COMPILE2DECIDE: DISTILLING FRONTIER REASONING MODELS INTO FAST CALIBRATED DECISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large reasoning models have broad semantic competence but are computationally inefficient for repeated bounded decisions. We ask how much of their judgment can be compiled into a much smaller System One decision model that scores a dynamic candidate set directly, without generating explanations or reasoning traces, and can answer "none of the listed options" when the correct answer is absent. A reasoning-capable LLM teacher (Qwen3.8-27B, queried once per input with thinking disabled) scores the full label space plus none; its distribution is projected onto each run-time subset by sending the mass of unlisted labels to none and supplies soft supervision to a ModernBERT-base cross-encoder student, while ground-truth labels anchor the probabilities. The components are standard; the contribution is a decision contract and a pre-registered matched-control measurement. On CLINC150 (confirmatory; BoolQ is descriptive) the student is compared with cross-entropy on ground-truth labels, label smoothing and a permuted-teacher control, and evaluated on calibration, accuracy, latency, throughput, GPU time and energy. Over three seeds, distillation lowered dynamic negative log-likelihood by 0.082, 0.092 and 0.027 nats against the three controls (99% family intervals exclude zero); temperature scaling removed about three quarters of the gap to the ground-truth controls, leaving about 0.02 nats, and accuracy was 0.85 points below label smoothing, so the pre-registered claim that the teacher adds information beyond matched controls is not met. The teacher had no accuracy headroom: the served student was more accurate (0.913 against 0.857), about 4.5 times faster per single request (median 38 against 171 ms) and used a quarter to a third of the teacher's GPU-seconds per decision, breaking even after about 46k–72k requests, an advantage that a label-only fine-tuned student shares; a constrained decoder fine-tuned on labels did not dominate it. What compiled was mainly calibrated probability structure rather than accuracy, which suggests that teacher headroom over labels bounds how much expensive generative intelligence can be converted into reusable, low-latency probabilistic decision functions. The study uses one teacher and one student backbone, and its test sets had been used in an earlier study.

1 INTRODUCTION

Large reasoning models are increasingly used as judges for bounded, repeated decisions: routing a request, selecting a tool, flagging an event. Generating and parsing tokens is an expensive way to produce such decisions. We ask how much of a reasoning model's judgment can be compiled into a small System One decision model that scores a run-time candidate set directly and returns a calibrated distribution, so that the teacher is not called at serving time (we call this approach Compile2Decide). The candidate set is chosen at run time, the correct label can be missing from it, and the model must be able to say that none of the listed options applies. We use *frontier reasoning model* for the teacher class; the instance studied is Qwen3.8-27B, the more accurate of two pre-specified reasoning-capable open-weight candidates, queried with thinking disabled through identifier-token probabilities; our results therefore concern what its direct judgment, not its reasoning traces, transfers.

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.

A common way to build such a model is to query an LLM once over the full label set and distill its answers into a small encoder. Prior work leaves the value of the LLM's *soft* targets uncertain. Part of distillation's effect can come from the regularization that any softened target supplies (Yuan et al., 2020; Tang et al., 2020; Müller et al., 2019), and a class-uniform "teacher" frequently yields the strongest improvements (Mason-Williams et al., 2025). With fine-tuned text teachers, distillation and label smoothing behave differently (Sultan, 2023); with LLM teachers, soft labels gave nearly the same accuracy as hard LLM labels (Ramírez et al., 2024b). KL-based distillation is also sensitive to teacher calibration (Kim et al., 2025; Fan et al., 2024), and a zero-shot LLM judge is not calibrated. None of these studies combines an LLM teacher, matched smoothing and permutation controls, a none option and run-time candidate sets.

We ask whether, when candidate sets vary and gold may be absent, soft targets from an LLM judge improve a student's probability quality beyond matched controls. Each control removes one possible source of benefit. Label smoothing supplies softening without a teacher. A permuted teacher keeps the teacher's gold and none mass and its entropy but destroys its similarity structure over wrong labels, following the DKPP control of Furlanello et al. (2018). Temperature scaling (Guo et al., 2017) tests whether any NLL advantage is only a matter of scale. Only an advantage over all three indicates that the teacher's ranking of wrong candidates carries usable information. We use a probability endpoint because the statistical account of distillation places the teacher's value in its class-probability estimates (Menon et al., 2021).

The study makes the following contributions, all of which concern design and reporting rather than method:

- a decision contract for distilling an LLM judge into a student that sees candidate subsets with an explicit none option, in which the teacher target is the dustbin coarsening of Hinton et al. (2015, Sec. 5.2) and the teacher's argmax on a subset is a restriction-induced decision (Section 3);
- a pre-registered comparison of standard distillation against three matched controls with a dynamic NLL endpoint, one simultaneous family of raw and temperature-scaled comparisons and a per-view breakdown that separates known-absent from out-of-scope-absent views (Section 4);
- a measured quality and cost comparison against a constrained decoder fine-tuned on gold labels, with a break-even rule that applies only when the student stays within a fixed accuracy margin of the teacher (Section 4.4).

Over three seeds on CLINC150, distillation had lower dynamic NLL than all three controls, with about 0.02 nats surviving temperature scaling, but accuracy was slightly below label smoothing, so the information claim does not hold. The teacher had no headroom, and a label-only fine-tuned student shares the student's serving-cost advantage over it (Section 5).

2 Related work

Distillation, smoothing and controls. Standard knowledge distillation trains a student on the teacher's softened outputs (Hinton et al., 2015). It has been described as learned label smoothing (Yuan et al., 2020), split into uniform, domain and instance components (Tang et al., 2020), and analyzed as bias-variance or variance reduction (Menon et al., 2021; Zhou et al., 2021; Safaryan et al., 2023). Label smoothing improves calibration but erases inter-class similarity (Müller et al., 2019), and the distribution over non-target classes is a main source of logit distillation's effect (Zhao et al., 2022). Two controls bear directly on our design. The DKPP control of Born-Again Networks permutes all non-argmax teacher outputs and still produced gains in several settings (Furlanello et al., 2018); Random Control Distillation replaces the teacher with a class-uniform target, equivalent to label smoothing, and frequently gives the strongest gains (Mason-Williams et al., 2025). With fine-tuned text teachers, label smoothing raises posterior entropy while distillation almost always lowers it (Sultan, 2023). Temperature scaling is a strong post-hoc calibrator (Guo et al., 2017; Desai & Durrett, 2020), and teacher calibration conditions distillation gains (Kim et al., 2025; Fan et al., 2024); Appendix G covers further work.

LLM teachers and abstain classes. Hard LLM labels can reduce human annotation for text classifiers (Wang et al., 2021; Pangakis & Wolken, 2024; Kuzman & Ljubešić, 2025; Palo et al., 2024). For soft signals, Ramírez et al. (2024b) report nearly equal accuracy for soft and hard LLM labels, Wang et al. (2026b) report lower calibration error when LLM soft labels with verbalized confidence train an encoder, and Xia et al. (2025) distill LLM candidate-set annotations; none compares against matched label smoothing or includes a none option. The closest precedent for a distilled abstain class is Rawat et al. (2021), where a student distills a fixed class subset with an extra abstain output whose target is one-hot when gold lies outside the subset. Our projection instead puts the teacher's out-of-subset mass into none, as the dustbin of Hinton et al. (2015) and the residual-mass bucket of Anshumann et al. (2025) do; unlike there, this output is also the correct answer whenever gold is absent.

Out-of-scope intents and restricted candidate sets. CLINC150 pairs 150 intents with out-of-scope (OOS) queries (Larson et al., 2019). OOS detection uses an extra none class (Zhan et al., 2021), decision boundaries (Zhang et al., 2021) or confidence scores (Hendrycks & Gimpel, 2017); Lang et al. (2022) use soft labels only for pseudo-OOS samples with a fixed candidate set. Per-candidate cross-encoders over variable label sets follow entailment-based zero-shot classification (Yin et al., 2019). LLM OOS behavior depends on the size of the label space (Arora et al., 2024), aligned LLMs often choose an invalid option when none is correct (Góral et al., 2025), and on CLINC150 a zero-shot LLM and a fine-tuned encoder reach similar accuracy at very different cost (Rodrigues & Vas, 2026). Cost and amortization studies (Kang et al., 2023; Lambert & Luccioni, 2026) motivate our break-even count.

3 Decision contract and probability semantics

Contract. Let $\mathcal{U}$ be a task's full label set, for example the 150 in-scope intents of CLINC150; it does not contain the none option $\varnothing$. Each input s has at most one gold label $y \in \mathcal{U}$, and OOS inputs have $y = \varnothing$. At decision time the model receives a candidate set $\mathcal{A} \subseteq \mathcal{U}$ and outputs a distribution over $\mathcal{A} \cup \{\varnothing\}$ with target

$$y_{\mathcal{A}} = y \text{ if } y \in \mathcal{A}, \qquad y_{\mathcal{A}} = \varnothing \text{ otherwise.} \tag{1}$$

The canonical view presents $\mathcal{A} = \mathcal{U}$. The subset views are gold-present ($y \in \mathcal{A}$) and gold-absent ($y \notin \mathcal{A}$); a gold-absent view is known-absent when $y \in \mathcal{U}$ and OOS-absent when $y = \varnothing$. The none option has one fixed text, "none of the listed options," for the teacher, the student and the decoder in every view.

Teacher distribution and projection. The teacher sees the full label set once, in a fixed order with none last, and returns $q_{\mathcal{U}}$ over $\mathcal{U} \cup \{\varnothing\}$: the softmax of its next-token logits restricted to one identifier token per option. $q_{\mathcal{U}}$ is conditional on template, order and label set, and it is not a calibrated probability of being correct. For a presented subset $\mathcal{A}$ the teacher target is

$$r_{\mathcal{A}}(a) = q_{\mathcal{U}}(a) \quad (a \in \mathcal{A}), \qquad r_{\mathcal{A}}(\varnothing) = q_{\mathcal{U}}(\varnothing) + \textstyle\sum_{b \in \mathcal{U} \setminus \mathcal{A}} q_{\mathcal{U}}(b), \tag{2}$$

so $r_{\mathcal{A}}$ sums to one and $r_{\mathcal{U}} = q_{\mathcal{U}}$. Equation 2 is the dustbin coarsening of Hinton et al. (2015, Sec. 5.2), where all classes outside a specialist's subset share one output. There the dustbin is never a correct answer; here it is the gold label on every gold-absent view. Because the gold label of a gold-absent view lies in $\mathcal{U} \setminus \mathcal{A}$, $r_{\mathcal{A}}(\varnothing) \geq q_{\mathcal{U}}(y)$: teacher mass on the gold label always reaches the correct output.

Restriction-induced decisions. $r_{\mathcal{A}}$ post-processes one full-set prediction and is not the distribution the teacher would produce if shown only $\mathcal{A}$. We therefore call $\arg\max_a r_{\mathcal{A}}(a)$ a *restriction-induced decision*, never the teacher's judgment on the subset. Equating the two would require independence of irrelevant alternatives (IIA), which LLM judges need not satisfy (Arora et al., 2024); we do not test IIA. Renormalizing $q_{\mathcal{U}}$ over $\mathcal{A}$ would discard the mass on $\mathcal{U} \setminus \mathcal{A}$ and under-weight none, so that variant is not a comparison.

4 EXPERIMENTAL DESIGN

4.1 TASKS, SPLITS AND VIEWS

CLINC150 (Larson et al., 2019) is the primary task because its 150-way label set and genuine OOS queries produce both known-absent and OOS-absent views. BoolQ (Clark et al., 2019) is secondary, with $\mathcal{U} = \{\text{no}, \text{yes}\}$; its subset views hold a single option, so a gold-absent view is a verification item and BoolQ is not evidence about dynamic multi-way decisions. Test data are the official CLINC150 test set (5,498 examples, 1,000 OOS) and the official BoolQ validation set (3,270), both used in an earlier study (anonymized for review) (Section 7). The official training data are split once (seed 17), grouped by leakage key, into student_train (0.80; student and decoder training), model_tune (0.05; learning-rate selection and a reported diagnostic), prob_cal (0.025; temperature fitting) and two held-out roles that this study does not use (0.10 and 0.025). model_tune$_T$, a fixed draw of at most 512 model_tune examples, is the only tuning cohort with teacher calls. Training rows sharing a leakage key with any test row are removed, and the leakage group is the resampling unit for all intervals (Appendix A).

Views are materialized once (seed 17) and shared by all methods. Each training example has a canonical view and one fixed subset view. On CLINC150 ($k = 8$), half of the in-scope examples receive gold plus seven random intents and half eight random non-gold intents; OOS examples receive only gold-absent views. BoolQ uses $k = 1$. Evaluation examples have a canonical view, one gold-present view (in-scope only) and one gold-absent view. No candidate is truncated.

4.2 TEACHER, STUDENT AND OBJECTIVES

The teacher is one of two pre-specified open-weight candidates (Qwen3.8-27B (Qwen Team, 2026) and Qwen3-4B (Yang et al., 2025)), chosen before any student is trained by higher mean canonical accuracy on 150 training examples per task. Qwen3.8-27B was selected (mean pilot accuracy 0.883 against 0.767). It is used zero-shot with thinking disabled, one prompt template and one option order, and no chain-of-thought or verbalized confidence; scoring uses vLLM (Kwon et al., 2023), certified against a reference forward pass (argmax agreement on 8 of 8 rows per task). Prompt, identifiers and failure handling are in Appendix B; no teacher request failed.

The student is a ModernBERT-base cross-encoder (Warner et al., 2025) initialized from pretrained weights. Each candidate, and none as a last pseudo-candidate, is concatenated with the input and scored; a softmax over the $|\mathcal{A}| + 1$ scores gives $p_\theta(\cdot \mid s, \mathcal{A})$. With $\mathbf{e}_{y_\mathcal{A}}$ the one-hot target,

$$\mathcal{L}_{\text{GT-CE}} = \text{CE}\big(\mathbf{e}_{y_\mathcal{A}}, p_\theta\big), \qquad \mathcal{L}_{\text{GT-LS}} = \text{CE}\big((1-\epsilon)\mathbf{e}_{y_\mathcal{A}} + \tfrac{\epsilon}{|\mathcal{A}|+1}\mathbf{1},\, p_\theta\big),\ \epsilon = 0.1, \tag{3}$$

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{GT-CE}} + \lambda\, D_{\text{KL}}\big(r_\mathcal{A} \,\|\, p_\theta\big), \quad \mathcal{L}_{\text{KD-PERM}} = \mathcal{L}_{\text{GT-CE}} + \lambda\, D_{\text{KL}}\big(\tilde{r}_\mathcal{A} \,\|\, p_\theta\big), \tag{4}$$

with $\lambda = 0.5$ and $T = 1$ fixed. KD is standard distillation with the projected target of Equation 2. KD-PERM (CLINC150 only) replaces $q_\mathcal{U}$ per example by $\tilde{q}_\mathcal{U}$, which keeps $q_\mathcal{U}(y)$ and $q_\mathcal{U}(\varnothing)$ and permutes the remaining values among the non-gold labels of $\mathcal{U}$ (among all of $\mathcal{U}$ for OOS examples) with a permutation seeded by the example identifier and fixed across epochs, seeds and learning rates; $\tilde{r}_\mathcal{A}$ is its projection. This adapts the DKPP control of Furlanello et al. (2018) by also fixing the gold and none mass: KD-PERM keeps the teacher's confidence and entropy profile and removes its structure over wrong labels.

All student methods share one recipe (Appendix C): AdamW (Loshchilov & Hutter, 2019) with separate backbone and head learning rates, constant rate, gradient clipping, 32 views per step, exactly three epochs and the last checkpoint, with no early stopping. The base learning rates come from configurations selected for this backbone in a prior study. Every method receives the same two candidates, base and half, trained with seed 17, and keeps the one with lower dynamic NLL on model_tune; seeds 29 and 43 are also trained at that rate, so every method on both tasks has seeds 17, 29 and 43. A pre-registered diagnostic on model_tune (KD at least 2% lower relative dynamic NLL than both ground-truth controls, accuracy within one point) is reported only.

Decoder baseline. DEC is Qwen3-0.6B (Yang et al., 2025) with LoRA (Hu et al., 2021) (rank 16), fine-tuned on gold labels without teacher targets: on the same student_train views it is trained

with cross-entropy to emit the gold identifier token from the teacher prompt format restricted to the view's candidates plus none; its probabilities are the softmax over the view's identifier tokens. It uses a fixed recipe from the prior study and one learning rate, on CLINC150 only, whereas each student method chooses between two; DEC is outside the inferential family.

4.3 METRICS AND STATISTICS

Let V_i be the subset views of test example i (two for in-scope examples, one for OOS examples). The primary endpoint,

$$\mathrm{DNLL} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|V_i|}\sum_{v\in V_i} -\log p_\theta\big(y_{\mathcal{A}_v} \mid s_i, \mathcal{A}_v\big), \tag{5}$$

weights every original example once. $\mathrm{TS}(\cdot)$ denotes DNLL after dividing each model's logits by its own temperature, fitted on the canonical prob_cal views. Secondary metrics are canonical NLL, NLL per view type, canonical accuracy, none accuracy on gold-absent views, and canonical Brier score and 15-bin ECE (Appendix D).

The inferential family is on CLINC150 and has five comparisons: KD − GT-CE, KD − GT-LS and KD − KD-PERM on raw DNLL, and $\mathrm{TS}(\mathrm{KD}) - \mathrm{TS}(\text{GT-CE})$ and $\mathrm{TS}(\mathrm{KD}) - \mathrm{TS}(\text{GT-LS})$. Intervals are two-sided 99% percentile intervals (Bonferroni for five comparisons at $\alpha = 0.05$) from a grouped bootstrap with 10,000 resamples (seed 17), in which leakage groups are drawn with all their rows and views and the statistic is the difference averaged over seeds 17, 29 and 43. We call KD better than a control only if the upper bound is below zero and all three per-seed differences have the sign of the estimate. The intervals cover test sampling; training randomness enters only through the three-seed average and the sign rule. An interval containing zero is reported as "no difference detected, inconclusive," never as equality; "no effect larger than δ" requires the interval to lie within $\pm\delta$, with $\delta = 0.02\times$ the model_tune DNLL of GT-CE ($\delta = 0.0017$ nats), fixed before the test was read. Canonical accuracy is non-inferior to GT-CE or GT-LS only if the 99% lower bound of the difference and all three per-seed differences exceed -1 point. The statement that the teacher's soft targets add information beyond matched controls requires KD to be better than GT-CE, GT-LS and KD-PERM on raw DNLL and non-inferior in accuracy to GT-CE and GT-LS; if the temperature-scaled comparisons (or a descriptive variant with the temperature fitted on subset views) are not better, it carries the qualifier that the advantage can be replaced by post-hoc temperature scaling. BoolQ comparisons (KD − GT-CE, KD − GT-LS) use the same interval level and seeds and are descriptive.

4.4 COST AND BREAK-EVEN

Serving cost is measured on CLINC150 for the seed-17 KD student, the teacher and DEC at batch sizes 1 (B1) and 32 (B32) on one dedicated GPU, from raw request text to decision, as throughput, GPU-seconds, joules (NVML energy counters) and latency percentiles (protocol in Appendix F). The break-even count is

$$N^* = \frac{C_{\text{teacher annotation}} + C_{\text{student training}} + C_{\text{preparation}}}{c_{\text{teacher}} - c_{\text{student}}}, \tag{6}$$

where the numerator counts the teacher record on student_train, both seed-17 KD learning-rate runs and itemized preparation, and c is per-request cost, in GPU-seconds and joules and for B1 and B32 separately. A non-positive denominator means no break-even. N^* is read as a teacher-replacement break-even only if the 95% grouped bootstrap lower bound of student minus teacher canonical test accuracy exceeds -1 point; otherwise only the cost ratio is reported. N^* is a break-even count, not a realized saving, and it supports no cost claim for distillation itself unless the information claim holds; a counterfactual N^* for the GT-CE student, which needs no teacher annotation, is reported beside it. If DEC matches or exceeds the student on canonical accuracy, B32 throughput and B1 median latency, the paper makes no deployment-benefit claim for the student. A system's measured costs are claim-bearing only if its B1 decisions, its B32 decisions and its evaluated test decisions on the measurement cohort agree pairwise on at least 99% of requests.

Table 1: Teacher headroom on model_tune$_T$ (exploratory; rule net ≥ 0.01 and gain ≥ 0.03). Cells count examples by correctness of teacher (T) and reference (R). Prior-study references are contaminated; the BoolQ fresh row (seed-17 GT-CE, below the majority-class rate) carries no evidence about the teacher.

Task	Reference	n	Both	T only	R only	Neither	Gain	Harm	Net	Acc. T	Acc. R
CLINC150	prior study	512	440	0	71	1	0.000	0.139	-0.139	0.859	0.998
	GT-CE (fresh)	512	436	4	63	9	0.008	0.123	-0.115	0.859	0.975
BoolQ	prior study	399	331	28	26	14	0.070	0.065	0.005	0.900	0.895
	GT-CE (fresh)	399	204	155	13	27	0.388	0.033	0.356	0.900	0.544

4.5 Pre-registration and amendments

The protocol (splits, views, objectives, recipe, endpoints, decision rules and compute budgets; Appendix C) was frozen at the end of teacher selection, before the headroom diagnostic, student training and any test evaluation. The frozen protocol is identified by SHA-256 prefix `5d40f833`. This pre-registration was internal rather than deposited in an external registry. Amendments are listed in Appendix H; none changed the endpoints, the comparisons or the test data.

Amendments. A1, made after a probe of teacher accuracy on 256 training rows suggested that the original stop rule (net gain of at least one point and gain of at least three points over a cheap prior-study reference on model_tune$_T$) would halt the study, turned that rule into a reported diagnostic; the probe rows came from a provisional split that proved byte-identical to the frozen split, with all probe rows in student_train, and no claim of usable headroom is made unless the rule passes. Later amendments added KD-Perm, the temperature-scaled comparisons, the five-comparison family with BoolQ outside it, the equivalence margin, the GT-CE counterfactual for N^*, descriptive teacher diagnostics, an engine certification, the agreement rule of Section 4.4, a fail-closed handling rule for abnormal executions and, finally, cost measurement on a second GPU of the same model when exclusive access to the pre-registered device was unavailable (Appendix H).

5 Results

Students are means over seeds 17, 29, 43 at each method's selected learning rate (Appendix E). KD passed the model_tune diagnostic.

Teacher headroom (reported diagnostic). On CLINC150 the teacher has no headroom (Table 1): against the prior-study reference its gain was 0.000 (net -0.139). That reference had been trained on 511 of the 512 rows, so the comparison is contaminated; the pre-registration expected the bias to run against the teacher, but the diagnostics treat its direction as not established, and so do we. Against the seed-17 GT-CE student of this study, which never trained on them, the teacher also had almost no gain (0.008) and lower accuracy. The reported rule fails. The teacher labelled 73 of 80 OOS training rows as none (0.913, 95% CI $[0.828, 0.964]$); its calibration is in Appendix D.

Distillation against matched controls. Tables 2 and 3 give the test metrics and the comparisons. On CLINC150 the dynamic NLL of KD was 0.082 nats, against 0.164 for GT-CE, 0.174 for GT-LS and 0.109 for KD-Perm. All three raw differences have 99% intervals below zero with all three seeds of the same sign, so KD is better than each control. The intervals cover test sampling only, and the size of the GT-CE gap depends on the seed: per-seed KD $-$ GT-CE differences ranged from -0.114 to -0.021, because the seed-43 GT-CE run was far better (dynamic NLL 0.189, 0.203 and 0.100 for seeds 17, 29, 43). The difference to KD-Perm is attributable to teacher non-gold structure, since KD-Perm changes only which wrong labels receive which teacher values. After temperature scaling fitted on canonical prob_cal views, KD remained better than GT-CE and GT-LS by 0.021 and 0.025 nats (subset-fitted temperatures give -0.018 and -0.019). The advantage is not replaceable by temperature scaling, but scaling removes about 74% and 73% of the raw gap. Descriptively (not pre-registered), KD-Perm had lower raw dynamic NLL than both ground-truth controls but not after scaling (0.109 against 0.102 and 0.106), while $\mathrm{TS}(\mathrm{KD}) - \mathrm{TS}(\text{KD-Perm}) =$

Table 2: Test metrics at the selected learning rates (LR); students are means over seeds 17, 29, 43. DEC and the teacher (projected $q_{\mathcal{U}}$, no temperature) are single descriptive systems. Rounded to 0.001.

Task	Method	LR	DNLL	TS-DNLL	Canon. NLL	Canon. acc.	Brier	ECE
CLINC150	GT-CE	half	0.164	0.102	0.635	0.910	0.156	0.066
	GT-LS	base	0.174	0.106	0.579	0.924	0.194	0.175
	KD	base	0.082	0.081	0.305	0.915	0.133	0.020
	KD-PERM	half	0.109	0.109	0.366	0.915	0.137	0.028
	DEC	–	0.144	0.120	0.810	0.886	0.202	0.076
	Teacher	–	0.251	–	0.551	0.857	0.213	0.041
BoolQ	GT-CE	base	0.951	0.658	1.522	0.595	0.627	0.203
	GT-LS	half	0.692	0.624	0.961	0.660	0.499	0.091
	KD	half	0.679	0.633	0.911	0.691	0.461	0.100
	Teacher	–	0.298	–	0.367	0.899	0.163	0.025

Table 3: CLINC150 differences of KD minus control (mean over three seeds) with 99% grouped bootstrap intervals, which cover test sampling only; third line: range of the per-seed differences. b: family member, "better"; ni: non-inferior; nn: non-inferiority not shown (interval below zero). NLL rounded to 0.001, accuracy in points to 0.01. BoolQ: Appendix E.

Task	Comparison	DNLL	TS-DNLL	Canon. acc. (points)
CLINC150	KD − GT-CE	-0.082^{b}	-0.021^{b}	$+0.55^{\text{ni}}$
		$[-0.102, -0.064]$	$[-0.028, -0.014]$	$[-0.05, +1.16]$
		seeds -0.114 to -0.021	seeds -0.045 to -0.003	seeds $+0.42$ to $+0.64$
	KD − GT-LS	-0.092^{b}	-0.025^{b}	-0.85^{nn}
		$[-0.099, -0.084]$	$[-0.033, -0.017]$	$[-1.44, -0.26]$
		seeds -0.099 to -0.084	seeds -0.027 to -0.020	seeds -0.96 to -0.73
	KD − KD-PERM	-0.027^{b}	–	–
		$[-0.035, -0.020]$		
		seeds -0.035 to -0.017		

-0.028 matches the raw KD − KD-PERM gap: the part attributable to teacher non-gold structure survives calibration.

Canonical accuracy of KD was non-inferior to GT-CE ($+0.55$ points, $[-0.05, +1.16]$). Against GT-LS it was -0.85 points $[-1.44, -0.26]$: KD's accuracy is lower than GT-LS's (the 99% interval excludes zero and all three seeds are negative; a directional reading, not a pre-registered test), the deficit is not shown to exceed the 1-point margin, and non-inferiority is not shown. Because the information claim requires non-inferiority against both ground-truth controls, *it does not hold*. The family supports lower dynamic NLL than each control, with and without temperature scaling, for KD with dustbin projection ($\lambda = 0.5$, $T = 1$) from this teacher. On gold-present training views the projected teacher target put a mean mass of 0.158 on none, where none is wrong. On BoolQ (descriptive; Appendix E) the large KD − GT-CE gaps compare against an undertrained control: GT-CE accuracy (0.595) is below the test majority-class rate (0.622, 2,033 of 3,270 "yes"). Against GT-LS (0.660), KD's NLL advantage is not resolved (-0.013 $[-0.028, +0.002]$) and disappears after temperature scaling ($+0.009$ $[+0.000, +0.018]$), while accuracy is 3.0 points higher, with per-seed signs that disagree.

Gold-absent and out-of-scope views. KD had lower NLL than both ground-truth controls on every view type (Table 12 in Appendix E). Its advantage over KD-PERM lay on canonical (0.305 against 0.366) and gold-present views (0.142 against 0.212), while KD-PERM was as good on gold-absent views (0.034 and 0.049 against 0.036 and 0.050), whose targets are mostly projected none mass that the permutation leaves nearly unchanged. None accuracy on gold-absent views lay between 0.974 and 0.997 for every system. The teacher's restriction-induced decisions reach an NLL of 0.014 on known-absent views but 0.587 on gold-present views.

Table 4: Serving cost (512 CLINC150 test requests, five blocks, one GPU) and break-even counts N^* in GPU-s, rounded (exact: Appendix F). Agreement: minimum pairwise decision agreement. $N^*_{\text{GT-CE}}$ uses the KD student's cost.

		Decisions/s		GPU-s/decision		J/decision		B1 latency (ms)
System	Agreement	B1	B32	B1	B32	B1	B32	p50 / p95
KD student	1.000	25.6	25.1	0.039	0.040	8.5	10.6	38 / 49
DEC	0.990	19.5	63.8	0.055	0.016	10.3	5.8	55 / 82
Teacher	0.992	5.8	7.9	0.174	0.127	90.2	70.7	171 / 206
N^* (KD student)				46k	72k	–	–	
$N^*_{\text{GT-CE}}$ (counterfactual)				29k	45k	–	–	

Serving cost and the decoder. Costs are measured for the seed-17 KD student (accuracy 0.913). All three systems passed the agreement rule (minimum 1.000, 0.990 and 0.992 for student, DEC and teacher), so their costs are claim-bearing (Table 4). The student used 0.039 GPU-s per decision at B1 and 0.040 at B32, against 0.174 and 0.127 for the teacher; it does not gain from batching in this serving implementation, which forwards at most 8 requests (each with 151 candidate sequences) per model call. DEC, fine-tuned on gold labels, is cheaper at B32 (0.0157 against 0.0402 GPU-s per decision; 63.8 against 25.1 decisions/s) but 2.8 points less accurate (0.886 against 0.913) and slower at B1 (median 55 against 38 ms, but block medians ranged 35–71 ms for DEC and 30–41 ms for the student). It therefore does not dominate the student under the pre-registered rule, which requires all three conditions; non-dominance rests mainly on accuracy, since the B1 latency difference is unstable across blocks. DEC used one fixed learning rate whereas each student method chose between two, so this comparison does not isolate architecture. The student's deployment-benefit claim stands at both batch sizes: it is more accurate than the teacher at a quarter to a third of its GPU cost per decision.

Break-even. The accuracy condition for reading N^* as a teacher-replacement break-even is met (measured KD student minus teacher canonical test accuracy $+5.7$ points, 95% grouped bootstrap interval $[+4.7, +6.7]$). The one-time cost of 6,212 GPU-s (teacher annotation of student_train 2,134; all attempts of the two seed-17 KD learning-rate candidate runs 4,077, as pre-registered; preparation 0) balances the per-decision saving after about 46k requests at B1 and 72k at B32. Teacher B1 cost ranged 0.159–0.189 GPU-s per decision across blocks, which moves the B1 N^* between about 41k and 52k. The counterfactual GT-CE student, which needs no annotation (3,933 GPU-s) and is assumed to serve at the KD student's cost, breaks even after about 29k and 45k requests. Distillation thus costs 2,279 GPU-s more than training GT-CE for a dynamic NLL reduction of 0.082 nats; because the information claim does not hold, we make no amortization claim for distillation itself. Joule-based N^* is unavailable (training energy was recorded for the training loop only); all N^* are break-even counts, not realized savings.

6 DISCUSSION

What compiles. What compiled was probability structure rather than accuracy. KD beat smoothing without a teacher, the scrambled-teacher control and, after temperature scaling, both ground-truth controls. The KD-PERM gap survives scaling ($\text{TS(KD)} - \text{TS(KD-PERM)} = -0.028$), so where the teacher puts its non-gold mass matters, not only how much. Most of the raw advantage over the ground-truth controls, however, is scale that one temperature recovers, and KD was about 0.9 points less accurate than GT-LS, so the information claim fails. Soft supervision anchored by ground truth is best read here as a calibration source with a small residual of teacher-specific structure.

Temperature scaling. GT-CE was overconfident (mean fitted temperature 2.43), and scaling lowered its dynamic NLL from 0.164 to 0.102; KD's temperature was 1.02. Scaling thus removed about 74% and 73% of the raw gaps to GT-CE and GT-LS, leaving 0.021 and 0.025 nats. After scaling, GT-CE was slightly lower than KD on gold-present views (0.133 against 0.141), and KD's residual

advantage lay on gold-absent views (Appendix E). Temperatures were fitted on canonical prob_cal views, about 0.7% OOS against 18% in the test set, which may penalize the scaled controls on OOS-absent test views.

Teacher headroom bounds compilation. The capability that can be compiled is bounded by what the teacher knows that the labels do not. With thinking disabled, the teacher had no headroom on CLINC150, yet the calibration gain still arose, in line with reports that weaker teachers can improve students (Yuan et al., 2020; Stanton et al., 2021). On BoolQ the teacher was far more accurate than every student (0.899 against 0.595–0.691), but BoolQ views hold one option and the task is descriptive here. The decisive test needs tasks where the teacher, possibly reasoning before it scores, exceeds label-trained students, or where labels are scarce.

Where the deployment gain comes from. Replacing a generative judge with a direct-scoring student gave the large effect; soft rather than hard supervision gave the small one (2,279 extra GPU-s for 0.082 nats of raw dynamic NLL). The GT-CE student is also more accurate than the teacher and breaks even sooner, so the costs favor a fine-tuned student over the teacher, not distillation over fine-tuning; and because DEC is cheaper at B32, we make no claim that direct scoring dominates constrained generation. All statements concern KD with dustbin projection ($\lambda = 0.5$, $T = 1$) from Qwen3.8-27B, a ModernBERT-base student and CLINC150 with eight-candidate subsets; since distillation gains depend on teacher calibration (Kim et al., 2025; Fan et al., 2024), the effect may differ for other teachers.

Future work. First, domains whose ground truth arrives from future outcomes (financial microstructure, blockchain, traffic, maritime operations, scientific event detection and agent routing), with transfer within and across them. Second, teachers that reason before scoring, or that propose the candidate space and the decision decomposition, which could supply the missing headroom. Third, a confidence-gated cascade that calls the teacher only when the student is uncertain (Chen et al., 2023; Ong et al., 2025), evaluated by selective risk (Geifman & El-Yaniv, 2017), escalation fraction and an energy break-even; calibrated, none-aware probabilities are what such a gate thresholds.

7 LIMITATIONS

The CLINC150 test set and the BoolQ validation set were used in an earlier study (anonymized for review), which shaped some choices here; the protocol was fixed before any test results were inspected, but the test sets are not unseen. The headroom reference was trained on 511 of the 512 model_tune$_T$ rows, so that comparison is contaminated, with the direction of bias not established. The teacher, the decoder's base model and the student backbone may have seen CLINC150 or BoolQ in pretraining. Every method uses three seeds; the intervals cover test sampling, and seed variation enters only through the average and the sign rule. One teacher with one template and option order, one student backbone, one decoder and one dynamic task at one subset size are used, and we do not test IIA. Learning-rate selection and the diagnostic share model_tune, and DEC has no learning-rate choice. A1 removed a stop rule after a probe of teacher accuracy. Costs are measured on one GPU type (not the pre-registered one; a disclosed amendment), one serving implementation per system and canonical requests only. More seeds, tuned λ, T and DEC remain open; closed frontier models and reasoning-enabled teacher decoding were not tested.

8 CONCLUSION

Compile2Decide asked how much of a reasoning model's judgment a small System One decision model can absorb when it scores run-time candidate sets with a none option. With Qwen3.8-27B and CLINC150, soft supervision anchored by ground truth gave lower dynamic NLL than all three matched controls over three seeds, with about 0.02 nats surviving temperature scaling, but no accuracy benefit over label smoothing, from a teacher with no headroom. A fine-tuned student, distilled or not, replaced the teacher at a quarter to a third of its GPU cost per decision. What compiled was calibrated probability structure; how much judgment can be compiled when the teacher knows more than the labels remains an open question.

Reproducibility statement

Section 3 defines the contract, projection and targets, and Section 4 gives the splits, views, objectives, metrics, statistics, cost protocol and decision rules. The appendices give split construction and support (Appendix A), the teacher prompt, identifier scheme and selection (Appendix B), the training recipe, model revisions and compute (Appendix C), metric, permutation, bootstrap and teacher diagnostics (Appendix D), per-seed results and learning-rate selection (Appendix E), the measurement protocol (Appendix F), extended related work (Appendix G) and the amendment log (Appendix H). The protocol was frozen by recording its hash (SHA-256 prefix `5d40f833`). Code and instructions are provided in the supplementary material. Additional reproduction artifacts, including the prior-system dependency, the split and view identifiers and the teacher records, will be released with the final version.

AI use statement

Generative AI tools were used to assist with retrieving and discovering related literature, code implementation, analysis, writing and polishing. The research ideas originated with the authors. The authors reviewed and verified AI-assisted content and take full responsibility for the paper's accuracy, originality and conclusions. Qwen3.8-27B, ModernBERT-base and Qwen3-0.6B are objects of study, not writing aids.

References

Ilias Aarab. BTZSC: A Benchmark for Zero-Shot Text Classification Across Cross-Encoders, Embedding Models, Rerankers and LLMs. In *ICLR 2026*, 2026.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse Logit Sampling: Accelerating Knowledge Distillation in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), Volume 1: Long Papers*, 2025. doi: 10.18653/v1/2025.acl-long.885.

Gaurav Arora, Shreya Jain, and Srujana Merugu. Intent Detection in the Age of LLMs. In *Proceedings of EMNLP 2024: Industry Track*, pp. 1559–1570, 2024. doi: 10.18653/v1/2024.emnlp-industry.114.

Ryan Brown and Chris Russell. Task-Specific Knowledge Distillation via Intermediate Probes. *arXiv preprint arXiv:2603.12270*, 2026.

Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. Intent Recognition and Out-of-Scope Detection using LLMs in Multi-party Conversations. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2025)*, pp. 504–512, 2025. doi: 10.18653/v1/2025.sigdial-1.41.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, and Ngai-Man Cheung. Revisiting Label Smoothing and Knowledge Distillation Compatibility: What was Missing? In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022), PMLR 162:2890-2916*, 2022.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv preprint arXiv:2305.05176*, 2023.

Peng Chen and Bang Wang. On Collaborating Small and Large Models For Few-shot Intent Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13937–13953, 2025. doi: 10.18653/v1/2025.findings-emnlp.749.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300.

Shrey Desai and Greg Durrett. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, 2020. doi: 10.18653/v1/2020.emnlp-main.21.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. Revisit the Essence of Distilling Knowledge through Calibration. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235:12882-12894*, 2024.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018), PMLR 80:1607-1616*, 2018.

Parishruthi Ganesh, Gerry Dozier, and Cheryl Seals. Selecting Open-Weight Language Models for Zero-Shot Intent Classification: A Systematic Evaluation of 41 Models. *arXiv preprint arXiv:2607.27421*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Alberto Andres Valdes Gonzalez. Cost-Aware Model Selection for Text Classification: Multi-Objective Trade-offs Between Fine-Tuned Encoders and LLM Prompting in Production. *arXiv preprint arXiv:2602.06370*, 2026.

Gracjan Góral, Emilia Wiśnios, Piotr Sankowski, and Paweł Budzianowski. Wait, that's not an option: LLMs Robustness with Incorrect Multiple-Choice Options. In *Proceedings of the 63rd Annual Meeting of the ACL (ACL 2025), Volume 1: Long Papers*, 2025. doi: 10.18653/v1/2025.acl-long.75.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017), PMLR 70:1321-1330*, 2017.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR 2017*, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. doi: 10.18653/v1/2023.findings-acl.507.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Junmo Kang, Wei Xu, and Alan Ritter. Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers*, 2023. doi: 10.18653/v1/2023.acl-long.622.

Suyoung Kim, Seonguk Park, Junhoo Lee, and Nojun Kwak. The Role of Teacher Calibration in Knowledge Distillation. *IEEE Access*, 2025. doi: 10.1109/ACCESS.2025.3585106.

Taja Kuzman and Nikola Ljubešić. LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, 13:35621–35633, 2025. doi: 10.1109/ACCESS.2025.3544814.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Katherine Lambert and Sasha Luccioni. Towards Resource-Efficient LLMs: End-to-End Energy Accounting of Distillation Pipelines. *arXiv preprint arXiv:2605.13981*, 2026.

Hao Lang, Yinhe Zheng, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. Estimating Soft Labels for Out-of-Domain Intent Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 261–276, 2022. doi: 10.18653/v1/2022.emnlp-main.18.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of EMNLP-IJCNLP 2019*, 2019. doi: 10.18653/v1/D19-1131.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR 2019*, 2019.

Viviana Luccioli, Rithika Iyengar, Ryan Panley, Flora Haberkorn, Xiaoyu Ge, Leland Crane, Nitish Sinha, and Seung Jung Lee. LLM on a Budget: Active Knowledge Distillation for Efficient Classification of Large Text Corpora. In *2025 IEEE International Conference on Big Data (BigData)*, 2025. doi: 10.1109/bigdata66926.2025.11401628.

Israel Mason-Williams, Gabryel Mason-Williams, and Helen Yannakoudakis. A Functional Perspective on Knowledge Distillation in Neural Networks. *arXiv preprint arXiv:2510.12615*, 2025.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), PMLR 139:7632-7642*, 2021.

Ishan Mishra, Sethu Vamsi Krishna, and Deepak Mishra. Distilling Calibrated Student from an Uncalibrated Teacher. *arXiv preprint arXiv:2302.11472*, 2023.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online Cascade Learning for Efficient Inference over Streams. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235:38071-38090*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to Route LLMs with Preference Data. In *ICLR 2025*, 2025.

Flavio Di Palo, Prateek Singhi, and Bilal Fadlallah. Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 2024. doi: 10.18653/v1/2024.emnlp-main.215.

Nicholas Pangakis and Samuel Wolken. Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pp. 113–131, 2024. doi: 10.18653/v1/2024.nlpcss-1.9.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, Fan Zhou, Hao Luo, Haofeng Huang, Haoran Lian, Haoyan Huang, Hongqing Chen, Jianwei Zhang, Jing Xu, Junjie Wang, Langshi Chen, Liangyu Wang, Linlang Jiang, Man Yuan, Minmin Sun, Peng Jin, Siqi Zhang, Siyu Wang, Xingzhang Ren, Yakai Wang, Yi Zhang, Yiming Dong, Yizhong Cao, Yubo Ma, Yunfei Mao, Bo Zheng, and Dayiheng Liu. On the Design of Qwen3.8-Next Architecture: Evaluation, Efficiency, and Training Stability. *arXiv preprint arXiv:2608.30320*, 2026.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork. `https://qwen.ai/blog?id=qwen3.8`, 2026. Qwen Team blog; cited by the Qwen/Qwen3.8-27B model card. Model card: `https://huggingface.co/Qwen/Qwen3.8-27B`, revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0.

Guillem Ramírez, Alexandra Birch, and Ivan Titov. Optimising Calls to Large Language Models with Uncertainty-Based Two-Tier Selection. In *First Conference on Language Modeling (COLM 2024)*, 2024a.

Guillem Ramírez, Matthias Lindemann, Alexandra Birch, and Ivan Titov. Cache & Distil: Optimising API Calls to Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024b. doi: 10.18653/v1/2024.findings-acl.704.

Ankit Singh Rawat, Manzil Zaheer, Aditya Krishna Menon, Amr Ahmed, and Sanjiv Kumar. When in Doubt, Summon the Titans: Efficient Inference with Large Models. *arXiv preprint arXiv:2110.10305*, 2021.

Carson Rodrigues and Oysturn Vas. When Do LLMs Replace Fine-Tuned NLU? A Decision Framework for Intent Detection in Production Conversational Systems. *arXiv preprint arXiv:2608.20371*, 2026.

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge Distillation Performs Partial Variance Reduction. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. doi: 10.52202/075280-3288.

Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR 2021*, 2021.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew G. Wilson. Does Knowledge Distillation Really Work? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Md Arafat Sultan. Knowledge Distillation $\approx$ Label Smoothing: Fact or Fallacy? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pp. 4469–4477, 2023. doi: 10.18653/v1/2023.emnlp-main.271.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. Understanding and Improving Knowledge Distillation. *arXiv preprint arXiv:2002.03532*, 2020.

Guanghui Wang, Kaiwen Lv Kacuila, Zhiyong Yang, Zitai Wang, Jin-Wen Wu, Longtao Huang, Qianqian Xu, and Qingming Huang. The Bridge-Garden Dilemma in LLM Distillation: Why Mixing Hard and Soft Labels Works. *arXiv preprint arXiv:2605.26246*, 2026a.

Jinyuan Wang, Ningyuan Deng, and Yi Yang. Assessing and Mitigating Miscalibration in LLM-Based Social Science Measurement. *arXiv preprint arXiv:2605.11954*, 2026b.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-emnlp.354.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the ACL (ACL 2025), Volume 1: Long Papers*, 2025. doi: 10.18653/v1/2025.acl-long.127.

Qianhui Wu, Huiqiang Jiang, Haonan Yin, Börje F. Karlsson, and Chin-Yew Lin. Multi-Level Knowledge Distillation for Out-of-Distribution Detection in Text. In *Proceedings of the 61st Annual Meeting of the ACL (ACL 2023), Volume 1: Long Papers*, 2023. doi: 10.18653/v1/2023.acl-long.403.

Mingxuan Xia, Haobo Wang, Yixuan Li, Zewei Yu, Jindong Wang, Junbo Zhao, and Runze Wu. Prompt Candidates, then Distill: A Teacher-Student Framework for LLM-driven Data Annotation. In *Proceedings of the 63rd Annual Meeting of the ACL (ACL 2025), Volume 1: Long Papers*, 2025. doi: 10.18653/v1/2025.acl-long.139.

Albert Xu, Xiang Ren, and Robin Jia. Contrastive Novelty-Augmented Learning: Anticipating Outliers with Large Language Models. In *ACL 2023*, 2023.

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. *arXiv preprint arXiv:2505.09388*, 2025.

Lehan Yang and Jincen Song. Rethinking the Knowledge Distillation From the Perspective of Model Calibration. *arXiv preprint arXiv:2111.01684*, 2021.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of EMNLP-IJCNLP 2019*, 2019. doi: 10.18653/v1/D19-1404.

Amirhossein Yousefiramandi and Ciaran Cooney. Fine-Tuning Causal LLMs for Text Classification: Embedding-Based vs. Instruction-Based Approaches. *arXiv preprint arXiv:2512.12677*, 2025.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting Knowledge Distillation via Label Smoothing Regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/cvpr42600.2020.00396.

Álvaro Zaera, Diana Nicoleta Popa, Ivan Sekulic, and Paolo Rosso. Efficient Out-of-Scope Detection in Dialogue Systems via Uncertainty-Driven LLM Routing. In *Proceedings of the 63rd Annual Meeting of the ACL (ACL 2025), Volume 6: Industry Track*, pp. 328–335, 2025. doi: 10.18653/v1/2025.acl-industry.25.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y.S. Lam. Out-of-Scope Intent Detection with Self-Supervision and Discriminative Training. In *Proceedings of ACL-IJCNLP 2021 (Volume 1: Long Papers)*, 2021. doi: 10.18653/v1/2021.acl-long.273.

Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep Open Intent Classification with Adaptive Decision Boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence 35(16)*, 2021. doi: 10.1609/aaai.v35i16.17690.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip Yu. Are Pre-trained Transformers Robust in Intent Classification? A Missing Ingredient in Evaluation of Out-of-Scope Intent Detection. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pp. 12–20, 2022. doi: 10.18653/v1/2022.nlp4convai-1.2.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled Knowledge Distillation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi: 10.1109/cvpr52688.2022.01165.

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. In *ICLR 2021*, 2021.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. In *ICLR 2024*, 2024.

A SPLIT CONSTRUCTION AND SUPPORT

Roles. The official training data of each task are split once (seed 17) into five roles.

Role	Share	Use in this study
student_train	0.80	Student and decoder training; teacher records for KD and KD-PERM
router_train	0.10	Held out; unused
model_tune	0.05	Learning-rate selection and the reported model_tune diagnostic (students only)
model_tune$_T$	$\leq$512 rows	Teacher headroom diagnostic; the only tuning cohort with teacher calls
prob_cal	0.025	Temperature fitting for the temperature-scaled comparisons
policy_cal	0.025	Held out; unused
test	official	One evaluation after all models are frozen

Construction. Leakage groups are the normalized utterance (CLINC150) and the transitive closure of shared normalized questions and passages (BoolQ). Groups are allocated per label stratum, defined by the group's majority target, in a seeded group order; each group goes to the role furthest below its target share. CLINC150 OOS rows keep the official-training proportion without resampling. Training rows in a group that shares a key with any test row are removed (test priority), duplicate inputs within a group keep the first identifier, and inputs with conflicting targets are quarantined in all roles. model_tune$_T$ is a draw of at most 512 model_tune identifiers with a generator seeded from the string `model_tune_T`, the task name and seed 17, applied to the sorted identifier list. Every join across split, view, teacher and prediction files fails on a missing or extra identifier; teacher failures remain as rows. The official CLINC150 validation set is unused in every role.

Probe rows (amendment A1). The 256 pre-freeze teacher-probe rows were drawn from the student_train role of a provisional split generated with the same split procedure and seed as the frozen split. The frozen CLINC150 split was byte-identical to this provisional split, and all 256 probe rows fall in student_train.

Table 5: Retained examples per role and population. Exclusions: CLINC150, 4 inputs with conflicting duplicate targets; BoolQ, 1,439 training rows whose leakage group shares a key with a test row. The 3,100 official CLINC150 validation examples are unused. Leakage groups: 20,596 (CLINC150, one per example) and 10,138 (BoolQ); the BoolQ test set has 2,938 groups.

Task	Population	stud._train	rout._train	m_tune	m_tune$_T$	prob_cal	pol._cal	test
CLINC150	in-scope	11,998	1,500	750	508	450	300	4,498
	OOS	80	10	5	4	3	2	1,000
	min. per intent	79	10	5	–	3	2	29
BoolQ	yes	3,995	499	250	250	125	125	2,033
	no	2,396	300	149	149	75	74	1,237

Views. Each training example has a canonical view and one fixed subset view. On CLINC150 ($k = 8$) half of the in-scope examples receive a gold-present view (gold plus seven random other intents) and half a gold-absent view (eight random non-gold intents); OOS examples receive only gold-absent views of eight random intents. BoolQ ($k = 1$) uses half $\{y\}$ and half $\{\text{non-gold}\}$. Evaluation roles use the canonical view, one gold-present view (in-scope only) and one gold-absent view. The per-example generator is derived from a hash of task, identifier, role and seed 17, so views do not depend on file order. No candidate or label is truncated.

No view was truncated; the longest tokenized student input was 66 tokens (CLINC150) and 1,098 tokens (BoolQ).

Table 6: View counts per role and view type (CLINC150 $k = 8$; BoolQ $k = 1$).

Task	Role	Canonical	Gold-present	Known-absent	OOS-absent
CLINC150	student_train	12,078	5,999	5,999	80
	model_tune	755	750	750	5
	prob_cal	453	450	450	3
	test	5,498	4,498	4,498	1,000
BoolQ	student_train	6,391	3,196	3,195	–
	model_tune	399	399	399	–
	prob_cal	200	200	200	–
	test	3,270	3,270	3,270	–

B TEACHER PROMPT TEMPLATE AND IDENTIFIERS

Template. The teacher receives a single user message with the following content, where id_j is the j-th identifier code and the $|\mathcal{U}|$ task options appear in official label order, followed by none as the last option.

```
State:
{state}

Question:
{query}

Choices:
{id_1}: {text of option 1}
{id_2}: {text of option 2}
...
{id_|U|}: {text of option |U|}
{id_|U|+1}: none of the listed options

Return only the exact identifier of the best choice. Do not explain.
```

The message is wrapped in the model's chat template with the generation prompt appended and thinking disabled (`enable_thinking=False`). There is one fixed option order, no ensemble over orders and no chain-of-thought.

Task fields. For CLINC150 the state is the user utterance, the query is "Which intent best describes this request?", option texts are the intent names with underscores replaced by spaces, and the order is the sorted list of intent names. For BoolQ the state is the passage, the query is the question, and the options are "No" and "Yes" in that order. The none text is identical for the teacher, the student and DEC.

Identifier scheme. Candidate codes are proposed in the order: the digits `0–9`, the uppercase letters `A–Z`, the lowercase letters `a–z`, then two-letter strings over the same 52 letters in lexicographic order (`AA`, `AB`, ...). A code is kept only if the tokenizer encodes it as exactly one token, that token is not used by an earlier kept code, and decoding the token returns the code. The first $|\mathcal{U}| + 1$ kept codes are used. For CLINC150 this gives 151 codes starting at `0` and ending at `CF`; for BoolQ the codes are `0`, `1` and `2`. Single-token status of every code was certified at the generation boundary for both teacher candidates before teacher selection.

Teacher distribution and inference. $q_{\mathcal{U}}$ is the softmax of the next-token logits over the $|\mathcal{U}| + 1$ identifier tokens after one prefill of the full prompt. Inference uses vLLM (Kwon et al., 2023) (0.28.1rc1.dev580+g385dce36b) in bfloat16 with a 4,096-token context, generation restricted to the allowed identifier tokens, log-probabilities normalized over the allowed tokens, temperature 0, one output token and prefix caching. Up to four independent single-GPU engines process disjoint shards. A row over the prompt limit or failing after one retry is recorded with status fail or timeout, counts as a teacher error in the headroom diagnostic and receives only the cross-entropy term in KD and

KD-PERM; coverage is reported. Recorded fields per row are the identifier, role, status, $q_{\mathcal{U}}$, argmax, prompt token count and attributed batch GPU-seconds.

Teacher candidates and selection. The two candidates are

`Qwen/Qwen3.8-27B`	revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`
`Qwen/Qwen3-4B`	revision `1cfa9a7208912126459214e8b04321603b3df60c`

both Apache-2.0 and run on one GPU in bfloat16 (Qwen3.8-27B: Qwen Team, 2026; Qwen3-4B: Yang et al., 2025). A larger model of the same generation (Qiu et al., 2026) was not eligible because it is released only in FP8 and did not fit in memory on two GPUs. The selected teacher is Qwen3.8-27B. Selection uses 150 student_train examples per task, drawn with a generator seeded from `s3_pilot`, the task name and seed 17: the candidate with higher mean canonical accuracy over the two tasks is selected, and a gap below one point is broken by lower per-example GPU-seconds.

Selection outcome and engine certification. Pilot canonical accuracy was 0.867 (CLINC150) and 0.900 (BoolQ) for Qwen3.8-27B and 0.667 and 0.867 for Qwen3-4B, means 0.883 and 0.767; the selected teacher is Qwen3.8-27B (per-example GPU-seconds 0.073 against 0.015; the tie-break was not needed). Immediately after selection and before the protocol freeze, the vLLM distribution was compared on 8 pilot rows per task with an HF transformers bfloat16 forward pass on the same prompt tokens, using pre-specified acceptance criteria of argmax agreement $\geq 7/8$, mean maximum absolute difference ≤ 0.02 and mean difference in $1-q_{\text{top}} \leq 0.02$. Argmax agreed on 8 of 8 rows for both tasks; the mean maximum absolute difference was 0.0057 (CLINC150) and 0.0010 (BoolQ), and the mean $1 - q_{\text{top}}$ difference 0.0056 and 0.0010. Before the pilot, the engine memory fraction was reduced from 0.88 to 0.78; this changes only key-value-cache capacity, not logits. Every teacher request on every role succeeded (coverage 1.0, no failures or timeouts).

C TRAINING RECIPE, EXPERIMENT MATRIX AND COMPUTE BUDGETS

Selection and the model_tune diagnostic. Learning rates are selected per method as the candidate with the lower dynamic NLL on the full model_tune role, with ties going to base. The pre-registered model_tune diagnostic evaluates whether KD improves model_tune dynamic NLL by at least 2% relative to both GT-CE and GT-LS while reducing canonical accuracy by no more than one point against either control. The diagnostic is reported descriptively and does not affect model selection or subsequent experiments.

Experiment matrix. Seed 17 covers every method and both learning rates: 8 training runs on CLINC150 and 6 on BoolQ. Seeds 29 and 43 at the selected rates add 8 CLINC150 runs and 6 BoolQ runs. Including DEC, the study contains 29 training runs within the pre-registered limit of 30.

Compute used. All 14 student runs completed. DEC completed after one technical retry following an initial 0.06-GPU-hour infrastructure failure. A post-processing validation defect was detected only after the student runs had completed; it did not affect their outputs or any claim-bearing result. Under amendment A8, the completed student outputs were retained and DEC was rerun with the same configuration, using 0.78 GPU-hours. At this stage, total use was 5.59 training GPU-hours and 2.97 teacher-inference GPU-hours, below the pre-registered limits of 90 and 48 GPU-hours. The seed-29/43 CLINC150 runs used 4.38 GPU-hours (0.49–0.63 per run), and the corresponding BoolQ runs used 0.32 GPU-hours (0.053 per run).

Compute budgets and failure handling. The pre-registered limits were 30 training runs, 3 GPU-hours per training run and 90 training GPU-hours; 48 GPU-hours of teacher inference, including at most 2 GPU-hours for teacher selection; 8 GPU-hours of cost measurement; and at most two technical retries. If a compute limit were reached, experiments not yet started would be omitted in a pre-specified order, beginning with the BoolQ student runs, and reported as not run rather than interpreted as negative results. A technical failure could be retried once with the same configuration and seed; a non-finite loss was treated as a failed run without retry, whereas a finite run below the majority baseline remained a valid result.

Table 7: Training recipe. Student settings are shared by GT-CE, GT-LS, KD and KD-PERM.

Item	Setting
Student backbone	`answerdotai/ModernBERT-base` (Warner et al., 2025), revision `8949b909ec900327062f0ebf497f51aef5e6f0c8`; pre-trained initialization, full fine-tuning
Student input	`State:\n{state}\n\nQuestion:\n{query}` followed by `\n\nCandidate:\n{candidate text}`; none is the last pseudo-candidate; softmax over $\lvert\mathcal{A}\rvert + 1$ scores
Optimizer	AdamW (Loshchilov & Hutter, 2019), weight decay 0.01, separate backbone and head parameter groups
Schedule	Constant learning rate, no warmup, no scheduler
Base learning rate	CLINC150: backbone 3×10^{-5}, head 3×10^{-4}; BoolQ: backbone 3×10^{-5}, head 1×10^{-4} (configurations selected for this backbone in a prior study)
Learning-rate candidates	base and half (both group rates $\times0.5$)
Batch	32 training views per optimizer step; micro-batching by token budget with the same window loss
Gradient clipping	1.0
Precision	float32 weights, bfloat16 autocast
Truncation	none
Epochs and checkpoint	3 epochs; last checkpoint; no early stopping or epoch selection
Data order	all training views of student_train (canonical and subset) per epoch in an order shuffled by run seed and epoch
Seeds	17 for every method and learning rate; 29 and 43 at the selected rate for every method on both tasks
Objectives	GT-CE; GT-LS with $\epsilon = 0.1$ over $\mathcal{A}\cup\{\varnothing\}$; KD with $\lambda = 0.5$, $T = 1$ and target $r_{\mathcal{A}}$; KD-PERM as KD with target $\tilde{r}_{\mathcal{A}}$ (Appendix D); failed teacher rows use cross-entropy only
Methods by task	CLINC150: GT-CE, GT-LS, KD, KD-PERM; BoolQ: GT-CE, GT-LS, KD
DEC model	`Qwen/Qwen3-0.6B` (Yang et al., 2025), revision `c1899de289a04d12100db370d81485cdf75e47ca`
DEC adaptation	LoRA (Hu et al., 2021), rank 16, $\alpha = 16$, dropout 0, all linear layers
DEC prompt and loss	teacher template restricted to the view's candidates plus none; cross-entropy on the gold identifier token and the end-of-sequence token; evaluation probabilities are the softmax over the view's identifier tokens
DEC optimization	learning rate 3×10^{-5} (no head group), 32 examples per step, bfloat16 autocast, 3 epochs, last checkpoint, seed 17, no learning-rate choice; CLINC150 only

D METRIC, PERMUTATION AND BOOTSTRAP DETAILS

Metrics. For a view v of example i with candidate set $\mathcal{A}_v$ and target $y_{\mathcal{A}_v}$, the view NLL is $-\log p_\theta(y_{\mathcal{A}_v} \mid s_i, \mathcal{A}_v)$. Dynamic NLL averages view NLL over the subset views of each example and then over examples (Section 4.3). Canonical NLL averages over canonical views. Per-view-type NLL averages over all views of one type, with gold-absent views split into known-absent and OOS-absent by the example's population. Canonical accuracy is the fraction of canonical views whose argmax equals the target. None accuracy is the fraction of gold-absent views whose argmax is none, overall and separately for known-absent and OOS-absent views. The canonical Brier score is $\frac{1}{N}\sum_i\sum_c(p_{ic} - \mathbb{1}[c = y_i])^2$ over the $|\mathcal{U}| + 1$ outputs. ECE uses the maximum probability as confidence, 15 equal-width bins on $[0, 1]$, and the bin-size-weighted mean absolute gap between confidence and accuracy, on native probabilities.

Temperature scaling. For each trained model one temperature is fitted by minimizing canonical NLL on the prob_cal canonical views (Guo et al., 2017), using golden-section search on the inverse temperature over $[1/20, 1/0.05]$ with 80 iterations; boundary values are also evaluated and boundary hits are disclosed. The fitted temperature divides the logits of all test views, and the resulting dynamic NLL is $\mathrm{TS}(\cdot)$. Two temperature-scaled comparisons, $\mathrm{TS(KD)} - \mathrm{TS(GT\text{-}CE)}$ and $\mathrm{TS(KD)} - \mathrm{TS(GT\text{-}LS)}$, belong to the inferential family.

Permuted teacher. For an in-scope example with gold y, $\tilde{q}_{\mathcal{U}}(y) = q_{\mathcal{U}}(y)$ and $\tilde{q}_{\mathcal{U}}(\varnothing) = q_{\mathcal{U}}(\varnothing)$, and the values $\{q_{\mathcal{U}}(b)\}_{b \in \mathcal{U} \setminus \{y\}}$ are reassigned to the labels of $\mathcal{U} \setminus \{y\}$ by a permutation π_i. For an OOS example, $\tilde{q}_{\mathcal{U}}(\varnothing) = q_{\mathcal{U}}(\varnothing)$ and the values over all of $\mathcal{U}$ are permuted. π_i is drawn from a generator seeded from the string `kd_perm`, the example identifier and seed 17, and is identical across epochs, seeds and learning rates. $\tilde{r}_{\mathcal{A}}$ is the projection (Equation 2) of $\tilde{q}_{\mathcal{U}}$. The permutation preserves the multiset of teacher values, hence entropy and the gold and none mass on the full label set, and removes the association between teacher values and specific wrong labels.

Grouped bootstrap. Each resample draws the G leakage groups of the test set with replacement (G draws, seed 17, 10,000 resamples); a group contributes all its examples and views with its multiplicity. In each resample the dynamic NLL of each model is recomputed as a multiplicity-weighted mean over examples, the KD minus control difference is taken per seed, and the statistic is its mean over seeds 17, 29 and 43 on both tasks. On the CLINC150 test set every example is its own group ($G = 5{,}498$); on BoolQ $G = 2{,}938$ for 3,270 examples. The 99% interval is the 0.005 and 0.995 percentile pair of the resampled statistics, the Bonferroni level for the five CLINC150 comparisons. Canonical accuracy differences (KD − GT-CE, KD − GT-LS) use the same resamples and level for non-inferiority. BoolQ comparisons use the same level and are descriptive.

Seed-sign rule and equivalence. As pre-registered, "better" requires the upper bound below zero and all per-seed point differences of the same sign as the estimate, and "non-inferior" requires the lower bound above -0.01 and all per-seed accuracy differences above -0.01. All five CLINC150 family comparisons and both accuracy comparisons had three per-seed differences of the same sign (Table 10). The equivalence margin is $\delta = 0.02 \times 0.0861 = 0.0017$ nats, from the model_tune dynamic NLL of GT-CE at its selected rate; "no effect larger than δ" requires the 99% interval within $[-\delta, +\delta]$. No family comparison met it; all five were "better".

Intervals outside the family. The accuracy condition for N^* uses a 95% grouped bootstrap interval of student minus teacher canonical test accuracy with the same resampling scheme.

Headroom diagnostic. On model_tune$_T$ each example falls in one of four cells by the correctness of the teacher (canonical argmax of $q_{\mathcal{U}}$) and the cheap reference. Gain is the fraction correct only for the teacher, harm the fraction correct only for the cheap reference, and net is gain minus harm; teacher failures count as teacher errors. The reported rule is net ≥ 0.01 and gain ≥ 0.03 on point estimates. Also reported are the teacher's none recall on OOS rows, the OOS count and the mean teacher none mass by population. The cheap reference is chosen by canonical accuracy on model_tune$_T$ among prior-study models, after excluding models whose recorded per-request cost is not below the teacher's and models whose clean retraining would exceed 3 GPU-hours. Its none text is its original abstain text ("None of the available options is a correct answer to the question."), since these models are not modified. The selected references were a classifier (CLINC150) and a dual encoder (BoolQ). The CLINC150 reference had been trained on 511 of the 512 rows of model_tune$_T$ (BoolQ: 322 of 399), so the comparison is contaminated. The pre-registration stated an expected downward bias (against the teacher), but the diagnostic files treat the direction as not established, and we use that direction-agnostic wording; the diagnostic is exploratory. As a clean companion, the same cells are computed against the seed-17 GT-CE student of this study, which never trained on model_tune (Table 1; teacher none recall on the 4 OOS rows of CLINC150 model_tune$_T$ was 4/4, 95% Clopper–Pearson $[0.398, 1.000]$). On CLINC150 the teacher has no headroom against either reference and is less accurate than both. The large BoolQ companion net reflects the seed-17 GT-CE student (accuracy 0.544 on this cohort, below its majority-class rate) and carries no evidence about teacher information.

Teacher calibration and none behavior. On canonical views the teacher's NLL, Brier score and ECE were 0.467, 0.199 and 0.052 on CLINC150 model_tune$_T$ (0.368, 0.171 and 0.040 on BoolQ) and 0.551, 0.213 and 0.041 on the CLINC150 test set (0.367, 0.163 and 0.025 on BoolQ). Its mean none mass was 0.066 on in-scope and 0.925 on OOS rows of CLINC150 model_tune$_T$, and 0.065 and 0.852 on student_train, where it labelled 73 of 80 OOS rows as none (0.913, 95% Clopper–Pearson $[0.828, 0.964]$). Mean off-gold teacher probability by identifier position on CLINC150 student_train has a Pearson correlation of 0.052 with position and a max-to-min ratio of 1,127 across

positions; this diagnostic is descriptive and does not bear on the wording of KD − KD-PERM, which is always "teacher non-gold structure".

Teacher mass log. Table 8 gives the mean projected teacher mass off the gold label and on none, per view type, from the CLINC150 student_train records, as a descriptive account of what the KD and KD-PERM targets contain.

Table 8: Mean projected teacher mass per view type on CLINC150 student_train (no failed rows). Off-gold: mass on presented labels other than gold; none: $r_{\mathcal{A}}(\varnothing)$; target: $r_{\mathcal{A}}(y_{\mathcal{A}})$, which equals the none mass on gold-absent views.

		KD target			KD-PERM target		
View type	Views	Off-gold	None	Target	Off-gold	None	Target
Canonical	12,078	0.0976	0.0704	0.8377	0.0976	0.0704	0.8377
Gold-present	5,999	0.0035	0.1582	0.8383	0.0036	0.1581	0.8383
Known-absent	5,999	0.0055	0.9945	0.9945	0.0052	0.9948	0.9948
OOS-absent	80	0.0023	0.9977	0.9977	0.0066	0.9934	0.9934

E PER-SEED RESULTS AND LEARNING-RATE SELECTION

All results are means over seeds 17, 29 and 43. Table 9 gives the per-seed test results, Table 10 the per-seed differences behind the sign rule, Table 11 temperatures and per-view NLL after temperature scaling, and Table 12 per-view NLL and none accuracy without scaling; no fitted temperature hit a search boundary.

Table 9: Per-seed test results at the selected learning rates, with temperatures fitted on canonical (T) and subset (T_{sub}) prob_cal views; none accuracy on gold-absent views.

Task	Method	Seed	DNLL	TS-DNLL	Canon. acc.	None acc.	T	T_{sub}
CLINC150	GT-CE (half)	17	0.189	0.122	0.907	0.979	2.844	2.034
		29	0.203	0.104	0.914	0.991	2.788	2.534
		43	0.100	0.082	0.908	0.980	1.656	1.385
	GT-LS (base)	17	0.169	0.104	0.921	0.993	0.734	0.592
		29	0.173	0.108	0.927	0.996	0.705	0.624
		43	0.178	0.106	0.924	0.996	0.727	0.608
	KD (base)	17	0.077	0.077	0.913	0.985	0.913	0.855
		29	0.089	0.089	0.918	0.988	1.097	1.044
		43	0.079	0.079	0.914	0.987	1.042	0.993
	KD-PERM (half)	17	0.111	0.111	0.917	0.993	0.987	0.940
		29	0.120	0.122	0.921	0.994	0.923	0.920
		43	0.096	0.095	0.908	0.991	0.935	0.859
BoolQ	GT-CE (base)	17	0.827	0.671	0.552	0.517	3.743	3.733
		29	1.013	0.649	0.626	0.417	4.812	6.143
		43	1.012	0.653	0.606	0.500	5.673	5.969
	GT-LS (half)	17	0.765	0.657	0.596	0.534	2.958	3.064
		29	0.755	0.680	0.618	0.287	1.766	3.360
		43	0.555	0.536	0.766	0.652	1.461	1.503
	KD (half)	17	0.612	0.540	0.746	0.624	2.100	2.140
		29	0.734	0.628	0.658	0.486	2.232	2.620
		43	0.689	0.731	0.668	0.207	0.751	1.956

Descriptively (not pre-registered), KD-PERM has lower raw dynamic NLL than both ground-truth controls (0.109 against 0.164 and 0.174) but not after temperature scaling (0.109 against 0.102 and 0.106), and $\mathrm{TS(KD)} - \mathrm{TS(KD\text{-}PERM)}$ is -0.028 (per seed -0.035, -0.033, -0.017), close to the raw KD − KD-PERM difference of -0.027.

Table 10: Per-seed differences (KD minus control; accuracy as a proportion) and their three-seed mean, which is the bootstrap statistic. TS_{sub} is descriptive.

Task	Comparison	Endpoint	Seed 17	Seed 29	Seed 43	Mean [99% CI]
CLINC150	KD − GT-CE	DNLL	-0.113	-0.114	-0.021	-0.082 $[-0.102, -0.064]$
	KD − GT-LS	DNLL	-0.092	-0.084	-0.099	-0.092 $[-0.099, -0.084]$
	KD − KD-PERM	DNLL	-0.035	-0.031	-0.017	-0.027 $[-0.035, -0.020]$
	KD − GT-CE	TS-DNLL	-0.045	-0.015	-0.003	-0.021 $[-0.028, -0.014]$
	KD − GT-LS	TS-DNLL	-0.027	-0.020	-0.027	-0.025 $[-0.033, -0.017]$
	KD − GT-CE	TS_{sub}	-0.035	-0.015	-0.004	-0.018 $[-0.026, -0.010]$
	KD − GT-LS	TS_{sub}	-0.020	-0.017	-0.019	-0.019 $[-0.028, -0.010]$
	KD − GT-CE	Accuracy	$+0.006$	$+0.004$	$+0.006$	$+0.0055$ $[-0.0005, +0.0116]$
	KD − GT-LS	Accuracy	-0.007	-0.009	-0.010	-0.0085 $[-0.0144, -0.0026]$
BoolQ	KD − GT-CE	DNLL	-0.215	-0.278	-0.323	-0.272 $[-0.307, -0.239]$
	KD − GT-LS	DNLL	-0.153	-0.021	$+0.134$	-0.013 $[-0.028, +0.002]$
	KD − GT-CE	TS-DNLL	-0.131	-0.021	$+0.078$	-0.025 $[-0.035, -0.014]$
	KD − GT-LS	TS-DNLL	-0.117	-0.052	$+0.195$	$+0.009$ $[+0.000, +0.018]$
	KD − GT-CE	TS_{sub}	-0.131	-0.023	$+0.008$	-0.049 $[-0.056, -0.040]$
	KD − GT-LS	TS_{sub}	-0.117	-0.041	$+0.126$	-0.011 $[-0.017, -0.003]$
	KD − GT-CE	Accuracy	$+0.194$	$+0.032$	$+0.061$	$+0.0959$ $[+0.0793, +0.1124]$
	KD − GT-LS	Accuracy	$+0.150$	$+0.039$	-0.098	$+0.0304$ $[+0.0173, +0.0436]$

Table 11: Temperatures and test NLL after scaling with T, by view type (means over three seeds); $TS\text{-}DNLL_{sub}$ uses T_{sub} (descriptive). DEC: $T = 1.399$, TS-DNLL 0.120.

Task	Method	T	T_{sub}	Canon.	Gold-pres.	Known-abs.	OOS-abs.	TS-DNLL	$TS\text{-}DNLL_{sub}$
CLINC150	GT-CE	2.429	1.984	0.386	0.133	0.078	0.090	0.102	0.100
	GT-LS	0.722	0.608	0.408	0.164	0.061	0.078	0.106	0.100
	KD	1.018	0.964	0.304	0.141	0.036	0.050	0.081	0.082
	KD-PERM	0.948	0.906	0.363	0.216	0.030	0.047	0.109	0.110
BoolQ	GT-CE	4.743	5.282	0.998	0.592	0.724	–	0.658	0.658
	GT-LS	2.062	2.642	0.871	0.468	0.781	–	0.624	0.620
	KD	1.694	2.239	0.814	0.394	0.872	–	0.633	0.609

BoolQ differences (descriptive). Over seeds 17, 29 and 43 (Table 10), KD − GT-CE was -0.272 $[-0.307, -0.239]$ in dynamic NLL, -0.025 $[-0.035, -0.014]$ after temperature scaling and $+9.59$ points $[+7.93, +11.24]$ in accuracy; KD − GT-LS was -0.013 $[-0.028, +0.002]$, $+0.009$ $[+0.000, +0.018]$ and $+3.04$ points $[+1.73, +4.36]$ (99% grouped bootstrap intervals). Only GT-CE (accuracy 0.595) is below the test majority-class rate of 0.622; GT-LS (0.660) is above it. Against GT-LS, KD's NLL advantage is not resolved and disappears after temperature scaling, and the per-seed signs disagree for every KD − GT-LS endpoint and for the scaled KD − GT-CE difference; the large KD − GT-CE gaps compare against an undertrained control. BoolQ is descriptive and outside the family.

Model_tune diagnostic. At the selected rates, KD had 29.1% lower dynamic NLL than GT-CE and 58.2% lower than GT-LS on CLINC150 model_tune, with canonical accuracy not lower than either (differences 0.000 and $+0.001$), so the reported diagnostic passed. On BoolQ (recorded only) the relative improvements were 25.7% and 16.5%.

Precision projection (amendment A4). Before test results were inspected, the 99% bootstrap half-widths of the seed-17 model_tune differences were scaled by $\sqrt{n_{\text{model_tune}}/n_{\text{test}}} = \sqrt{755/5{,}498}$ to project test half-widths: 0.0145 for KD − GT-CE (model_tune estimate -0.025 $[-0.068, +0.010]$) and 0.0066 for KD − GT-LS (-0.085 $[-0.104, -0.068]$); for the in-scope stratum alone, 0.0156 and 0.0077. All exceed $\delta = 0.0017$, so the test was not expected to be able to

Table 12: CLINC150 test NLL by view type and none accuracy on gold-absent views: known-absent (gold in $\mathcal{U}$ but not presented; 4,498 views) and OOS-absent (gold is none; 1,000 views). Students: mean over seeds 17, 29, 43; no temperature scaling. Descriptive.

	NLL by view type				None accuracy	
Method	Canonical	Gold-present	Known-abs.	OOS-abs.	Known-abs.	OOS-abs.
GT-CE	0.635	0.263	0.086	0.119	0.984	0.981
GT-LS	0.579	0.157	0.190	0.173	0.997	0.988
KD	0.305	0.142	0.036	0.050	0.987	0.986
KD-PERM	0.366	0.212	0.034	0.049	0.994	0.988
DEC	0.810	0.206	0.112	0.079	0.974	0.986
Teacher	0.551	0.587	0.014	0.029	0.996	0.990

Table 13: Learning-rate selection on the full model_tune role (seed 17): dynamic NLL and canonical accuracy for both candidates; the rate with lower dynamic NLL is selected (ties to base).

		Dynamic NLL		Canon. acc.		
Task	Method	Base	Half	Base	Half	Selected
CLINC150	GT-CE	0.108	0.086	0.966	0.967	half
	GT-LS	0.146	0.178	0.966	0.948	base
	KD	0.061	0.073	0.967	0.956	base
	KD-PERM	0.077	0.076	0.966	0.963	half
BoolQ	GT-CE	0.846	1.375	0.544	0.714	base
	GT-LS	0.762	0.752	0.607	0.596	half
	KD	0.714	0.628	0.602	0.747	half

show "no effect larger than δ". model_tune contains about 0.7% OOS examples and the test set about 18%; δ was not changed after test results were inspected.

F MEASUREMENT PROTOCOL

Systems and paths. Three systems are measured on CLINC150: the seed-17 KD student at its selected learning rate, the teacher and DEC. The student is served with float32 weights and bfloat16 autocast, and DEC with an unmerged LoRA adapter, float32 weights and bfloat16 autocast, in both cases the path used for evaluation. The teacher is served with the inference configuration of Appendix B.

Requests and blocks. The request list is a fixed draw of canonical CLINC150 test requests with a generator seeded from the string `measure` and seed 17. Each system runs at batch size 1 (B1) and 32 (B32) in five blocks of 512 requests after 32 warm-up requests, giving 30 measured blocks. Timing starts at the serving entry from raw request text and includes tokenization, candidate packing, host-device transfer and output selection.

Recorded quantities. Decisions per second; GPU-seconds per decision, counting all GPUs used; joules per decision from NVML energy counter differences; B1 per-request median and 95th-percentile latency; and B32 batch-time median and 95th percentile, never reported as per-request latency. Model loading, identifier certification and warm-up are reported separately as one-time costs.

Measurement window. Each block is measured with exclusive access to one GPU and its NUMA-local CPU resources. Measurements with competing GPU activity are deferred or repeated at most twice within the 8-GPU-hour measurement budget. Any block containing an invalid decision is excluded from claim-bearing cost measurements. Under amendment A9, all cells were measured on a second GPU of the same model (NVIDIA RTX PRO 6000 Blackwell) on the same NUMA

node because exclusive access to the pre-registered device was unavailable; all exclusivity, deferral, block, cohort and precision rules were unchanged.

Serving–evaluation agreement (amendments A6 and A7). For each system, the B1 and B32 decisions of every costed block on the 512-request cohort are compared with each other (same block index) and with the system's evaluated test decisions (argmax of the saved canonical test predictions, lowest index on ties; failed teacher records never agree). A system is claim-bearing only if all three agreements, taken as minima over blocks, are at least 0.99; otherwise its accuracy and costs are descriptive, the dominance decision is undetermined, and for the student and the teacher the replacement, deployment-benefit and break-even readings are withheld. Student serving forwards at most 8 requests per model call inside the timed call.

Break-even inputs and decoder dominance. The numerator of N^* is the teacher annotation cost of the student_train record, the cost of both seed-17 KD learning-rate runs and itemized preparation, with CPU-only preparation counted as zero GPU-seconds. Per-request costs come separately from the B1 and B32 measurements, in GPU-seconds and joules. A non-positive denominator means no break-even. The counterfactual $N^*_{\text{GT-CE}}$ uses the same formula for the seed-17 GT-CE student with zero annotation cost, both GT-CE learning-rate runs as training cost and, as an assumption, the measured per-request cost of the KD student (same architecture and serving path); it is the cost of replacing the teacher with any fine-tuned student. Because the information claim does not hold, no amortization claim is made for distillation itself. Sensitivity: teacher B1 cost per decision ranged 0.1586 to 0.1890 GPU-s across its five blocks; with the numerator of 6,211.8 GPU-s and the student's B1 cost, the B1 N^* would range from 41,437 to 51,975. DEC dominance is checked on point estimates of canonical accuracy, B32 decisions per second and B1 median latency; DEC uses one learning rate while the student chooses between two, and this asymmetry is stated with the result. The accuracy condition for the replacement reading is met (student minus teacher canonical accuracy $+0.057$, 95% grouped bootstrap $[+0.047, +0.067]$, 10,000 resamples).

Outcome. All 30 blocks (three systems, two batch sizes, five blocks) were valid on their first attempt, with no invalid decisions and no postponement; measurement used 0.47 GPU-hours including loading and warm-up. Agreements were: student 1.000 (B1–B32, B1–evaluated and B32–evaluated); DEC 0.990, 0.990 and 0.992; teacher 0.996, 0.992 and 0.992. All three systems are therefore claim-bearing. One-time costs, excluded from per-decision costs, were about 11–13 s of model loading for the student and DEC (identifier certification 0.3–0.4 s for DEC) and 154–160 s for the teacher. Student training energy was recorded only for the training loop (0.60 and 0.62 MJ for the two KD runs), not for whole processes, so joule-based N^* is unavailable and cost claims are made in GPU-seconds only. Decoder dominance fails on accuracy (0.886 against 0.913) and B1 median latency (55 against 38 ms; block medians 35–71 ms for DEC and 30–41 ms for the student, so the latency difference is unstable) although DEC has higher B32 throughput (63.8 against 25.1 decisions/s); the deployment-benefit and replacement readings are allowed at both batch sizes, and the distillation amortization reading is not, because the information claim does not hold.

Table 14: Serving measurements on 512 CLINC150 test canonical requests (mean over five valid blocks; second GPU of the same model, amendment A9). Latency in seconds: B1 per-request and B32 per-batch percentiles (never per-request for B32). Agreement: B1 row, B1–B32 / B1–evaluated; B32 row, B32–evaluated.

System	Batch	Decisions/s	GPU-s/decision	J/decision	Latency p50	Latency p95	Agreement
KD student	B1	25.6	0.0391	8.53	0.038	0.049	1.000 / 1.000
	B32	25.1	0.0402	10.63	1.256	1.740	1.000
DEC	B1	19.5	0.0546	10.33	0.055	0.082	0.990 / 0.990
	B32	63.8	0.0157	5.75	0.502	0.516	0.992
Teacher	B1	5.76	0.1743	90.18	0.171	0.206	0.996 / 0.992
	B32	7.88	0.1269	70.67	4.090	4.140	0.992

Table 15: Break-even inputs and N^* (requests) for the KD student and the counterfactual GT-CE student, whose per-request cost is assumed equal to the measured KD student's. Training: both seed-17 learning-rate runs (KD 1,850 and 2,228 GPU-s; GT-CE 2,121 and 1,812), whole-process GPU time of all attempts. Preparation: view construction is CPU-only. Joule N^* is unavailable (whole-process training energy not measured). Incremental distillation cost over GT-CE: 2,279 GPU-s, for a three-seed dynamic NLL difference of -0.082 $[-0.102, -0.064]$ and an accuracy difference of $+0.55$ points $[-0.05, +1.16]$. Break-even counts, not realized savings.

Student	Unit	Batch	Annotation	Training	Preparation	c_{teacher}	c_{student}	N^*
KD	GPU-s	B1	2,134	4,077	0	0.1743	0.0391	45,937
KD	GPU-s	B32	2,134	4,077	0	0.1269	0.0402	71,697
GT-CE	GPU-s	B1	0	3,933	0	0.1743	0.0391	29,084
GT-CE	GPU-s	B32	0	3,933	0	0.1269	0.0402	45,393
KD, GT-CE	J	B1 / B32	922,752 / 0	n/a	0	90.2 / 70.7	8.5 / 10.6	n/a

G Extended related work

Teacher calibration and variance. Calibrating the teacher improves distillation (Kim et al., 2025), larger teachers give better students once calibrated (Yang & Song, 2021), KL-based distillation is sensitive to teacher calibration (Fan et al., 2024), and calibrated students can be obtained from uncalibrated teachers with augmentation (Mishra et al., 2023). The interaction of label smoothing with distillation is qualified by Shen et al. (2021) and Chandrasegaran et al. (2022). We do not calibrate the teacher; this bears on the scope of the results (Section 6).

LLM teacher signals. Rationale supervision (Hsieh et al., 2023) and task-targeted distillation (Zhou et al., 2024) train small generative models from LLMs. Probe-based soft labels from teacher hidden states can distill better than output logits, because the vocabulary projection makes output distributions brittle (Brown & Russell, 2026); our teacher signal is such a vocabulary-restricted distribution. In generative distillation, mixing hard and soft labels outperforms either alone (Wang et al., 2026a). An LLM can also supply outlier examples to train a classifier's abstention (Xu et al., 2023). Distillation for OOS scoring uses fine-tuned teachers (Wu et al., 2023). In-domain OOS inputs remain hard for intent classifiers (Zhang et al., 2022).

Intent detection with LLMs and encoders. Encoder-to-LLM routing and small-large collaboration address intent and OOS decisions (Zaera et al., 2025; Chen & Wang, 2025; Castillo-López et al., 2025). A systematic evaluation of open-weight zero-shot intent classifiers reports inconsistent effects of instruction tuning on calibration (Ganesh et al., 2026). Zero-shot cross-encoders, embedding models, rerankers and LLMs have been compared across text classification benchmarks (Aarab, 2026), and embedding-head and instruction-tuned LoRA decoders have been compared for classification (Yousefiramandi & Cooney, 2025).

Cost, cascades and amortization. Distillation is usually a more cost-efficient use of a fixed budget than annotation (Kang et al., 2023); energy accounting treats teacher work as a near-fixed overhead whose break-even depends on reuse (Lambert & Luccioni, 2026); active selection reduces the number of teacher-labeled examples (Luccioli et al., 2025). Fine-tuned encoders have been compared with prompted LLMs on cost (Gonzalez, 2026; Palo et al., 2024). Cascades and routers reduce LLM cost by deferring only some inputs (Chen et al., 2023; Ong et al., 2025; Nie et al., 2024; Ramírez et al., 2024a), a problem related to selective classification (Geifman & El-Yaniv, 2017). Our comparison point is a fine-tuned small constrained decoder rather than a prompted LLM.

H Amendment log

A1–A5 are also in the frozen protocol file. A1–A8 were made before the protocol freeze at the end of teacher selection and before any production training; A9 was made after the model_tune diagnostics and before inspection of test results and any cost measurement.

ID	Change
A1	After the teacher probe, the headroom stop rule was converted to a reported diagnostic; subsequent experiments were no longer conditioned on its outcome. This amendment is disclosed as post-probe.
A2	After design review: seeds regardless of the diagnostic; KD-PERM; inferential temperature-scaled comparisons; five-comparison CLINC150 family at 99%; BoolQ outside the family; seed-sign rule; 10,000 resamples.
A3	After design review: GT-CE counterfactual for N^* and the distillation cost-claim rule; equivalence margin δ; OOS recall intervals and population-shift disclosure; teacher calibration and position diagnostics; fresh-student headroom companion; stated contamination bias direction.
A4	"Teacher non-gold structure" wording for KD − KD-PERM; precision projection before inspection of test results; gold-present none mass and per-view NLL reported next to KD − GT-CE and KD − GT-LS; null or negative outcomes worded as properties of KD with dustbin projection from this teacher.
A5	One-time engine certification of the teacher (vLLM against HF transformers) before the freeze; serialized first BoolQ run for projection.
A6	Decoder batch-equivalence check and 0.99 agreement threshold for DEC costs; corrections to a capacity note; population shift stated next to δ.
A7	Serving–evaluation agreement (≥ 0.99) for DEC, the student and the teacher; block validity and costing rules; at most 8 student requests per call.
A8	Added a fail-closed rule for abnormal executions: an affected stage stops, outputs are retained only after documented validation, and invalid executions do not contribute to claim-bearing cost estimates; compute budgets remain upper limits.
A9	Cost measurement was moved to a second GPU of the same model on the same NUMA node because exclusive access to the pre-registered device was unavailable; all other measurement rules were unchanged.