TASK-SPECIFIC KNOWLEDGE DISTILLATION VIA INTERMEDIATE PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard knowledge distillation from large language models (LLMs) assumes the teacher's output distribution is a high-quality training signal. On reasoning tasks, this assumption is frequently violated as a model's intermediate representations may encode the correct answer, yet this information is lost or distorted through the vocabulary projection, where prompt formatting and answer-token choices create a brittle, noisy interface.

We introduce PROBE-KD, a distillation framework that bypasses this bottleneck by training lightweight probes on frozen teacher hidden states and using the probe's predictions, rather than output logits, as supervision for student training. This simple change yields consistent improvements across four reasoning benchmarks (AQuA-RAT, ARC Easy/Challenge, and MMLU), with gains most pronounced under limited data.

The key mechanism is that probes trained on intermediate representations provide cleaner labels than the teacher's own outputs, effectively denoising the distillation signal. PROBE-KD requires no architectural changes to student or teacher, is architecture-agnostic, and adds minimal compute since probe training is cheap and teacher representations can be cached. By tapping into internal representations, PROBE-KD enables practitioners to extract more value from large teacher models without additional training data or architectural complexity.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable performance across diverse reasoning tasks, yet deploying these models at scale remains prohibitively expensive. Knowledge distillation (Hinton et al., 2015) offers a path forward by training compact student models to mimic larger teachers on specific target domains. The dominant paradigm matches the student's output distribution to the teacher's predicted probabilities (Sanh et al., 2020; Jiao et al., 2020; Sun et al., 2019).

In this work, we address the challenge of combining domain-specific annotations alongside model distillation to substantially improve the performance of small BERT-style classifiers on domain-specific benchmarks. We find that our approach substantially outperforms both standard distillation and classical supervised approaches.

The need for additional annotations is clear, as general-purpose reasoners, LLMs often demonstrate limited success on new benchmarks, and when used as teachers, their output becomes a noisy form of supervision. On multiple-choice reasoning tasks, LLMs frequently assign probability mass to incorrect answers, not because their internal representation is insufficiently rich, but because the mapping from internal representations to specific answer tokens (A, B, C, D) is suboptimal for these benchmarks. The teacher's output layer was optimized for general next-token prediction, not for expressing task-specific knowledge.

Indeed, prior work has demonstrated that LLM hidden states encode substantially richer information than their outputs reveal: probes can recover latent knowledge even when models output incorrect answers (Burns et al., 2024; Azaria & Mitchell, 2023), and internal representations encode task-relevant structure that the output layer fails to express (Zou et al., 2025; Alain & Bengio, 2018).

We exploit this observation for more effective knowledge transfer. Our approach, PROBE-KD (Probe-based Knowledge Distillation), is a simple two-stage procedure. In Stage 1, we extract hidden states

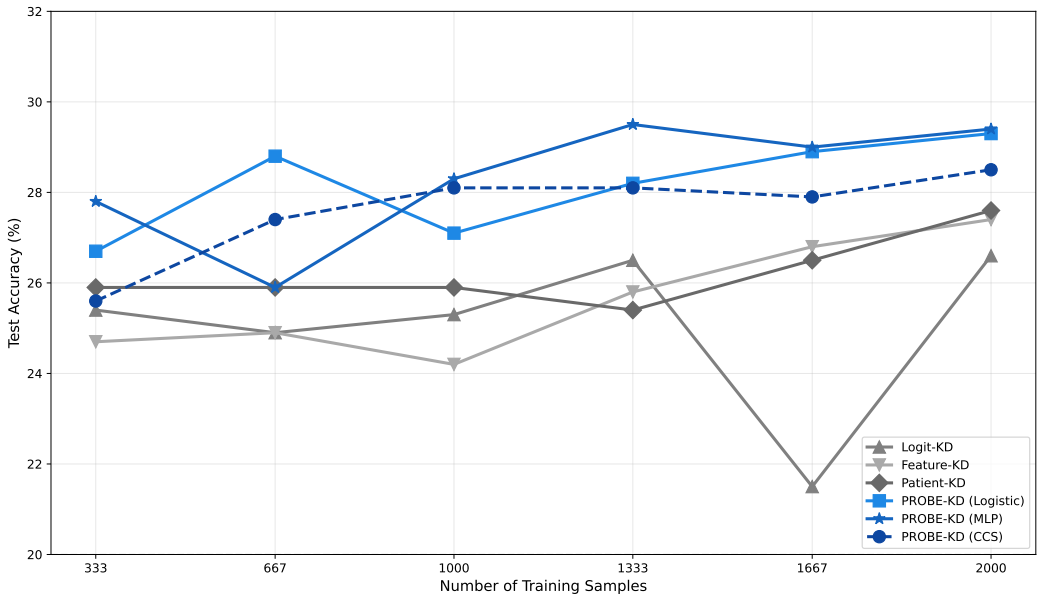

Figure 1: Data efficiency comparison on AQuA-RAT. Test accuracy (%) as a function of training data percentage for different knowledge distillation methods. We distill from Qwen2.5-7B-Instruct (teacher) to DeBERTa-v3-base (student, 86M parameters). PROBE-KD variants (Logistic, MLP, and CCS) consistently outperform standard distillation baselines (Logit-KD, Feature-KD, Patient-KD) across all data regimes, with gains most pronounced in low-data settings.

from all layers of the teacher for each training example, and train a lightweight MLP probe to predict task labels from these representations. In Stage 2, we freeze the probe, compute its soft predictions for each example, and use these as supervision for training a compact student model via KL-divergence distillation.

Compared to standard distillation, PROBE-KD allows exploitation of additional ground-truth annotations for almost no additional compute. Unlike fine-tuning a model, the most compute intensive component of PROBE-KD is the generation of hidden states, which is a part of generating model outputs for standard distillation. Furthermore, when labels are unavailable, unsupervised probing methods such as CCS (Burns et al., 2024) can replace supervised probe training.

PROBE-KD yields two complementary benefits over supervised learning with hard labels. First, soft probability distributions provide richer supervision per example: beyond indicating which answer is correct, they encode which incorrect answers are plausible, capturing inter-class relationships that improve generalization (Hinton et al., 2015; Müller et al., 2020; Yuan et al., 2021). Mandal et al. (2024) formalize this advantage, proving that soft-label training requires $O(1/\gamma^2)$ neurons to achieve a given loss versus $O(1/\gamma^4)$ for hard labels, a substantial benefit when student capacity is limited. Second, soft labels provide implicit regularization that prevents overconfidence and improves calibration (Zhou et al., 2021; Szegedy et al., 2015). These benefits compound in our setting, where task-specific data is limited and student capacity is severely constrained.

Compared to standard logit-based distillation, PROBE-KD addresses noise from the teacher's output layer. This final projection was never optimized for the downstream task, it maps rich internal representations to arbitrary answer tokens through a decoder trained for next-token prediction. Teacher logits therefore mix useful distributional information with decoder noise. PROBE-KD bypasses this bottleneck by learning a task-specific decoder (the probe) directly on hidden states, producing soft labels that retain dark knowledge while eliminating output-layer artifacts.

The probe does not inject new knowledge, it can only decode information already present in the teacher's representations. Evidence for this comes from probe accuracy: on AQuA-RAT, an MLP probe achieves 52% versus 45% for the teacher's own outputs (Chételat et al., 2025). This gap would be impossible if hidden states did not allow recovery of the correct answer, since the probe has access

only to the teacher's internal representation, not the original input. The training labels serve solely to learn the optimal projection from latent space to label space (Alain & Bengio, 2018). This aligns with findings that LLMs encode knowledge they fail to output correctly (Burns et al., 2024; Azaria & Mitchell, 2023); the probe simply provides better task alignment.

Our approach is inherently task-specific, with the probe trained on task data and its outputs reflecting task-relevant structure. This is consistent with evidence that task-specific distillation outperforms task-agnostic approaches (Jiao et al., 2020; Liu et al., 2022), particularly for compact models where capacity must be allocated carefully. Multiple-choice reasoning is not merely an evaluation format but a common inference primitive: classification, reranking, and decision-support systems all reduce to selecting among constrained options, making compact specialists for this task structure broadly applicable.

We evaluate PROBE-KD against supervised learning and standard logit distillation on four multiple-choice reasoning benchmarks: AQuA-RAT, ARC-Challenge, ARC-Easy, and MMLU. Here, PROBE-KD achieves state-of-the-art knowledge distillation (Figure 1). Our contributions include:

- We introduce PROBE-KD, a distillation framework that fuse domain-specific annotations with LLM internal states via probe predictions. These predictions are used as soft supervision, combining dark knowledge transfer with task-specific optimization.
- We provide a conceptual framework distinguishing *latent knowledge* (contained in the hidden states) from *the teacher's answers* (outputs), showing that distilling the former yields superior students.
- PROBE-KD We demonstrate that probe architecture impacts distillation quality. MLP probes consistently outperform linear probes, suggesting sufficient capacity is necessary to decode task-relevant structure from hidden states

2 RELATED WORK

Knowledge Distillation. Knowledge distillation trains student networks to match the teacher's soft labels, leveraging the richer supervision they provide over hard labels (Hinton et al., 2015; Buciluǎ et al., 2006). Extensions include feature-based methods that align intermediate representations (Romero et al., 2015), attention transfer (Zagoruyko & Komodakis, 2017), and relational approaches that preserve pairwise similarities (Park et al., 2019; Tian et al., 2022). For language models, DistilBERT (Sanh et al., 2020) and TinyBERT (Jiao et al., 2020) combine logit matching with hidden-state alignment, while Patient Knowledge Distillation (Sun et al., 2019) distills from multiple intermediate layers. MiniLM (Wang et al., 2020) transfers self-attention distributions rather than hidden states. Recent work extends these ideas to LLMs: MiniLLM (Gu et al., 2025) uses reverse KL divergence to avoid overestimating low-probability tokens, while Distilling Step-by-Step (Hsieh et al., 2023) extracts rationales alongside labels.

Feature-based methods require the student to replicate the teacher's hidden states directly (e.g., minimizing $\|h_{\text{student}} - h_{\text{teacher}}\|^2$). This creates architectural coupling. The student and teacher must share compatible hidden dimensions, and assumes teacher representations are optimal targets for the student. PROBE-KD avoids both issues: we train a probe *on top of* teacher hidden states to produce improved soft labels, then distill from the probe's output distribution. The student never sees teacher hidden states, enabling arbitrary student architectures while leveraging the probe's denoising effect.

Chain-of-Thought Distillation. A parallel line of work distills reasoning capabilities by training students on teacher-generated rationales (Wei et al., 2023; Magister et al., 2023; Shridhar et al., 2023). These methods require the teacher to produce explicit chain-of-thought explanations, which the student learns to generate before answering. Deng et al. (2023) propose implicit chain-of-thought reasoning, distilling reasoning into vertical computation across layers rather than horizontal token generation. Ho et al. (2023) show that fine-tuning on LLM-generated rationales can enable small models to outperform few-shot prompted large models. While effective for tasks where rationales are available or can be generated, these approaches do not apply to settings with only answer labels. PROBE-KD operates in the latter regime: we require only hidden states and task labels, making our approach applicable to any classification task without rationale annotation.

Probing Neural Networks. Linear probes have been extensively used to understand what neural networks encode (Alain & Bengio, 2018; Belinkov & Glass, 2019). For transformers, probing reveals syntactic structure (Hewitt & Manning, 2019), semantic roles (Tenney et al., 2019), and world knowledge (Petroni et al., 2019). Hewitt & Liang (2019) introduce control tasks to measure probe selectivity, showing that high probe accuracy does not necessarily indicate information is encoded, powerful probes can learn the task themselves. This concern is less relevant for our setting given that we want the probe to solve the task well, using whatever information the hidden states provide. The question is not whether the probe memorizes versus extracts, but whether probe-derived soft labels improve student training, an empirical question we answer affirmatively.

Recent work extends probing to factual knowledge (Meng et al., 2023) and reasoning (Stolfo et al., 2023). We build on this tradition but shift the goal. Rather than using probes to *analyze* what models encode, we use them to *extract* improved training signal for distillation.

Latent Knowledge in LLMs. Several studies demonstrate that LLM hidden states encode richer information than their outputs reveal. Burns et al. (2024) show that unsupervised probes can recover latent knowledge even when models output incorrect answers, achieving accuracy above zero-shot baselines. Azaria & Mitchell (2023) find that internal states encode whether the model is generating truthful content, independent of the actual output. The logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2025) decode hidden states at intermediate layers into vocabulary distributions, revealing how predictions are refined across layers. Li et al. (2024) show that middle layers often contain the most task-relevant features, not final layers. These works establish that a gap exists between what LLMs encode and what they output. We are the first to exploit this gap for knowledge distillation. Prior work uses probes and lenses for analysis and interpretation; we show that probe predictions provide superior supervision for training compact student models. The tuned lens decodes to the vocabulary space using the model's own unembedding matrix; our probe decodes to task-specific label space using learned projections trained on task data. The teacher's unembedding matrix was optimized for next-token prediction, not for the downstream task, introducing noise that our probe avoids.

3 METHOD: PROBE-KD

We now describe PROBE-KD in detail. Let $\mathcal{T}$ denote a large teacher model with L layers, and $\mathcal{S}$ a compact student model. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of input-label pairs, our goal is to train $\mathcal{S}$ to perform well on the task while being significantly smaller than $\mathcal{T}$.

3.1 STAGE 1: PROBE TRAINING

For each input x, we extract hidden states from all L layers of $\mathcal{T}$ and concatenate them: $\mathbf{h} = [\mathbf{h}^{(1)}; \mathbf{h}^{(2)}; \ldots; \mathbf{h}^{(L)}] \in \mathbb{R}^{L \cdot d}$, where d is the hidden dimension and we use the representation at the last token position from each layer.

We train a probe $\mathcal{P} : \mathbb{R}^{L \cdot d} \to \mathbb{R}^{C}$ to predict task labels. We consider two architectures:

Linear Probe (Logistic): $\mathcal{P}(\mathbf{h}) = W\mathbf{h} + b$, where $W \in \mathbb{R}^{C \times (L \cdot d)}$.

MLP Probe: $\mathcal{P}(\mathbf{h}) = W_2 \cdot \text{ReLU}(W_1\mathbf{h} + b_1) + b_2$, where $W_1 \in \mathbb{R}^{h \times (L \cdot d)}$, $W_2 \in \mathbb{R}^{C \times h}$, and h is a hidden dimension (512 in our experiments).

The probe is trained with cross-entropy loss: $\mathcal{L}_{\text{probe}} = -\sum_{i=1}^{N} \log \text{softmax}(\mathcal{P}(\mathbf{h}_i))_{y_i}$.

Unsupervised Variant (CCS). We also consider an unsupervised probe training approach using Contrast-Consistent Search (CCS) Burns et al. (2024). Unlike the supervised probes which require task labels, CCS discovers a truth direction in the teacher's representation space using only unlabeled contrast pairs. We extend this logic for MCQs by exploiting the fact that one answer must be correct.

For each question with C choices, we extract hidden states $\mathbf{h}_1, \ldots, \mathbf{h}_C$ and train a binary probe $\mathcal{P}_{\text{ccs}} : \mathbb{R}^{L \cdot d} \to \mathbb{R}$ that predicts whether each choice is correct. Let $p_c = \sigma(\mathcal{P}_{\text{ccs}}(\mathbf{h}_c))$ be the predicted probability for choice c. The CCS loss enforces two properties without using labels: (1) confidence predictions should be decisive, not 0.5, via $\mathcal{L}_{\text{conf}} = \frac{1}{C}\sum cp_c(1 - p_c)$; and (2) consistency,

Algorithm 1 PROBE-KD: Probe-Based Knowledge Distillation

Require: Teacher $\mathcal{T}$ with L layers, Student $\mathcal{S}$, Dataset $\mathcal{D}$
// Stage 1: Extract Hidden States
for $(x, y) \in \mathcal{D}$ **do**
  Extract $\mathbf{h} = [\mathbf{h}^{(1)}; \ldots; \mathbf{h}^{(L)}]$ from $\mathcal{T}(x)$
end for
// Stage 2: Train Probe
if supervised **then**
  Train probe $\mathcal{P}$ on $\{(\mathbf{h}_i, y_i)\}$ with cross-entropy
else {*(CCS: unsupervised)*}
  Train probe $\mathcal{P}$ on $\{\mathbf{h}_i\}$ with consistency + confidence loss
end if
// Stage 3: Distill to Student
for $(x, y) \in \mathcal{D}$ **do**
  Compute $p_{\text{probe}} = \text{softmax}(\mathcal{P}(\mathbf{h})/\tau)$
  Compute $p_{\mathcal{S}} = \text{softmax}(\mathcal{S}(x)/\tau)$
  $\mathcal{L} = \alpha \cdot \text{KL}(p_{\text{probe}} \| p_{\mathcal{S}}) + (1 - \alpha) \cdot \text{CE}(y, \mathcal{S}(x))$
  Update $\mathcal{S}$ with $\nabla\mathcal{L}$
end for
return Trained student $\mathcal{S}$

exactly one answer should be correct, via $\mathcal{L}\text{cons} = (\sum cp_c - 1)^2$. The probe is optimized with $\mathcal{L}\text{ccs} = \mathcal{L}\text{conf} + \mathcal{L}_{\text{cons}}$, requiring no labeled examples.

3.2 STAGE 2: PROBE-BASED DISTILLATION

Once the probe is trained, we use its soft predictions as supervision for the student. Given input x, we compute the probe's output distribution: $p_{\text{probe}} = \text{softmax}(\mathcal{P}(\mathbf{h})/\tau)$, where τ is the temperature parameter and $\mathbf{h}$ is the concatenated all-layer representation. The student is trained with a combination of distillation and task losses: $\mathcal{L} = \alpha \cdot \text{KL}(p_{\text{probe}} \| p_{\mathcal{S}}) + (1 - \alpha) \cdot \text{CE}(y, \mathcal{S}(x))$, where α balances the two objectives. Algorithm 1 summarizes the complete PROBE-KD procedure.

3.3 WHY PROBES PROVIDE BETTER SUPERVISION

Compared to standard distillation, the use of probes has several distinct advantages. Not only does training the probe on ground-truth data correct errors made by the existing model, the probe also serves as an intermediary that directly matches the output form of the student.

In contrast, when using the general LLM model as a teacher, the outputs are always overcomplete. The model must express its "belief" through specific tokens, competing with many alternatives most of which cannot be a valid answer . In contrast, the probe learns the optimal mapping from hidden states to labels using ground-truth supervision. The probe does not create "hidden knowledge"; it simply provides a better interface to the same representations the teacher uses internally.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We evaluate on four multiple-choice reasoning benchmarks. AQuA-RAT (Ling et al., 2017) contains algebraic word problems with 5 answer choices. ARC-Challenge and ARC-Easy (Clark et al., 2018) contain grade-school science questions with 4 choices; we use the official train/test splits (1.1K/1.2K and 2.3K/2.4K respectively). MMLU (Hendrycks & Dietterich, 2019) spans 57 subjects across STEM, humanities, and social sciences with 4 choices; since MMLU lacks a standard training set, we use the auxiliary train split provided by HuggingFace (2.2K examples). To study data efficiency, we train on $\{1\%, 10\%, 25\%, 50\%, 75\%, 100\%\}$ subsets of each training set.

For the primary teacher, we use Qwen2.5-7B-Instruct (Qwen et al., 2025) (7B parameters, 28 layers, hidden dim 3584), evaluated via 5-shot multiple-choice prompting. To test generalization across architectures, we also evaluate Phi-3-mini-4k-instruct (3.8B) and TinyLlama-1.1B-Chat-v1.0 (1.1B) as teachers. The primary student is DeBERTa-v3-base (He et al., 2023) (86M parameters) with a classification head over [CLS] representations. For the student architecture ablation, we compare four student architectures: DeBERTa-v3-base (86M), DeBERTa-v3-large (304M), ModernBERT-base (149M), and ModernBERT-large (395M).

Probes. We compare three probe architectures trained on concatenated hidden states from all L teacher layers. The **Logistic** probe is a linear projection $W \in \mathbb{R}^{C\times(L\cdot d)}$. The **MLP** probe is a two-layer network with hidden dimension 512. The **CCS** probe (Burns et al., 2024) is an unsupervised two-layer network that outputs a scalar score per choice, trained without labels. For MCQA, we adapt CCS by treating the C answer choices as the contrasting set: we extract hidden states for each "Question: {q} Answer: {choice}" prompt and train the probe to satisfy $\sum_{c=1}^{C} p_c \approx 1$ (exactly one answer is correct) while maximizing confidence via $\mathcal{L}_{\text{CCS}} = \text{Var}(p) + (\sum_c p_c - 1)^2$, where $p_c = \sigma(f(h_c))$ is the probe's prediction for choice c.

4.2 Methods Compared

We evaluate seven methods. As baselines, we include the untrained Student-Base and Teacher-MC (teacher 5-shot accuracy). For student training, we compare Supervised (gold labels only), Logit-KD (distillation from teacher output probabilities), and Feature-KD (student hidden states trained to match teacher hidden states via MSE loss (Jiao et al., 2020)). Our proposed methods are Probe-KD (Logistic) and Probe-KD (MLP), which distill from probe predictions rather than teacher outputs.

4.3 Training Details

Probe Training. Probes are trained for 20 epochs with AdamW (lr=1e-3, batch size 128, weight decay 0.01). We extract and concatenate hidden states from all layers (e.g., $28 \times 3584 = 100,352$ dimensions for Qwen2.5-7B).

Student Training. 3 epochs, AdamW (lr=2e-5), batch size 16, linear warmup (10% steps). For distillation: temperature $\tau = 2.0$, KD weight $\alpha = 0.7$.

Data Scaling. We train on {1%, 10%, 25%, 50%, 75%, 100%} of training data to study data efficiency.

Seeds. Main experiments use 5 seeds (42–46); ablations use 3 seeds.

5 Results

We evaluate Probe-KD against supervised learning, standard distillation methods, and additional baselines across four reasoning benchmarks. Our experiments address three questions: (1) Does Probe-KD improve over standard distillation? (2) When are the gains largest? (3) Does the approach generalize across teachers, students, and domains?

5.1 Main Results

Table 1 establishes that MLP probes trained on hidden-states consistently outperform the teacher's own outputs. The gap is non-trivial, up to +5.6% on AQuA-RAT, confirming that hidden states encode task-relevant information that the output layer fails to express. Linear probes underperform on some datasets, indicating that sufficient probe capacity might be necessary to decode this latent signal.

Table 2 presents our main comparison across training methods. Probe-KD (MLP) achieves the best student accuracy on AQuA (29.4%) and ARC-E (75.1%), and outperforms all distillation baselines on average: +2.2% over Logit-KD, +5.0% over Feature-KD, and +1.5% over Patient-KD.

Several comparisons merit discussion:

Table 1: Probe accuracy vs. teacher accuracy (%). MLP probes on hidden states outperform teacher 5-shot outputs, with the largest gap on AQuA-RAT.

Dataset	Teacher	Logistic	MLP	Δ
AQuA-RAT	44.7	50.6	50.3	+5.6
ARC-Easy	96.6	96.6	97.2	+0.6
ARC-Challenge	89.7	90.3	91.2	+1.5
MMLU	71.4	70.1	73.5	+2.1

Table 2: Test accuracy (%).

Method	AQuA	ARC-C	ARC-E	MMLU
Reference				
Teacher 5-shot	44.7	89.7	96.6	71.4
MLP Probe	50.3	91.2	97.2	73.5
Student: DeBERTa-v3-base (86M params)				
Supervised	29.3	52.3	73.5	31.8
+ Label Smoothing (Szegedy et al., 2015)	27.6	51.6	73.9	33.8
Distillation Methods				
Logit-KD (Hinton et al., 2015)	26.6	50.9	74.4	24.5
Feature-KD (Jiao et al., 2020)	27.4	38.6	69.7	29.6
Patient-KD (Sun et al., 2019)	27.6	**51.5**	74.6	25.7
Ours				
PROBE-KD (Logistic)	29.3	**51.5**	74.3	27.1
PROBE-KD (MLP)	**29.4**	50.1	**75.1**	**30.7**
PROBE-KD (CCS)	28.5	49.7	74.1	26.8

Feature-KD vs. PROBE-KD. Feature-based distillation (Jiao et al., 2020) trains the student to match teacher hidden states directly via MSE loss. This improves over Logit-KD but underperforms PROBE-KD. We attribute this gap to two factors: (1) Feature-KD requires architectural compatibility (we use a projection layer to match dimensions, following Jiao et al. (2020)), while PROBE-KD places no constraints on student architecture; (2) Feature-KD treats all information in hidden states as equally valuable, while the probe learns to extract task-relevant signal.

Label smoothing. Uniform label smoothing (Szegedy et al., 2015) improves over hard-label supervision but substantially underperforms PROBE-KD. Label smoothing provides uniform regularization; probe soft labels provide informed uncertainty based on the teacher's representations.

Teacher fine-tuning. A natural question arises: if probe training requires labels, why not simply fine-tune the teacher? Table 3 shows these approaches are complementary on AQuA-RAT. We fine-tune the Qwen2.5-7B teacher using LoRA (Hu et al., 2021) on the training set for 3 epochs, improving its 5-shot accuracy from 44.7% to 52.3%. However, standard Logit-KD from this fine-tuned teacher yields only 27.8% student accuracy—barely better than distilling from the base teacher (26.6%).

In contrast, training an MLP probe on the *base* teacher's hidden states achieves 29.4%, surpassing the fine-tuned teacher approach. Crucially, PROBE-KD can also be applied *on top of* fine-tuning: training a probe on the fine-tuned teacher's hidden states yields 28.7%. This suggests that probe-based distillation provides distinct benefits, extracting knowledge that fine-tuning alone cannot transfer through standard distillation.

Moreover, probe training is substantially more efficient: our 53M-parameter MLP probe trains in under 5 minutes on cached hidden states, compared to a few hours for LoRA fine-tuning a 7B model on a B200 GPU, a $>35\times$ speedup.

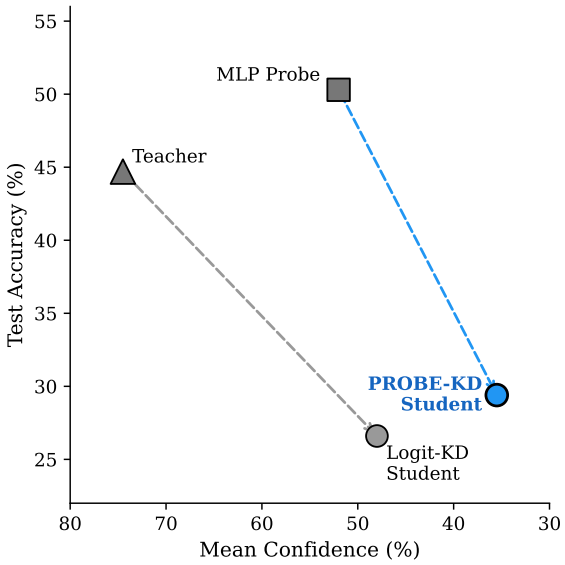

Figure 2: PROBE-KD Calibration analysis on AQuA-RAT. We plot test accuracy against mean prediction confidence, with the x-axis reversed so that lower confidence (better calibration) appears rightward. Arrows indicate the distillation path from source (Teacher or MLP Probe) to student.

Table 3: Fine-tuning vs. probe training on AQuA-RAT. The teacher is fine-tuned with LoRA for 3 epochs. PROBE-KD improves distillation from both base and fine-tuned teachers.

Method	Teacher Acc.	Student Acc.
Base Teacher (Qwen2.5-7B, no fine-tuning)		
Logit-KD	44.7	26.6
PROBE-KD (MLP)	44.7	**29.4**
PROBE-KD (CCS)†	44.7	28.5
Fine-tuned Teacher (+LoRA, 3 epochs)		
Logit-KD	52.3	27.8
PROBE-KD (MLP)	52.3	28.7

†CCS requires no labels for probe training.

5.2 WHEN DOES PROBE-KD HELP MOST?

We hypothesize that PROBE-KD provides the largest gains when teacher outputs are noisiest. We test this along data availability.

Data availability. Table 4 varies training set size on AQuA-RAT. The gap between PROBE-KD and Logit-KD is largest in the lowest data regime and decreases slightly at thousands of samples. The gap over Feature-KD follows a similar pattern. Cleaner supervision from probes is particularly valuable when training examples are scarce, as each example must carry more signal.

5.3 ABLATIONS

We verify that PROBE-KD generalizes across model choices.

Teacher architecture. Table 5 compares two additional teacher families on AQuA-RAT.

Student architecture. Table 6 varies student capacity. PROBE-KD benefits all configurations, with larger gains for larger students (+4.2% for 86M params vs. +5.6% for 395M over Logit-KD),

Table 4: Data scaling on AQuA-RAT. PROBE-KD consistently outperforms baseline distillation methods across data regimes.

	Training Samples		
Method	n=333	n=667	n=1000
Baseline Distillation			
Logit-KD	25.4	24.9	25.3
Feature-KD	24.7	24.9	24.2
Patient-KD	25.9	25.9	25.9
Ours			
PROBE-KD (Logistic)	26.7	**28.8**	27.1
PROBE-KD (MLP)	**27.8**	25.9	**28.3**
PROBE-KD (CCS)[†]	25.6	27.4	28.1

	Training Samples		
Method	n=1333	n=1667	n=2000
Baseline Distillation			
Logit-KD	26.5	21.5	26.6
Feature-KD	25.8	26.8	27.4
Patient-KD	25.4	26.5	27.6
Ours			
PROBE-KD (Logistic)	28.2	28.9	29.3
PROBE-KD (MLP)	**29.5**	**29.0**	**29.4**
PROBE-KD (CCS)[†]	27.5	27.9	28.5

[†] CCS requires no labels for probe training.

Table 5: Teacher ablation on AQuA-RAT. .

Teacher	Logit	Feature	PROBE-KD (MLP)
TinyLlama-1.1B	22.2	27.1	**27.5**
Phi-3-mini-3.8B	22.2	27.0	**27.8**

suggesting that higher-capacity students better exploit the cleaner supervision signal. The gap over Feature-KD is also consistent across architectures.

5.4 CALIBRATION

Beyond accuracy, we examine whether PROBE-KD produces better-calibrated students. We measure calibration via mean confidence: the average of $\max_c p(c \mid x)$ across test examples, where $p(c \mid x)$ is the model's predicted probability for choice c. A well-calibrated model has mean confidence approximately equal to its accuracy.

Figure 2 reveals the the teacher LLM is severely overconfident: 74.5% mean confidence despite only 44.7% accuracy, a calibration gap of nearly 30 percentage points. Standard Logit-KD transfers this miscalibration to the student, which predicts with 48% confidence while achieving only 26.6% accuracy.

In contrast, the MLP probe trained on hidden states is well-calibrated (52% confidence, 50.3% accuracy), and the PROBE-KD student inherits this property: its 35.5% mean confidence closely matches its 29.4% accuracy. This calibration improvement arises because probes produce soft labels that reflect genuine uncertainty in the hidden representations, rather than the teacher's overconfident token probabilities. Calibrated predictions are particularly valuable in downstream applications where prediction confidence informs decision-making, such as selective prediction.

Table 6: Student ablation on AQuA-RAT.

DeBERTa	Logit	Feat.	PROBE-KD
Base (86M)	27.8	27.6	**28.7**
Large (304M)	**21.6**	20.2	21.3
ModernBERT	Logit	Feat.	PROBE-KD
Base (149M)	27.4	26.8	**30.2**
Large (395M)	27.9	27.4	**28.6**

5.5 OVERVIEW

Notably, PROBE-KD is the only distillation method that consistently matches or exceeds fully supervised learning across all four benchmarks, a result that standard distillation methods fail to achieve. This suggests that probe-based supervision successfully integrates the benefits of soft-label training with task-specific optimization, yielding students that inherit the teacher's dark knowledge.

6 LIMITATIONS

Task scope. We evaluate on multiple-choice classification, where the probe maps hidden states to a small label set. Extending to generation tasks (e.g., open-ended QA, summarization) would require probes that decode to sequences, substantially increasing complexity. The multiple-choice setting is practically important—it underlies classification, reranking, and constrained generation, but does not cover all distillation scenarios.

Hidden state storage. While probe training is fast (minutes on cached hidden states), extracting and storing hidden states from all layers requires substantial memory: $O(N \cdot L \cdot d)$ for N examples, L layers, and hidden dimension d. For Qwen2.5-7B with 28 layers and $d = 3584$, this is $\sim$400KB per example. For large datasets, streaming or dimensionality reduction may be necessary.

Access requirements. PROBE-KD requires access to teacher hidden states, precluding black-box API-only teachers. This is a meaningful constraint as many powerful models (e,g. GPT-5.2) do not expose internal representations. However, the proliferation of open-weight models (Llama, Qwen, Mistral) makes this assumption practical.

Probe architecture. We compare linear and two-layer MLP probes; the optimal architecture may vary by task and teacher. More expressive probes could extract additional signal but risk overfitting. We leave systematic architecture search to future work.

7 CONCLUSION

We introduced PROBE-KD, a knowledge distillation framework that uses probe predictions on teacher hidden states as soft supervision. The key insight is that LLM hidden states encode richer task-relevant information than outputs reveal. By training a lightweight probe to decode this latent information, we obtain accurate soft labels that improve student training over standard logit distillation.

Probes trained on hidden states outperform teacher outputs (Table 1), and this improved supervision translates to better students (Table 2). PROBE-KD achieves state-of-the-art performance across difficult reasoning tasks and in low-data regimes (Table 4). PROBE-KD is consistent across teacher architectures, student architectures, and domains. This suggests distillation should target classifier outputs based on the latent space, rather than the output of the unembedding layer. For tasks where: teacher outputs are unreliable; multi-step reasoning; out-of-distribution inputs; tasks far from pretraining, PROBE-KD offers a principled way to extract cleaner supervision.

PROBE-KD requires no architectural changes to teacher or student, adds minimal overhead, and integrates with any soft-label distillation objective. We hope this work encourages further investigation into representation-based knowledge transfer, moving beyond the assumption that model outputs are the best available supervision signal.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, November 2018. URL http://arxiv.org/abs/1610.01644. arXiv:1610.01644.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying, October 2023. URL http://arxiv.org/abs/2304.13734. arXiv:2304.13734.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey, January 2019. URL http://arxiv.org/abs/1812.08951. arXiv:1812.08951.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, November 2025. URL http://arxiv.org/abs/2303.08112. arXiv:2303.08112.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, Philadelphia PA USA, August 2006. ACM. ISBN 9781595933393. doi: 10.1145/1150402.1150464. URL https://dl.acm.org/doi/10.1145/1150402.1150464.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, March 2024. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827.

Didier Chételat, Joseph Cotnareanu, Rylee Thompson, Yingxue Zhang, and Mark Coates. InnerThoughts: Disentangling Representations and Predictions in Large Language Models, January 2025. URL http://arxiv.org/abs/2501.17994. arXiv:2501.17994.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit Chain of Thought Reasoning via Knowledge Distillation, November 2023. URL http://arxiv.org/abs/2311.01460. arXiv:2311.01460.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models, November 2025. URL http://arxiv.org/abs/2306.08543. arXiv:2306.08543.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, March 2023. URL http://arxiv.org/abs/2111.09543. arXiv:2111.09543.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, March 2019. URL http://arxiv.org/abs/1903.12261. arXiv:1903.12261.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks, September 2019. URL http://arxiv.org/abs/1909.03368. arXiv:1909.03368.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large Language Models Are Reasoning Teachers, June 2023. URL http://arxiv.org/abs/2212.10071. arXiv:2212.10071.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes, July 2023. URL http://arxiv.org/abs/2305.02301. arXiv:2305.02301.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685. arXiv:2106.09685.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding, October 2020. URL http://arxiv.org/abs/1909.10351. arXiv:1909.10351.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, June 2024. URL http://arxiv.org/abs/2306.03341. arXiv:2306.03341.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems, October 2017. URL http://arxiv.org/abs/1705.04146. arXiv:1705.04146.

Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, Quzhe Huang, and Dongyan Zhao. Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10652–10658, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.729. URL https://aclanthology.org/2022.emnlp-main.729/.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching Small Language Models to Reason, June 2023. URL http://arxiv.org/abs/2212.08410. arXiv:2212.08410.

Saptarshi Mandal, Xiaojun Lin, and R. Srikant. A Theoretical Analysis of Soft-Label vs Hard-Label Training in Neural Networks, December 2024. URL http://arxiv.org/abs/2412.09579. arXiv:2412.09579.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When Does Label Smoothing Help?, June 2020. URL http://arxiv.org/abs/1906.02629. arXiv:1906.02629.

nostalgebraist. interpreting GPT: the logit lens — LessWrong. August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation, May 2019. URL http://arxiv.org/abs/1904.05068. arXiv:1904.05068.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language Models as Knowledge Bases?, September 2019. URL http://arxiv.org/abs/1909.01066. arXiv:1909.01066.

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, January 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets, March 2015. URL http://arxiv.org/abs/1412.6550. arXiv:1412.6550.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, March 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling Reasoning Capabilities into Smaller Language Models, May 2023. URL http://arxiv.org/abs/2212.00193. arXiv:2212.00193.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis, October 2023. URL http://arxiv.org/abs/2305.15054. arXiv:2305.15054.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient Knowledge Distillation for BERT Model Compression, August 2019. URL http://arxiv.org/abs/1908.09355. arXiv:1908.09355.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision, December 2015. URL http://arxiv.org/abs/1512.00567. arXiv:1512.00567.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline, August 2019. URL http://arxiv.org/abs/1905.05950. arXiv:1905.05950.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Representation Distillation, January 2022. URL http://arxiv.org/abs/1910.10699. arXiv:1910.10699.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, April 2020. URL http://arxiv.org/abs/2002.10957. arXiv:2002.10957.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL http://arxiv.org/abs/2201.11903. arXiv:2201.11903.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting Knowledge Distillation via Label Smoothing Regularization, March 2021. URL http://arxiv.org/abs/1909.11723. arXiv:1909.11723.

Sergey Zagoruyko and Nikos Komodakis. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer, February 2017. URL http://arxiv.org/abs/1612.03928. arXiv:1612.03928.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking Soft Labels for Knowledge Distillation: A Bias-Variance Tradeoff Perspective, February 2021. URL http://arxiv.org/abs/2102.00650. arXiv:2102.00650.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, March 2025. URL http://arxiv.org/abs/2310.01405. arXiv:2310.01405.