, the corpus provides no direct information about the teacher's dispersion over alternative semantic outcomes; students trained to larger distillation budgets increasingly concentrate on the recorded responses while their semantic uncertainty collapses. Multi-response distillation provides a direct intervention on this information bottleneck. Retaining independent teacher responses exposes alternative semantic outcomes that a single target cannot represent and progressively preserves hallucination detectability as the number of responses increases. In contrast, filtering uncertain examples and replacing them with deterministic abstention targets modify which responses the student is trained to reproduce without retaining the teacher's response distribution.