

Large Decision Models: Repurposing Large Language Models as Single-Pass Decision Makers

Ziyi Chen^{1*} Yan Zhang^{2*} Jianhui Wei^{3*} Jianing Yin⁴ Daoan Zhang⁵

¹University of Illinois Urbana-Champaign ²National University of Singapore

³Zhejiang University ⁴University of Pennsylvania ⁵University of Rochester

*Equal contribution.

Abstract

Large language models make decisions through parallel prefill and sequential decoding. Although effective, this approach incurs decoding costs and does not explicitly expose decision probabilities. Existing task-specific classifiers require per-task training and offer limited generalization to unseen decision schemas. We propose Large Decision Models (LDMs), repurposing language models to return explicit probabilities over labels specified at run time using one prefill forward pass. We first adapt models to this interface through Diverse Decision Supervision (DDS), covering varied domains, rules and labels. To further narrow the performance gap to a reasoning model, Self-Posed Reasoning Distillation (SPRD) adds tasks generated by a frozen teacher, filters them for answer consistency, and trains on label distributions obtained from repeated teacher reasoning. We evaluate on DecisionVista, pairing new-schema tests with held-out human-labelled sources. Across three base models, the recipe improves accuracy over a step-matched baseline by 1.8 and 1.5 percentage points, respectively, achieving a 182 $\times$ speedup in latency over reasoning rollout on Gemma-4-31B. Reasoning remains stronger on the new-schema test, motivating a confidence router for uncertain decisions. On development sets, routing recovers a median 92% of reasoning’s accuracy advantage while generating 27–58% of its tokens. Code is available at <https://github.com/JianhuiWei7/LDM>.

1. Introduction

Many applications require a decision: route a support ticket, judge whether a claim is supported, or rate an incident’s severity. Large Language Models (LLMs) acquire broad knowledge through pretraining and generalize across diverse

tasks. However, generating decisions through autoregressive decoding (*rollout*) incurs latency and computational costs and does not explicitly expose decision probabilities, limiting their suitability for tasks that require fast responses and confidence-aware choices. Conventional BERT-based classifiers offer efficient predictions, but task-specific finetuning ties their output heads to predefined labels (Devlin et al., 2019). To combine classification efficiency with the generality of pretrained language models, we study **Large Decision Models** (LDMs), which map a context, a question and a label set specified at run time to a distribution over the labels in a single forward pass—an operation we call *readout* (§2; Fig. 1). Existing systems such as Jev, Nimble, OpenJev and Laya offer this interface (TypeSafe, 2026; Bespoke Labs, 2026; OpenJev authors, 2026; Convai Innovations, 2026).

To adapt LLM rollout to LDM readout, we use *decision tuning*: given a context, a question and candidate labels, we fine-tune the model with LoRA to predict the correct label by applying cross-entropy to the normalized option-token logits at the final input position, without generating any sequence. After training, we evaluate *schema transfer*: accuracy on decision schemas, label sets and domains absent from finetuning, alongside calibration and single-pass cost. Schema transfer is feasible because pretraining already places most of the decision in the representation that a single forward pass reads. Without any decision training, the readout of an untuned 27B model reaches 97% of its tuned accuracy, and 1/16 of the tuning data recovers 88% of the tuning gain at 9B (§4). Decision tuning is therefore an *adaptation* of how that representation is read out, not the acquisition of new task knowledge.

Our diagnosis identifies two opportunities for improving transfer (§4). First, decision tuning improves source accuracy more reliably than new-schema accuracy, motivating supervision across broader schemas. Second, reasoning rollout reveals capabilities that the single-pass readout does not fully express: on two new-schema validation sets, the same untuned 27B model is 7.8–9.3 points more accurate

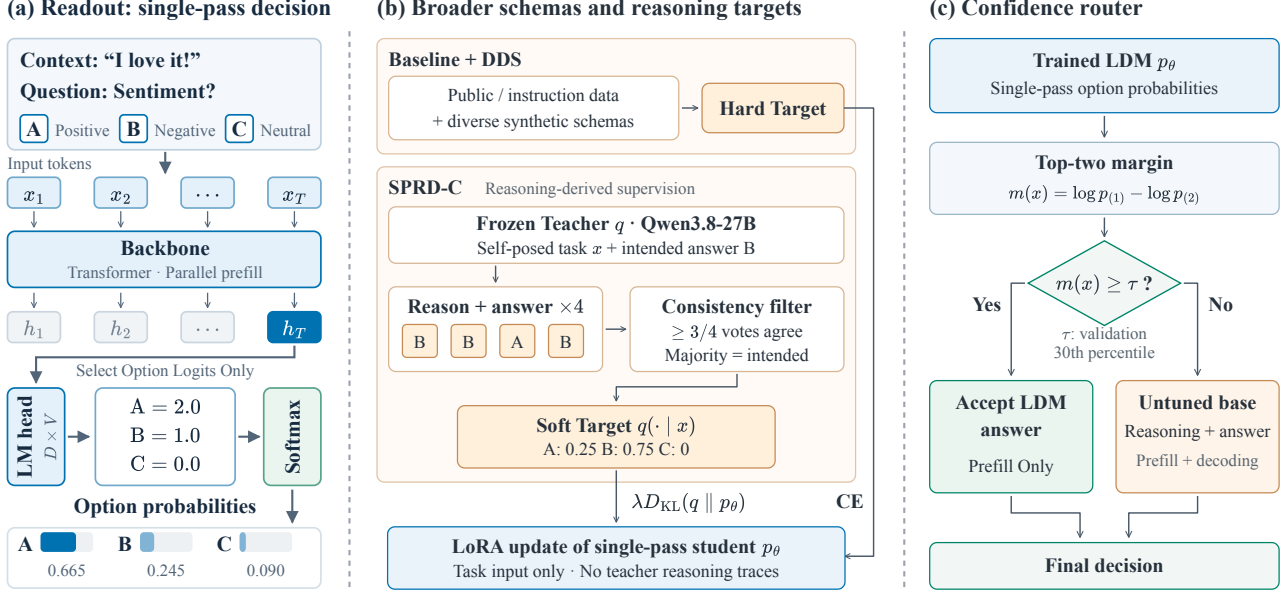


Figure 1. Overview. (a) The LDM processes the task in one prefill, projects the last input hidden state through the LM head, and normalises only the allowed option logits. (b) Baseline and DDS examples supply hard-label cross-entropy supervision. The frozen Qwen3.8-27B teacher generates four reasoning-and-answer samples per self-posed task. SPRD-C filters for consistency and distils answer-vote distributions with $\lambda D_{KL}(q || p_\theta)$. (c) A margin router accepts confident LDM predictions and otherwise invokes the student’s untuned base for one reasoning rollout.

when allowed to reason. This rollout–readout gap motivates two complementary uses of reasoning: distillation during training and selective computation at inference.

Our training recipe builds on these two opportunities (§5). **Diverse Decision Supervision (DDS)** uses synthetic decision schemas to promote generalization beyond familiar tasks and output patterns. Varying domains, rules and label meanings, alongside public and instruction data, encourages the readout to interpret each new schema. **Self-Posed Reasoning Distillation (SPRD)** aims to bring readout performance closer to that of the base model using reasoning rollout. It distils a reasoning teacher’s answer distributions on new self-posed tasks into the readout, making those decisions a training target while retaining single-pass inference.

At deployment, a **confidence router** combines fast readout with selective reasoning rollout. It accepts confident readout answers and sends uncertain decisions to the base model for reasoning rollout, using a threshold fixed on validation data (§5.3).

We evaluate DDS and SPRD separately and together across three base models from two families, with three seeds each (§6). We introduce **DecisionVista**, a benchmark suite pairing new decision schemas (**DecisionVista-Synth**) with human-labelled tasks from eight sources excluded from fine-tuning and method design (**DecisionVista-Human**). Our protocol separates model development from pre-registered

testing on these two sets (§3). Averaged across bases, the combined recipe improves accuracy over a step-matched baseline by 1.8 points on DecisionVista-Synth and 1.5 on DecisionVista-Human. Against a baseline selected on validation data, it gains 0.8 points on DecisionVista-Human and 2.1 on the public typed-decisions benchmark. Our 9B model outperforms Nimble-9B on the same base across transfer evaluations and Nimble-13. On Gemma-4-31B served on two H100 GPUs, tuned readout reduces median latency from 7.38 s to 40.6 ms, a $182\times$ speedup over reasoning rollout (Table 4). On validation sets, routing recovers a median 92% of the rollout–readout gap using 27–58% of full reasoning’s generated tokens (§6).

Our contributions are:

1. **LDM formulation:** single-pass decision models centered on generalization to unseen schemas.
2. **DDS and SPRD:** a training recipe combining synthetic schema diversity with distillation of a reasoning teacher’s decisions.
3. **DecisionVista:** synthetic and human-labelled benchmarks, with a protocol separating model development from pre-registered testing.
4. **Confidence routing:** a deployment strategy combining fast readout with selective reasoning rollout.

5. **Statistically validated gains:** our recipe improves accuracy over the step-matched baseline by 1.8 points on DecisionVista-Synth (95% CI [0.6, 3.1]) and 1.5 on DecisionVista-Human (95% CI [0.9, 2.2]). Tuned readout achieves a $182\times$ median-latency speedup over reasoning rollout on Gemma-4-31B.

2. Large Decision Models

A *Large Decision Model* (LDM) maps a context c and a decision schema specified at run time to a probability distribution over the schema’s allowed outcomes in a single forward pass. A schema consists of a question or decision rule q and a finite label set $\mathcal{Y}$ with natural-language descriptions:

$$p_{\theta}(\cdot \mid c, q, \mathcal{Y}) \in \Delta(\mathcal{Y}), \quad (1)$$

where $\Delta(\mathcal{Y})$ is the set of probability distributions over $\mathcal{Y}$; the predicted outcome is $\arg \max_{y \in \mathcal{Y}} p_{\theta}(y \mid c, q, \mathcal{Y})$. Schemas may express categorical choices, binary judgments or ordinal ratings, with label meanings and any ordering supplied at run time. This allows the same model to address new decisions without a task-specific output head or further training.

A pretrained model can use this interface zero-shot. Decision tuning aligns the model with this interface by adapting autoregressive rollout to single-pass readout. Jev ([TypeSafe, 2026](#)) serves this interface as a hosted service; Nimble-9B ([Bespoke Labs, 2026](#)) and OpenJev-27B ([OpenJev authors, 2026](#)) release LLM weights for it, and Laya ([Convai Innovations, 2026](#)) a 0.3–0.4B encoder with a trained decision head. An LDM should be judged first on *schema transfer*, accuracy on decision schemas, label sets and domains absent from fine-tuning, and then on calibration and cost.

3. Evaluating schema transfer

Testing schema transfer requires both comparison with existing systems and evidence that gains extend beyond the tasks used to develop the recipe. We therefore separate public comparisons, model selection and exploratory diagnosis from independent confirmation (Table 1).

Public benchmarks and development sets. Public benchmarks anchor comparisons with released decision models: Nimble’s full 13-subset benchmark (*Nimble-13*) ([Bespoke Labs, 2026](#)) and the typed-decisions test set reported for Jev and Laya (*Typed-Decisions*) ([Convai Innovations, 2026](#)); four workflows, labels from an LLM teacher. Typed-decisions contains 2,000 decisions across 400 cases (five per case); its authors report 73.5% repeatability of the teacher labels. No model in the primary study trains on their sources. Table 3 also reports Kev’s transfer-v4 test (*Kev-Transfer*;

Table 1. Evaluation sets and primary uses (DV: DecisionVista; Val.: validation suite). Validation sets guide development; the validation suite supports selection and calibration. Test sets were sealed and first read under pre-registered analyses after the models were frozen; later analyses are marked post hoc. No model in the primary study trains on these sets. ≥ 2 of 3 LLMs: kept iff at least two of three blind non-author labellers agree with the author; *DV-Pilot was sealed under its earlier rule (Appendix D). Grey rows appear only in the appendix.

Set	Items	Tasks	Labels	Primary use
Nimble benchmark	3,880	13	human	Public benchmark
Typed-decisions (test)	2,000	4	LLM teacher	Public benchmark
Validation suite	1,256	7	human	Validation
Transfer suite	2,955	8	human	Validation
DV-Dev-A/B	1,177	24	≥ 2 of 3 LLMs	Validation
DV-Synth	1,505	35	≥ 2 of 3 LLMs	Test
DV-Human	2,299	8	human	Test
DV-Stress	1,134	24	≥ 2 of 3 LLMs	Validation
DV-Pilot	1,311	29	3 LLMs agree*	Test

656 items from sources new to Kev, including MMLU and QNLI), read once.

The human-labelled *validation suite* (Val.) supports model selection, temperature calibration and routing thresholds. Other validation sets probe complementary forms of transfer: the human-labelled *transfer suite* tests generalization across sources, while synthetic DV-Dev-A/B test newly constructed decision schemas. These sets guide recipe development, so gains on them require confirmation on untouched sets. The human-suite sources are excluded from fine-tuning, and synthetic evaluation families are screened against training schemas. DV-Dev-A/B contain rule-based decisions written in contrastive groups by gpt-5.5 and kept when gpt-6-astra agreed blind. Source inventories and the earlier DV-Stress and DV-Pilot stages are in Appendices N and D.

DecisionVista tests. The two primary sets were constructed after the primary models were frozen, then sealed and first read under pre-registered analyses; no design decision used them. **DecisionVista-Synth** tests newly constructed schemas: 35 families over 18 decision operators, including capacity allocation, Pareto choice and counterfactual intervention. Four model families author the tasks; the three non-authors label each item blind, and at least two must agree with the author (Appendix K). Families are screened against earlier evaluation families and training schemas; we claim screened non-overlap, not freedom from pretraining exposure. Its labels are model agreement, not human judgement; a human audit protocol is released. We also evaluate without the author that verified synthetic DDS labels (Appendix M). **DecisionVista-Human** complements this test with 2,299 human-labelled items from eight sources absent from all training mixtures, earlier evaluation sets and Super-NaturalInstructions tasks (Appendix L). Here the held-out units are sources; related task types can

appear in training. Tables abbreviate the pair to DV-Synth and DV-Human.

Measurement and protocol. The primary statistics are mean accuracy across task families on DecisionVista-Synth and across sources on DecisionVista-Human, averaged equally across bases. Paired bootstrap intervals resample task families or items within each human-labelled source; a two-level sensitivity analysis also resamples training seeds (Appendix C). Nimble scores average its 13 subset accuracies; typed-decisions intervals resample its 400 cases. Calibration uses negative log-likelihood (NLL), Brier score and expected calibration error (ECE; 15 bins). Appendix D records which sets informed each choice; analyses specified after the sealed reads are marked *post hoc*. Pre-registrations, hashes, per-item predictions and code are in the supplementary material.

4. Diagnosing schema transfer

Our exploratory diagnosis asks why decision tuning transfers unevenly and how much accuracy the single-pass readout leaves unused. Two observations motivate broader schema supervision and reasoning-derived targets.

The decision is largely already in the representation. Across Qwen3.5 sizes from 0.8B to 27B, the untuned single-pass readout reaches 76–97% of the accuracy of the same model after decision tuning, the tuning gain shrinks from 15.8 to 2.3 points as the base grows, and 1/16 of the tuning data already recovers most of it (Appendix T). Pretraining thus supplies the representation; tuning adapts how it is read. The adaptation is light but must reach the backbone: LoRA updates under 0.5% of the parameters, whereas training only the option-letter head on frozen features recovers almost none of the gain (Appendix U).

Source gains do not ensure schema transfer. We rebuilt Nimble’s recipe and varied scale, data, objective and capacity (Appendix B). At 27B, training on 20k records from 36 public datasets raises validation accuracy from 79.3 to 81.8; adding the public benchmark’s training sources gives 89.1 on its eight choice and yes/no subsets, versus OpenJev’s 84.1. Yet neither recipe consistently improves new-schema accuracy: the public-data recipe scores 88.1/87.5 on DV-Dev-A/B and the source-augmented recipe 87.8/88.0, against the untuned model’s 89.1/87.5. Transfer depends on scale and evaluation set—baseline tuning does help at 9B—so this is a motivation to broaden schema coverage, not a claim that supervised tuning cannot transfer.

The rollout–readout gap. The same untuned 27B model is more accurate when allowed to reason (majority of four samples): 96.9 vs. 89.1 on DV-Dev-A, 96.8 vs. 87.5 on

DV-Dev-B and 85.1 vs. 79.3 on the validation suite (Table 7). Where the two modes disagree on DV-Dev-B, reasoning is right 90% of the time. A separate single-sample analysis finds the largest gains on exact multi-step computation (dates and numeric comparisons: 56–66% readout, 98–100% reasoning across the three bases), while reasoning can lose on NLI (Appendix P). Reasoning costs hundreds of generated tokens per decision. These observations motivate DDS to broaden the schemas encountered during training and SPRD to train the readout toward a reasoning teacher’s decisions, paying for teacher computation offline. Confidence routing retains access to reasoning for uncertain decisions at deployment.

5. Training LDMs for schema transfer

We train the readout to interpret unfamiliar decision schemas and to approximate decisions available from a reasoning teacher. DDS broadens schema coverage through synthetic decisions alongside public and instruction data; SPRD supplies reasoning-derived answer distributions on self-posed tasks. At deployment, a confidence router can send uncertain decisions to the student’s untuned base using reasoning rollout.

Implementation. For this study, we implement the LDM interface by normalising the logits of single-token option identifiers:

$$p_{\theta}(y \mid c, q, \mathcal{Y}) = \operatorname{softmax}_{y \in \mathcal{Y}} z_{\theta}(\ell_y \mid x(c, q, \mathcal{Y})), \quad (2)$$

where $x(\cdot)$ serialises the context, question and options, ℓ_y is the letter assigned to outcome y , and $z_{\theta}(\ell_y \mid x)$ is its logit at the final input position. We support *choice* (2–8 named options), *yes/no*, and *ordinal score* (3–6 ordered levels); these bounds are implementation choices, not part of the LDM definition. Each instruction-tuned base trains a LoRA adapter (Hu et al., 2022) on this readout, with option order re-shuffled at every draw. The baseline training data contain 22.4k public and contrastive records (Appendix H).

5.1. Diverse Decision Supervision (DDS)

Learning a common answer format does not by itself ensure that the readout can apply unfamiliar decision rules. DDS uses synthetic task construction to vary domains, rules and label meanings, aiming to encourage schema interpretation beyond the original training tasks. Public datasets and categorical instruction tasks (Wang et al., 2022) provide complementary coverage, while the readout objective stays fixed.

Each run uses the same 31.7k-record sample from the 63.4k pool: additional public sources, Super-NaturalInstructions tasks, and synthetic decisions built from a domain–decision

grid or grounded in real documents (Appendix J). Synthetic labels require blind verification, and synthetic schemas are audited against evaluation families. The pool excludes benchmark-source data and is 13-gram de-contaminated against evaluation items available at construction; the later DecisionVista sets are checked against the frozen mixtures (Appendices K, L).

5.2. Self-Posed Reasoning Distillation (SPRD)

SPRD turns the rollout-readout gap into supervision: the readout learns decisions reached with additional computation, without learning to generate the reasoning trace. Qwen3.8-27B is the frozen task author and reasoning teacher for all primary students; only the 27B student learns from its own untuned base.

Self-posed tasks and decision targets. The teacher proposes schemas across domains, reasoning skills and answer formats, then writes contrastive instances for schemas that pass an evaluation-family audit: 16,040 items from 696 schemas (Appendix N). It answers each item four times using reasoning rollout. The target is $q(y | x) = n_y/m$, where n_y counts votes for y and m is the number of parseable samples. Requiring the top vote share over all four samples to be at least 0.75 retains 14,829 items. The consistency-filtered variant, **SPRD-C**, further keeps the 10,331 items whose reasoning majority agrees with the label intended at item construction. This checks consistency between two computations of the same model, not correctness. The student receives answer distributions, never reasoning traces.

Joint training. Baseline data and DDS records supply supervised labels; pool records supply teacher distributions. We shuffle the enabled data streams together and minimise their mean per-record loss:

$$\ell_\theta(r) = \begin{cases} -\log p_\theta(y_r^* | x_r), & r \in \mathcal{D}_{\text{sup}}, \\ \lambda D_{\text{KL}}(q_r \| p_\theta(\cdot | x_r)), & r \in \mathcal{D}_{\text{SPRD}}, \end{cases} \quad (3)$$

where $\mathcal{D}_{\text{sup}}$ contains baseline data plus DDS when enabled, and $\lambda = 3$. Inference remains one forward pass.

5.3. Confidence routing to reasoning

The router accepts confident readout decisions and sends uncertain items to that student’s untuned base using reasoning rollout (one sample). Its score is the top-two margin $m(x) = \log p_\theta(y_{(1)} | x) - \log p_\theta(y_{(2)} | x)$, equal to the corresponding logit gap in Eq. 2. If $m(x) < \tau$, the reasoning answer is used; otherwise the readout’s answer stands at zero generated tokens. We fix τ at the b -quantile of margins on the validation suite, with $b = 30\%$ unless stated, before reading other sets. This sets validation coverage; the fraction routed on an unseen set can differ. The router requires

no additional training.

6. Results

We first test transfer on the pre-registered evaluations, then examine component effects, compare released models and evaluate confidence routing. We compare baseline training alone, with DDS, with SPRD-C, and with both on Qwen3.5-9B, Qwen3.8-27B and Gemma-4-31B-it, using the fixed Qwen teacher from §5.2. The primary comparator, **step-matched baseline**, trains for 4,032 steps against the full recipe’s 4,031 at batch size 16 (2.9 passes over baseline data). We report three seeds per variant; the pre-registered test estimates use seeds 17/29/41 at 9B and 17/29 at 27B/31B, excluding later seed-41 runs (Appendix H). The validation column averages accuracy on DV-Dev-A, DV-Dev-B and the transfer suite.

Primary transfer results. Both pre-registered primary tests succeed (Table 2). The combined recipe gains **+1.8 points** over the step-matched baseline on DecisionVista-Synth (95% CI [0.6, 3.1]; sign-flip $p=0.009$) and **+1.5** on DecisionVista-Human ([0.9, 2.2]). The gains are larger for Qwen: +2.3/+2.1 on the two sets at 9B and +2.4/+1.9 at 27B, compared with +0.7/+0.6 for Gemma, whose intervals include zero. Neither set is near ceiling (step-matched baseline: 64.5–80.0 on DecisionVista-Synth, 72.8–79.6 on DecisionVista-Human; Appendix Table 16).

Dependence on the baseline comparator. Matching steps controls training compute but can overtrain on the smaller baseline dataset. In a *post-hoc* comparison, we select the baseline on the validation suite from one-pass or step-matched training and full or 0.65-scaled updates, selecting the recipe only over full or 0.65-scaled updates on the same suite. The gains become +0.8 on DecisionVista-Synth ([−0.1, 1.7]), +0.8 on DecisionVista-Human ([0.2, 1.4]) and +2.1 on typed-decisions ([1.2, 3.0]; Appendix Table 35). A well-selected baseline recovers most of the synthetic-schema advantage; the human-labelled and typed-decisions gains remain above zero.

Component effects and validation results. The original single-component variants train fewer steps (DDS: 3,386; SPRD-C: 2,050), so their positive mean effects over the step-matched baseline are descriptive. On DecisionVista-Human, the combination gains +0.4 [0.0, 0.9] over DDS and +0.5 [0.1, 1.0] over SPRD-C. On DecisionVista-Synth, the pre-registered combination-versus-DDS test is not significant (+0.4 [−0.3, 1.1]); the ordered component-testing chain stops there (Appendix Table 17). *Post-hoc* single-component runs at 4,031 steps still beat step-matched baseline on DecisionVista-Human at 9B and 27B; at 27B both have similar mean gains to the combination on

Table 2. Accuracy gains over the step-matched baseline (points; 95% bootstrap intervals; bold: interval above zero). Shading marks the combined recipe. Its DV-Synth and DV-Human means across bases are the pre-registered primary tests; other contrasts are secondary or descriptive.

Base	vs. step-matched baseline	Development	Test (used in no design decision)		
		Transfer + DV-Dev-A/B	Public	Sealed, pre-registered	
			Typed-Decisions	DV-Synth	DV-Human
Qwen3.5-9B	+DDS	+3.2 [1.7, 4.7]	+2.4 [1.0, 3.8]	+1.6 [0.07, 3.2]	+2.2 [1.2, 3.2]
	+SPRD-C	+3.2 [1.6, 4.8]	+1.3 [−0.2, 2.8]	+2.1 [0.02, 4.0]	+2.1 [1.1, 3.1]
	+DDS + SPRD-C	+3.9 [2.4, 5.5]	+2.5 [0.9, 4.2]	+2.3 [0.6, 3.9]	+2.1 [1.1, 3.1]
Qwen3.8-27B	+DDS	+1.6 [0.5, 2.7]	+1.7 [0.5, 2.9]	+1.2 [−0.9, 3.2]	+1.0 [−0.04, 2.1]
	+SPRD-C	+1.9 [0.9, 3.0]	+3.1 [1.8, 4.4]	+1.6 [−1.0, 4.6]	+0.9 [−0.05, 1.9]
	+DDS + SPRD-C	+2.3 [1.4, 3.4]	+3.0 [1.7, 4.3]	+2.4 [−0.3, 5.3]	+1.9 [0.8, 3.0]
Gemma-4-31B	+DDS	+1.8 [0.8, 2.9]	+1.5 [0.5, 2.6]	+1.4 [−0.3, 3.1]	+0.1 [−1.0, 1.1]
	+SPRD-C	+1.9 [1.0, 2.9]	+0.9 [−0.3, 2.2]	−0.2 [−1.9, 1.4]	−0.1 [−1.0, 0.9]
	+DDS + SPRD-C	+2.5 [1.5, 3.6]	+1.3 [−0.2, 2.7]	+0.7 [−0.7, 2.2]	+0.6 [−0.4, 1.5]
Mean across bases	+DDS	+2.2	+1.9 [1.1, 2.6]	+1.4 [0.4, 2.4]	+1.1 [0.4, 1.8]
	+SPRD-C	+2.3	+1.8 [0.8, 2.7]	+1.1 [−0.1, 2.5]	+1.0 [0.4, 1.6]
	+DDS + SPRD-C	+2.9	+2.3 [1.3, 3.2]	+1.8 [0.6, 3.1]	+1.5 [0.9, 2.2]
Both vs. validation-selected baseline, mean (post hoc)		—	+2.1 [1.2, 3.0]	+0.8 [−0.1, 1.7]	+0.8 [0.2, 1.4]

DecisionVista-Synth, while only SPRD-C does so on typed-decisions (Appendix Table 34).

On the transfer suite and DV-Dev-A/B, the combination gains +3.9, +2.3 and +2.5 across the three bases, with all 27 per-seed differences positive; against the one-pass baseline the gains are +4.0, +1.6 and +1.4. These sets informed design. The pooled test gains are 52–62% of the validation gain (+2.9). Primary intervals remain positive under seed re-sampling, alternative novelty units, the author exclusion and inclusion of later seeds; dropping any one human-labelled source retains a positive mean gain (Appendix V).

Comparison with released decision models. Table 3 separates model bases and discloses benchmark training. All models we run receive identical inputs: our LLMs, Nimble-9B and OpenJev use Nimble’s scorer; Laya and Kev use their released packages, including Kev’s shipped temperature.

Without training on benchmark sources, our 9B recipe scores 76.2 (± 0.6 over seeds) on Nimble’s 13-subset benchmark versus Nimble-9B’s 74.8 on the same base. The lead is concentrated in Civil Comments and MultiNLI (per-subset scores and seed deviations in Appendix Table 12). The 9B recipe also improves on Nimble on the human-labelled transfer suite (71.9 vs. 67.5) and DecisionVista-Human (74.9 vs. 72.5); the 27B recipe scores 77.0 on DecisionVista-Human versus OpenJev’s 76.5. On typed-decisions, which no design decision used, the recipe gains +2.3 [1.3, 3.2] over the step-matched baseline across bases; Kev-27B remains stronger among 27B readouts, while Laya-TD uses this benchmark’s training split.

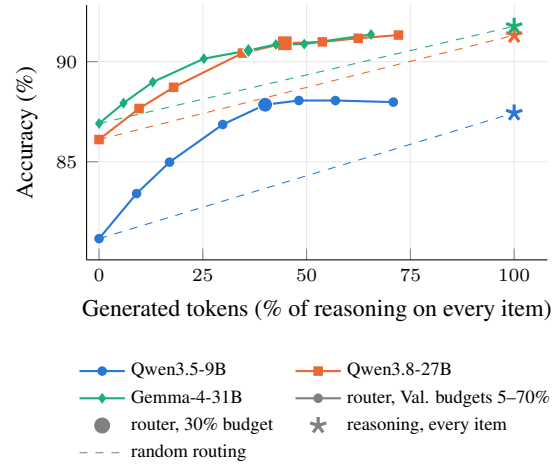


Figure 2. Confidence routing: accuracy (mean of the transfer suite, DV-Dev-A and DV-Dev-B; seed means) against generated tokens as a share of reasoning on every item. Router thresholds are fixed on the validation suite at routing budgets of 5–70%; large marker: 30%. Dashed: random routing, expected accuracy at the same generated-token share. Star: reasoning on every item.

Public-benchmark rankings depend on source exposure: adding the benchmark’s training splits gives DDS 80.8 at 9B and 81.7 at 27B on Nimble’s benchmark, with transfer-suite scores of 71.9 and 77.6. These source-trained runs belong to a different comparison regime; Appendix Table 13 reports them alongside the released systems and their training-data disclosures. The larger source-free recipes do not lead Nimble’s benchmark (Table 3).

Large Decision Models

Table 3. Decision accuracy (%; three-seed means for our tuned models; step-matched baselines). Bold: best per block, excluding reasoning references (°). Shaded: our recipe. *Confidence router with the same setting on every base: the 30% least confident items of the validation suite define the threshold, and routed items are answered by the untuned base using reasoning rollout (one sample); generated-token costs are in §6. °Untuned base using reasoning rollout on every item. ¶Trained on BoolQ and MultiNLI; ‡trained on typed-decisions (TD); §published results, not rerun. Benchmark and scoring details: §3 and §6.

Model	Public benchmarks (test)			Ours: development / sealed test	
	Nimble-13 Nimble; 13 tasks	Kev-Transfer Kev; new sources	Typed-Decisions Laya/Jev; workflows	Transfer suite dev.; 8 MC sources	DV-Human sealed; 8 sources
<i>Released decision models</i>					
Jev 1.13.0 [§]	76.0	–	72.7	–	–
Laya-TD [§]	–	–	76.6	–	–
Laya-TD, 0.4B [‡]	59.8	67.1	76.2	35.5	51.6
Laya-en (0.4B)	57.2	66.5	37.4	35.0	51.0
Kev-4B [¶]	74.3	83.8	66.7	65.6	69.2
<i>Qwen3.5-9B base</i>					
Untuned	73.7	77.0	63.5	66.2	69.4
Reasoning rollout [°]	76.1	–	72.2	75.0	76.3
Nimble-9B	74.8	82.5	65.0	67.5	72.5
Kev-9B [¶]	75.3	85.2	71.4	65.9	72.6
Baseline	74.1	84.9	69.8	69.7	72.8
DDS + SPRD-C	76.2	85.8	72.3	71.9	74.9
+ router*	76.7	–	73.2	76.2	76.4
<i>Qwen3.8-27B base</i>					
Untuned	74.4	85.5	72.0	75.4	73.2
Reasoning rollout [°]	72.8	–	75.2	80.6	76.0
OpenJev-27B	77.7	88.3	73.0	76.5	76.5
Kev-27B [¶]	77.8	89.5	75.9	73.5	75.5
Baseline	75.2	90.3	71.7	76.2	75.2
DDS + SPRD-C	75.7	90.2	74.7	77.5	77.0
+ router*	74.0	–	76.4	80.6	76.4
<i>Gemma-4-31B base</i>					
Untuned	77.8	88.6	75.0	77.8	79.2
Reasoning rollout [°]	77.0	–	74.8	82.2	80.9
Baseline	75.3	90.3	74.1	77.6	79.5
DDS + SPRD-C	76.3	90.2	75.4	79.0	80.2
+ router*	76.9	–	75.6	81.7	80.6

Table 4. Gemma-4-31B serving efficiency. Speedup uses median latency relative to reasoning; GPU-hours per million requests are extrapolated.

Mode	Median (ms)	Rate (req./s)	Speedup (×)	GPU-h (/1M)
Reasoning rollout	7,380	0.787	1.0	706.2
Tuned readout	40.6	44.2	181.8	12.6

Measured serving efficiency. On Gemma-4-31B served in BF16 on two H100 GPUs, tuned readout takes 40.6 ms median latency versus 7.38 s for reasoning rollout, with $56.2\times$ higher throughput (Table 4). These measurements exclude confidence routing; checkpoint and measurement details are in Appendix X.

Deployment through confidence routing. Table 3 includes each untuned base using reasoning rollout with one sample as a reference; median generation costs range from 357 to 2,124 tokens per decision across models. With the threshold set to route 30% of the validation suite, the router recovers 66–140% (median 92%) of the rollout–readout gap on the transfer suite and DV-Dev-A/B while generating 27–58% of reasoning’s tokens. Random routing of the same fraction of items recovers 20–35% (Fig. 2; Appendix Q). On the transfer suite it matches reasoning (+1.2, 0.0, −0.5 points). The test-set routing analyses are *post hoc* and descriptive: DecisionVista-Human needs 52–80% of reasoning’s tokens, while DecisionVista-Synth gains 15–18 points over the readout but remains 2–7 below reasoning. We count generated tokens only; routed items also require a second prefill, so end-to-end savings are smaller.

7. Analysis

We examine what the supervision labels contribute, how well readout probabilities calibrate, and how the rollout-readout gap varies across bases.

What the labels contribute. SPRD-C adds new items, a filter and reasoning labels at once. Holding the 10,331 filtered items and training steps fixed, replacing reasoning targets with the model’s own single-pass labels costs 27B 1.3–2.4 points on the validation suite and DV-Dev-A/B, and 1.2–1.9 on DecisionVista-Human and typed-decisions (Appendix Tables 33, 36). The advantage over self-training therefore comes from the labels, conditional on the reasoning-intent filter that selected these items. Hard and soft reasoning targets perform similarly.

Teacher identity and selection. Other capable labels (gpt-5.5 or the generator’s intended answer) perform similarly on DecisionVista-Synth and DecisionVista-Human. At 9B, SPRD-C exceeds an equal-sized random pool with generator labels by only +0.7 to +1.2 points at matched steps, with intervals touching zero. The clearest advantage is over readout self-training; at 27B the model’s own reasoning supplies the stronger labels without an additional labeller. A post-hoc Gemma control also gives comparable results with its own reasoning labels in place of Qwen’s (Appendix S). This is not a fixed-subset comparison: relabelling the same pool changes the retained items and adds about 2% more training steps. It does not support teacher family as the sole explanation for Gemma’s smaller gains.

Calibration. Longer baseline training makes Qwen models overconfident (transfer-suite ECE 23.8 at 9B, 18.5 at 27B), while the recipes stay at 8–12. One temperature fitted on the validation suite removes most of this gap and transfers to the test sets (Appendix Tables 30, 31). After scaling, every recipe retains a lower Brier score than the step-matched baseline on every base, and a lower NLL on both test sets; transfer-suite NLL at 27B is tied (0.71–0.73). Kev’s shipped temperature gives the lowest raw ECE but higher NLL and Brier than our Qwen recipes. Gemma’s raw readout remains strongly overconfident across training variants and requires calibration.

Mixture-of-experts bases. The smaller tested MoE bases leave larger rollout-readout gaps than the dense 27B model (Appendix R). Qwen3.5-122B-A10B scores 3.2 points higher using reasoning rollout on the validation suite but only ties its readout (79.8 vs. 79.3); 35B-A3B matches its reasoning while trailing its readout by two points. The largest tested MoE, Qwen3.5-397B-A17B, instead has the same transfer-suite gap as the dense base (5.2 points) and a 2.6–2.8-point readout advantage. On 35B-A3B, widen-

ing expert routing at inference or during tuning does not close the rollout-readout gap. One possible explanation is that generating multiple reasoning tokens engages different experts before the answer, but these comparisons do not isolate that mechanism or establish a general scaling law.

8. Related work

Decision interfaces and transfer. Label-aware classifiers such as TARS support unseen classes (Halder et al., 2020); class decisions can also be read from next-token scores (Schick & Schütze, 2021; Zhao et al., 2021). Decision services and models (TypeSafe, 2026; OpenJev authors, 2026; Bespoke Labs, 2026; Convai Innovations, 2026; Palmer, 2026) standardise the typed interface we name and evaluate. DDS applies instruction tuning and synthetic task construction (Wei et al., 2022a; Sanh et al., 2022; Wang et al., 2022; 2023b) to single-pass schema transfer.

Reasoning supervision. SPRD is closest to System-2 distillation (Yu et al., 2024) and implicit chain-of-thought (Deng et al., 2023; 2024), but targets final-answer distributions at one readout position over run-time label sets. Like on-policy (self-)distillation (Agarwal et al., 2024; Zhao et al., 2026), it learns from stronger teacher behaviour; supervision here is offline and uses no student-generated trajectories. Readout self-training (Lee, 2013; Xie et al., 2020) supplies a label-source control.

Selective reasoning. The router combines selective prediction (Geifman & El-Yaniv, 2017) with a same-base reasoning fallback, following the acceptance-and-escalation structure of LLM cascades (Chen et al., 2023; Ong et al., 2024). Its margin threshold is fixed on the validation suite; probability calibration separately uses temperature scaling (Guo et al., 2017). Appendix W gives the fuller comparisons.

9. Conclusion

Large Decision Models frame efficient decision making as transfer across schemas supplied at run time. DDS broadens schema supervision, while SPRD trains the readout toward reasoning-derived decisions; DecisionVista tests their combined effect after method development. The pre-registered transfer gains are concentrated in Qwen; a post-hoc comparison with validation-selected baseline yields smaller gains, with a remaining advantage on human-labelled DecisionVista-Human. Synthetic-schema labels measure model agreement rather than human judgement (Appendix A). Confidence routing provides a practical deployment pattern when a rollout-readout gap remains: accept confident readout answers and invoke reasoning rollout for uncertain cases.

Impact Statement

Large Decision Models are used for moderation, triage, routing and eligibility-style decisions, where errors affect people. Our work makes such models more accurate on decisions they were not trained for and studies their calibration, which can support abstention and human review; it does not make them suitable for consequential decisions without oversight. Several of our evaluation sets are written and labelled by language models, whose shared conventions can differ from human judgement; we therefore report a human-labelled test set, release all model-labelled items with their blind labels, and specify a human audit protocol. The confidence router lowers the compute of reasoning-based decisions. All data are public or synthetic and contain no personal information beyond what the public datasets already include.

References

- Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., and Bachem, O. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2306.13649>.
- Bespoke Labs. Nimble: open single-pass decision models. Model release and public benchmark, 2026.
- Chen, L., Zaharia, M., and Zou, J. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- Convai Innovations. Laya: A fast, non-autoregressive decision engine with calibrated probabilities. <https://huggingface.co/convaiinnovations/laya>, Apache-2.0, 2026.
- Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., and Shieber, S. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.
- Deng, Y., Choi, Y., and Shieber, S. From explicit CoT to implicit CoT: Learning to internalize CoT step by step. *arXiv preprint arXiv:2405.14838*, 2024.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL <https://aclanthology.org/N19-1423/>.
- Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- Halder, K., Akbik, A., Krapac, J., and Vollgraf, R. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3202–3213, 2020. URL <https://aclanthology.org/2020.coling-main.285/>.
- Ho, N., Schmid, L., and Yun, S.-Y. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
- Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
- Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. 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.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Huang, J., Gu, S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 2022.
- Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.
- Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., and Roberts, A. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

- Magister, L. C., Mallinson, J., Adamek, J., Malmi, E., and Severyn, A. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.
- Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024.
- OpenJev authors. OpenJev: an open-weights system-one decision model. Model release, 2026.
- Palmer, J. Kev: Calibrated typed decision models. <https://github.com/jaredpalmer/kev>, 2026.
- Sanh, V., Webson, A., Raffel, C., Bach, S. H., et al. Multi-task prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.
- Schick, T. and Schütze, H. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.
- TypeSafe. Jev: typed system-one decisions as a service. Hosted API, version 1.13.0, 2026.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023a.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., et al. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023b.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022b.
- Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- Yu, P., Xu, J., Weston, J., and Kulikov, I. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.
- Zelikman, E., Wu, Y., Mu, J., and Goodman, N. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.
- Zhao, S., Xie, Z., Liu, M., Huang, J., Pang, G., Chen, F., and Grover, A. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. URL <https://arxiv.org/abs/2601.18734>.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023.

Large Decision Models

Table 5. Supervised recipes run before SPRD. None improves new-schema accuracy (DV-Dev-A, DV-Dev-B) of the 27B base beyond seed noise; post-hoc update scaling recovers some. The eight-subset diagnostic averages the choice and yes/no subsets of Nimble, excluding Civil Comments and MultiNLI.

Base	Recipe	Change	8 subsets	Val.	DV-Dev-A	DV-Dev-B
27B	untuned	–	82.5	79.3	89.1	87.5
27B	Baseline	20k public-dataset records + Nimble data	82.6	81.8	88.1	87.5
27B	Baseline, full fine-tune	all weights, ZeRO-3	80.6	78.1	87.1	–
27B	Baseline + source splits	+ 21k benchmark-source train records	89.1	82.6	87.8	88.0
27B	Baseline+adversarial pairs	+ 6.5k LLM minimal pairs, flip margin	83.7	82.0	89.1	88.3
27B	Baseline, update $\times 0.5$	post-hoc scaling	82.8	82.4	89.9	87.5
27B	Baseline + source splits, update $\times 0.5$	post-hoc scaling	88.0	82.0	89.7	89.2
9B	untuned	–	77.6	71.7	79.9	74.8
9B	Nimble data only	Nimble’s recipe	80.8	72.9	–	–
9B	Baseline (3 seeds)		80.4–81.3	75.6–77.3	84.3–85.5	79.1
9B	+ permutation consistency	JS between two option orders	80.9	77.9	86.5	–
9B	+ ordinal loss	ranked-probability score on scores	80.3	75.8	83.5	–
9B	+ self-supervised decisions	12k denoising decisions	81.3	76.5	84.5	–
9B	+ distillation from OpenJev	KL to OpenJev readout	81.8	76.8	86.9	–
9B	+ synthetic task-matched	4k LLM items for suite tasks	81.1	77.5	87.6	–
9B	+ executable curriculum	12k program-labelled items	79.0	76.3	85.2	79.9
9B	+ retention KL + adversarial pairs		80.9	77.1	86.6	81.9
9B	rank 64, 3 epochs	capacity	80.0	75.4	84.5	–
9B	Baseline + source splits		88.2	76.0	84.7	–
OpenJev-27B (released)			84.1	84.1	88.6	86.1
Nimble-9B (released)			81.6	73.0	82.7	79.7

A. Limitations

Model-labelled evaluation. Our new-schema sets are labelled by agreement between models, not by humans. Part of SPRD’s gain on the earlier sets sits on items a third model family disputes (§7); DV-Pilot and DecisionVista-Synth therefore require multi-family agreement, but agreement is not correctness. The human-labelled DecisionVista-Human confirms the direction of the effect (+1.5), but not that DecisionVista-Synth’s labels are correct. We release a blinded human audit protocol whose pre-registered sample (items, sampling seed, sensitivity analysis) was fixed before the DecisionVista-Synth read; the audit is not part of this paper. **Comparator and post-hoc analyses.** The pre-registered comparator, baseline trained for the full recipe’s number of steps, is overtrained; against a baseline model selected on the validation suite the DecisionVista-Synth gain is not significant (+0.8 [−0.1, 1.7]). The validation-selected, one-pass, label-source, permutation and in-distribution comparisons were run after the sealed reads and are descriptive. **Novelty is screened, not guaranteed.** DecisionVista-Synth families were screened against all training schemas and earlier sets by model audit and embedding similarity; they may still resemble pretraining data. **Author-family confounding.** Each DecisionVista-Synth family has one author, so author-level results mix author style with family difficulty. **Scope.** We study English decisions with up to eight options, training contexts under 2k tokens, LoRA adapters, and three base models up to 31B that differ in family, size and post-training, so differences between them are not a controlled scaling study; we do not claim the results for full fine-tuning, long contexts, multilingual inputs or larger models. Gemma gains least; relabelling its SPRD-C pool with its own reasoning gives equivalent results (Appendix S), so teacher mismatch does not explain this, and the strength of its untuned readout is the likelier cause. The self-posed items are written by Qwen3.8-27B for every base. **Calibration of Gemma.** All Gemma variants, including step-matched baseline, remain overconfident (transfer-suite ECE 21–22); its fitted temperatures are 10–12 in every variant, and our recipe does not fix this; Gemma LDMs need post-hoc temperature scaling. **Router cost.** Router savings are in generated tokens; routed items also need the reasoning model’s prefill and a second model in serving, and we did not measure end-to-end routing latency or throughput. **Compute.** The SPRD teacher pass costs about 8 H100-hours per 16k items; seeds beyond three were not affordable for the 27B and 31B variants.

B. Supervised recipes behind the diagnosis

Table 5 lists every supervised recipe we ran before SPRD, with one seed each unless marked. All use the readout objective (cross-entropy over option letters with per-epoch option shuffling) and LoRA rank 16 unless marked.

Halving the learned baseline update after training (Wortsman et al., 2022) recovers 1.8 points on DV-Dev-A (88.1 to 89.9), a diagnostic motivation for reporting update-scaled baseline comparisons.

C. Statistics

Recipe comparisons (Tables 2, 16). For base b , recipe r and training seed s , let $A_{b,r,s}$ be an accuracy statistic. The effect of r at base b is $\Delta_b = \frac{1}{|S_b|} \sum_{s \in S_b} (A_{b,r,s} - A_{b,\text{baseline},s})$: each seed of the recipe is paired with the baseline variant of the same seed. The pooled effect is the unweighted mean over bases, $\Delta = \frac{1}{3} \sum_b \Delta_b$. Seeds: $S_{9B} = \{17, 29, 41\}$ everywhere; $S_{27B} = S_{31B} = \{17, 29\}$ for the test sets (as pre-registered) and $\{17, 29, 41\}$ for validation sets.

Statistics per set. Validation: A is the unweighted mean of the pooled item accuracies on DV-Dev-A, DV-Dev-B and the transfer suite. DV-Pilot and DecisionVista-Synth: family-macro accuracy (families weighted equally). DecisionVista-Human: source-macro accuracy (eight sources weighted equally).

Intervals. We resample the unit that defines novelty, with the same resample indices for every variant, seed and base, so that differences are paired: task families for the synthetic DecisionVista sets (contrastive groups stay intact), families and held-out source subjects for the exploratory statistic (each of the three sets resampled separately), and items within each source for DecisionVista-Human, whose eight sources are fixed. 10,000 resamples; percentile intervals; two-sided bootstrap p -values. For DecisionVista-Synth we also report the pre-registered sign-flip test over family-level differences (100,000 flips). The intervals condition on the trained models; seed variation is shown separately by the per-seed effects (all 27 exploratory per-seed differences are positive; DecisionVista-Synth per-seed effects at 9B are +2.7, +4.7 and -0.6 , at 27B +2.1 and +2.6, and at 31B +0.6 and +0.8).

Multiplicity. Only the two primary contrasts are confirmatory. On DecisionVista-Synth the per-scale tests are Holm-corrected and gatekept by the primary, followed by the ordered component tests; all other cells are descriptive.

Earlier comparisons. Analyses of the SPRD-only experiments (DV-Dev-A–DV-Pilot, Table 33 and §7) use a paired bootstrap over generation batches or task families with 5,000–20,000 resamples, as reported with each result.

D. Protocol chronology

Table 6 lists every decision that touched a test set. The time-stamped planning log in which each entry was written before the corresponding read, including the DV-Pilot file hash, is part of the released artifacts. DV-Dev-A, DV-Dev-B and DV-Stress are validation evidence; DV-Pilot is confirmatory for SPRD-C and DecisionVista-Synth for the combined recipe.

E. Cost

The teacher pass over the 16k-item pool generated about 59M tokens (a median of 2,051 tokens per item for four samples) and took about 8 GPU-hours on H100s with vLLM in FP8; schema and item generation took under 2 GPU-hours. In the earlier SPRD study the 14.8k pool records made training about $1.7\times$ as long as baseline training (4.6 vs. 2.8 hours for 27B on one H100); in the controlled study the full recipe and step-matched baseline train the same budget (4,031 and 4,032 steps), and the single-component variants fewer. Inference is unchanged: one prefill and no generated tokens, against a median of 335–513 generated tokens per sample for the reasoning rollout that the student imitates.

F. Additional analysis tables

G. All scores

H. Training configurations

Everything in Tables 3, 2, 16 and 9 comes from the variants below; the earlier recipe study (Appendix B) and the 4B DDS ablations used other settings and are reported separately.

Table 6. Order of decisions relative to test-set reads.

Step	Event
1	DV-Dev-A and DV-Dev-B built and used for earlier recipes (Appendix B); DV-Stress built and sealed.
2	Pool generated; DV-Stress family descriptions used only in the exclusion list.
3	λ , τ chosen on the validation suite and DV-Dev-A (9B); SPRD variants and DV-Stress checkpoints frozen and recorded.
4	DV-Stress read once (+1.0 over baseline).
5	Third-family (Magistral) labels for DV-Dev-A, DV-Dev-B and DV-Stress; gains concentrate on contested items.
6	Step-matched baseline, SPRD-C and other variants trained; DV-Pilot families proposed and audited, protocol recorded.
7	DV-Pilot items generated and sealed after three-way agreement (file hash recorded).
8	Variant chosen on the validation suite, DV-Dev-A, DV-Dev-B and validated subsets (SPRD-C) and recorded.
9	DV-Pilot read once for the SPRD rows.
10	DDS mixture chosen after a 4B screen scored on the validation suite, DV-Dev-A, DV-Dev-B, the eight-subset diagnostic and the transfer suite; models trained (seeds 17/29 at every base, 41 at 9B; never read on DV-Stress, DV-Pilot or DV-Synth).
11	DV-Synth designed with advisors; families built by decision operator from four authors (three not used in any training data); pre-registration written (primary contrast DDS+SPRD-C vs. step-matched baseline), then amended to place the human audit at the end.
12	DV-Synth items generated and labelled; hashes recorded.
13	Seed 41 at 27B and 31B trained (after the pre-registration, before the read); variant manifest (adapter hashes) and analysis code frozen; DV-Synth read once for all variants and references.
14	After internal review, DV-Human built from datasets no variant or design decision had used, hashed and pre-registered, then read once for all variants.
15	Post-hoc, descriptive: one-pass baselines at 27B and 31B; Laya references.
16	Post-hoc amendment for uniform labels: DV-Dev-A, DV-Dev-B and DV-Stress relabelled blind by GLM-5.2, DeepSeek-V4 and Kimi-K3 with DV-Synth’s prompt; items kept iff at least two of three agree with the author (2,311 of 2,374, 97.3%); all variants re-scored from stored predictions. The factorial’s development contrasts and score tables use these kept items; the diagnosis, recipe-study and routing analyses report the full sets.

Table 7. Rollout–readout gap of untuned Qwen3.8-27B. Reasoning = majority of four thinking-mode samples; readout = single-pass letter readout.

Validation source	Readout	Reasoning	Gap
WinoGrande	67.5	93.5	+26.0
MedMCQA	65.0	72.5	+7.5
TweetEval irony	69.5	76.0	+6.5
SMS spam	97.5	98.5	+1.0
ARC-Challenge	96.0	96.5	+0.5
CaseHOLD	78.5	78.0	−0.5
CommitmentBank	85.7	69.6	−16.1
Val. (all, 1,256)	79.3	85.1	+5.8
DV-Dev-A (613)	89.1	96.9	+7.8
DV-Dev-B (592)	87.5	96.8	+9.3

Experiment summary. Table 11 maps every trained configuration that appears in a table to its data, budget, seeds and analysis status. “Pre-reg.” rows are covered by the DecisionVista-Synth and DecisionVista-Human pre-registrations; all other comparisons on the test sets are post hoc.

Large Decision Models

Table 8. Agreement (%) of other label sources with the reasoning-teacher majority on the self-posed pool, by the teacher’s vote share (15,980 of the 16,040 items have at least one parseable teacher answer).

Vote share	Items	Readout	gpt-5.5	Gen. intent
1.0	13,191	85.9	93.9	73.9
[0.75, 1)	1,638	54.8	64.0	43.2
< 0.75	1,151	47.3	51.9	40.0
All	15,980	79.9	87.8	68.3

Table 9. Accuracy (%) and ECE (%), pooled over Val., DV-Dev-A, DV-Dev-B and the transfer suite; lower is better); seed means. Best per column within each block in bold; our recipe shaded. The eight-subset diagnostic averages Nimble’s choice and yes/no subsets excluding Civil Comments and MultiNLI. It, Nimble-13 (the full benchmark), Val. and the transfer suite are human-labelled; synthetic DecisionVista sets are new decision schemas labelled by model agreement. No tuned row trains on any benchmark source. Val., the transfer suite, DV-Dev-A and DV-Dev-B informed design or selection decisions (validation evidence); DV-Pilot, DecisionVista-Synth (family-macro) and DecisionVista-Human (DV-Human, source-macro; human-labelled) were sealed and read once for all rows after the variants were frozen; before that final read, no DDS variant had been scored on DV-Pilot. [‡]Fine-tuned by its authors on the training split of their typed-decisions benchmark.

Model	Seeds	Public benchmarks		Validation				Test				ECE↓
		8 subsets	Nimble-13	Val.	Transfer	DV-Dev-A	DV-Dev-B	DV-Pilot	DV-Synth	DV-Human		
<i>Released decision encoders</i>												
Laya-en (0.4B)	1	61.3	57.2	44.3	35.0	54.6	44.1	54.2	33.3	–	17.8	
Laya-multilingual (0.3B)	1	63.3	57.9	40.0	30.8	51.4	38.0	45.0	30.8	–	30.5	
Laya-typed-decisions [‡] (0.4B)	1	61.1	59.8	48.3	35.5	55.6	46.1	57.9	33.8	–	7.3	
<i>Qwen3.5-9B</i>												
Untuned	1	77.3	73.7	71.7	66.2	80.5	75.6	90.8	52.5	69.4	14.0	
Nimble-9B	1	81.7	74.8	72.7	67.5	83.5	82.2	93.2	63.6	72.5	12.4	
Baseline, step-matched	3	80.0	74.1	75.9	69.7	84.1	79.3	91.8	64.5	72.8	20.3	
+ DDS	3	80.8	71.6	77.1	71.6	87.2	83.8	96.6	66.1	75.0	7.5	
+ SPRD-C	3	81.2	76.1	77.4	71.1	87.6	84.0	96.1	66.5	74.9	9.4	
+ DDS + SPRD-C	3	80.7	76.2	77.1	71.9	87.1	85.8	96.8	66.8	74.9	7.8	
<i>Qwen3.8-27B</i>												
Untuned	1	82.5	74.4	79.3	75.4	89.2	88.8	97.6	68.3	73.2	3.1	
OpenJev-27B	1	83.9	77.7	83.0	76.5	90.0	89.1	98.3	75.3	76.5	4.5	
Baseline, step-matched	3	82.0	75.2	81.6	76.2	89.4	87.1	97.1	76.1	75.2	15.0	
+ DDS	3	82.3	74.9	83.3	77.4	90.5	89.6	98.5	77.7	76.3	5.7	
+ SPRD-C	3	83.1	76.3	82.9	76.9	91.2	90.3	98.9	77.9	76.7	8.2	
+ DDS + SPRD-C	3	83.0	75.7	83.7	77.5	91.1	91.1	98.7	78.2	77.0	6.5	
<i>Gemma-4-31B</i>												
Untuned	1	84.3	77.8	82.3	77.8	90.8	90.3	98.9	77.2	79.2	18.3	
Baseline, step-matched	3	82.1	75.3	82.4	77.6	89.5	87.3	97.9	79.9	79.5	18.8	
+ DDS	3	82.6	75.6	83.6	78.9	91.1	89.8	99.3	81.0	79.5	17.2	
+ SPRD-C	3	83.0	76.6	83.2	78.0	91.4	90.6	99.3	79.6	79.4	17.7	
+ DDS + SPRD-C	3	83.4	76.3	84.1	79.0	91.3	91.6	99.4	80.8	80.2	16.9	

I. Benchmark-source training comparisons

Adding the public benchmark’s training splits makes the Nimble benchmark in-distribution for our source-augmented runs. Table 13 reports this separate comparison regime, including source exposure and trainer differences. With the controlled study trainer, DDS plus source splits reaches 80.8 over the full 13-subset benchmark at 9B and 81.7 at 27B, compared with Nimble-9B’s 74.8 and OpenJev’s 77.7. Adding SPRD-C does not improve this benchmark further.

J. The DDS corpus

Ablations of the corpus at 4B (single seed unless noted; step-matched 44k baseline) are in Table 15: synthesised decision tasks provide most of the gain; additional public data and Super-NaturalInstructions add little on new schemas but help the knowledge-heavy transfer suite; label cleaning is within noise; removing any single field (medicine, law, science, preference,

Table 10. Configuration of every variant in the controlled study.

Item	Setting
Bases	Qwen3.5-9B; Qwen3.8-27B (rev. 1d4bf0f2); Gemma-4-31B-it
Adapter	LoRA rank 16, $\alpha=32$, dropout 0.05, on every linear layer of the language model; bf16; final step, no checkpoint selection
Optimiser	AdamW, $\text{lr } 5 \times 10^{-5}$, linear schedule with 5% warm-up, gradient clipping 1.0, effective batch 16 (9B: 4×4 accumulation; 27B/31B: 2×8), max length 2,048
Objective	Eq. 3: cross-entropy on baseline and DDS records; $\lambda=3$ KL to SPRD-C vote targets; option order re-shuffled per draw
Data	baseline = 20k public records + Nimble’s 2.7k training items (22.4k); DDS = half of the 63.4k mixture (31.7k, fixed sample); SPRD-C = 10,331 items labelled by Qwen3.8-27B for every base
Steps	baseline, step-matched: 4,032 (2.9 passes); +DDS: 3,386; +SPRD-C: 2,050; +both: 4,031 (one pass each); one-pass baseline: 1,405
Seeds	17, 29, 41 (seed 41 at 27B/31B trained after the DV-Synth pre-registration)
Readout	Eq. 2 with Nimble’s scorer, canonical option order at test time, argmax, max input 8,192 tokens
Manifest	adapter SHA-256 of all 36 variants, analysis scripts and set hashes are frozen in the released MANIFEST file

Table 11. Experiment summary. Steps at effective batch 16. Seeds: 9B / 27B / 31B. Status: *pre-reg.*, primary or secondary contrast fixed before the sealed reads; *descr.*, descriptive; *post hoc*, specified after the sealed reads.

Configuration	Training data	Steps	Seeds	Status	Where
Baseline, step-matched	baseline 22.4k	4,032	3 / 3 / 3	pre-reg. comparator	Tables 3, 2, 9
+DDS + SPRD-C	baseline + DDS 31.7k + SPRD-C 10.3k	4,031	3 / 3 / 3	pre-reg. primary	Tables 3, 2, 9
+DDS	baseline + DDS	3,386	3 / 3 / 3	pre-reg. secondary; descr. vs. baseline	Tables 2, 9
+SPRD-C	baseline + SPRD-C	2,050	3 / 3 / 3	pre-reg. secondary; descr. vs. baseline	Tables 2, 9
Baseline, one pass	baseline	1,405	2 / 2 / 2	descr.	Table 9
Validation-selected baseline and recipe	as above, update scale 1 or 0.65	–	3 / 2 / 2	post hoc	Tables 2, 35
+DDS or +SPRD-C, 4,031 steps	as the individual components	4,031	3 / 2 / –	post hoc	Table 34
Pool with generator labels	baseline + 10,331 random pool items	2,050	3 / – / –	post hoc	§7
Label-source controls	baseline + 14.8k pool, per label source	1.66 passes	– / 2 / –	post hoc (earlier study)	Table 36
+ benchmark source splits	recipe + public-suite training splits	–	1 / 2 / –	descr.	Table 13
Confidence router	no training; validation threshold	–	as recipe	post hoc on test sets	Fig. 2, Table 20

math/logic) changes neither the average nor that field’s own held-out benchmark beyond noise.

K. Decision Vista-Synth construction

Operators and authors. Eighteen decision operators, each realised by two task families written by two different authors: eligibility with exceptions, evidence sufficiency, priority scheduling, capacity allocation, Pareto choice, sequential stopping, matching, diagnosis, counterfactual intervention, risk-sensitive choice, committee aggregation, state tracking, numeric and date reasoning, version change, rubric grading, intent and pragmatics, reference resolution, and source consistency. Authors were gpt-6-astra (10 families), GLM-5.2 (7), DeepSeek-V4 (9) and Kimi-K3 (9); two GLM slots were reassigned after five proposal rounds produced no candidate passing the checks below, and one reassigned slot (intent and pragmatics) still produced none after a sixth round, leaving 35 families. Author and family are confounded by design.

Screening. Candidate families were de-duplicated, retained only if both non-author models confirmed no overlap with every earlier evaluation family, the 696 SPRD pool schemas and the 1,985 DDS synthesised task families, checked by the same models for schema validity (exclusive and exhaustive labels, strictly ordered levels, no numeric gaps, at least two interacting constraints), and compared by embedding similarity with all 5,930 DDS schema rows, the SPRD schemas and earlier families.

Items and labels. Each author wrote 45 candidate items per family in contrastive groups of three; the three non-author models labelled every item blind with a rationale, and an item was kept when at least two agreed with the author (1,505 of 1,546 candidates; 1,472 with all three blind labels agreeing, 33 with two of three; every family has at least 15 items). No item was filtered on the performance of any evaluated model, and all candidates, blind labels and rationales are released. The pre-registration, its amendment and deviations, and the file hashes are released with the set.

Example: one contrastive group. Family *ferry dog travel permission* (operator: eligibility with exceptions; author gpt-6-astra; all three blind labellers agree with the author on every item shown). The three items share the dog record, the operator policy and the question, and differ only in the facts in the last column.

Large Decision Models

Table 12. Nimble’s 13-subset public benchmark (accuracy, %; three-seed means for tuned rows, seed standard deviation shown when ≥ 3). No tuned row trains on any of these sources. HelpSteer2 and SummEval are 5-level ratings; SummEval-consistency (144 items) is unstable across seeds of the same recipe (13–84), which dominates the seed variance of the 13-subset average. Avg-13 is the full-benchmark score. Avg-11 is an exploratory sensitivity average excluding Civil Comments (89% majority class) and MultiNLI, retained to show dependence on these two subsets. Jev 1.13.0’s published 13-subset average is 76.0.

Model	VitC	MAS-en	MAS-de	PubMed	BoolQ	SQuAD2	PAWS	Aegis2	MNLI	CivilC	HelpS2	SE-rel	SE-con	Avg-13	Avg-11
<i>Qwen3.5-9B</i>															
Untuned	75.3	84.0	78.3	65.6	80.3	84.0	75.2	76.8	86.0	80.0	38.5	51.7	82.6	73.7	72.0
Nimble-9B	76.6	86.9	83.7	74.8	86.0	80.9	83.2	81.2	86.0	70.3	38.5	49.6	75.0	74.8	74.2
Baseline, step-m.	67.9	86.6	84.3	74.4	86.9	82.7	75.7	81.2	90.3	86.9	35.5	52.2	58.1 ± 30.6	74.1	71.4
+DDS	72.5	87.8	84.4	77.2	86.8	78.6	76.8	82.1	88.2	89.9	36.4	36.4 ± 8.8	34.3 ± 26.4	71.6	68.5
+SPRD-C	74.2	86.8	83.0	75.6	86.6	79.7	81.5	82.3	90.2	88.9	36.9	47.9 ± 3.3	75.9 ± 12.2	76.1	73.7
+both	70.6	86.9	82.6	74.8	86.9	80.7	81.3	81.6	88.2	89.6	38.6	46.9 ± 3.7	81.9 ± 3.0	76.2	73.9
<i>Qwen3.8-27B</i>															
Untuned	75.0	85.7	85.4	71.2	86.3	87.6	85.2	83.2	84.6	76.3	38.5	21.2	86.8	74.4	73.3
OpenJev-27B	80.5	87.7	86.9	81.2	88.3	83.6	82.4	80.8	89.0	86.7	43.0	35.0	85.4	77.7	75.9
Baseline, step-m.	72.4	86.7	87.5	77.7	88.3	81.3	79.2	82.5	92.1	87.9	38.4	22.9 ± 4.6	80.3	75.2	72.5
+DDS	76.6	88.0	87.9	77.5	88.9	79.2	77.2	83.3	91.3	90.2	36.4	13.5 ± 8.9	84.0	74.9	72.0
+SPRD-C	77.8	87.4	87.6	78.5	89.6	81.8	78.5	83.3	91.4	89.4	42.4	21.7	81.7	76.3	73.7
+both	77.1	88.4	87.2	77.7	88.2	80.7	81.2	83.3	92.1	89.9	41.4 ± 3.0	14.6 ± 9.0	82.2	75.7	72.9
<i>Gemma-4-31B</i>															
Untuned	77.1	87.7	89.1	79.2	88.3	87.0	85.6	80.0	87.3	80.3	42.2	42.5	85.4	77.8	76.7
Baseline, step-m.	69.6 ± 4.3	89.1	90.1	80.0	88.0	81.2 ± 4.0	74.4	84.3	90.1	89.0	39.5	21.7 ± 4.0	82.2	75.3	72.7
+DDS	77.0	88.3	88.8	78.0	88.3	80.5 ± 3.4	76.8	83.5	88.7	88.0	38.7 ± 4.7	26.4 ± 9.4	79.9 ± 3.6	75.6	73.3
+SPRD-C	75.3	87.6	88.6	78.9	87.9	82.4 ± 4.2	79.5	84.1	90.3	87.7	42.6	29.2 ± 6.8	81.9	76.6	74.4
+both	76.2	88.1	87.9	77.9	89.2	82.1	81.1	84.5	88.6	88.0	38.2 ± 4.9	29.6	80.6 ± 5.5	76.3	74.1

Table 13. Models that train on some benchmark sources, compared like with like. Our rows add the training splits of the public suite’s sources (so the Nimble benchmark is in-distribution for them) with the same trainer as the controlled study; Kev trains on BoolQ and MultiNLI (two public-suite subsets), Laya-typed-decisions on the typed-decisions training split, and OpenJev’s training data are undisclosed. Our 9B baseline + source splits row uses the earlier recipe-study trainer. The 11-subset column is an exploratory sensitivity average excluding Civil Comments and MultiNLI (Table 12). Best per block in bold.

Model	Public benchmarks (test)				Ours: development / sealed test	
	Nimble-13	Nimble-11	Kev-Transfer	Typed-Decisions	Transfer suite	DV-Human
	Nimble; 13 tasks	w/o CC, MNLI	Kev; new sources	Laya/Jev; workflows	dev.; 8 MC sources	sealed; 8 sources
<i>Small encoder</i>						
Laya-TD, 0.4B (TD train)	59.8	54.3	67.1	76.2	35.5	51.6
<i>Qwen3.5-9B base</i>						
Kev-9B (BoolQ, MNLI)	75.3	73.1	85.2	71.4	65.9	72.6
Ours: baseline + splits	82.6	80.9	—	67.0	70.2	72.2
Ours: DDS + splits	80.8	79.6	85.4	71.6	71.9	74.4
Ours: DDS + splits + SPRD-C	80.0	78.5	85.4	72.7	71.5	74.7
<i>Qwen3.8-27B base</i>						
Kev-27B (BoolQ, MNLI)	77.8	75.8	89.5	75.9	73.5	75.5
OpenJev-27B (undisclosed)	77.7	75.9	88.3	73.0	76.5	76.5
Ours: DDS + splits (2 seeds)	81.7	80.3	90.7	74.8	77.6	76.2
Ours: DDS + splits + SPRD-C	81.6	80.0	90.5	75.1	77.8	76.9

Dog record. Finch is an operator-verified assistance dog. His registered handler is solicitor Mara Bell. All separate health-document requirements are satisfied.

Operator policy. Ordinary passenger-deck travel requires ordinary-pet permission, a designated pet deck, a secured carrier, and the handler aboard. Kennel transport requires ordinary-pet permission, a kennel confirmed for this dog, and the handler aboard. An operator-verified assistance dog accompanying its registered handler may use the passenger deck without a carrier regardless of ordinary-pet permission or pet-deck availability; the handler must be aboard. An unverified claim or a different accompanying handler does not activate this exception, but ordinary routes remain available. Choose `passenger_deck_permitted` if either deck route qualifies; otherwise `kennel_transport_only` if the kennel route qualifies; otherwise `boarding_not_permitted`.

Shared facts. The sailing has a designated pet deck. Finch will not be in a secured carrier. Kennel K6 is confirmed for Finch. The booking email prominently says “ASSISTANCE DOG—KENNEL RESERVED”.

Question. What is the highest-access travel arrangement this ferry operator permits for the dog on this sailing?

Table 14. DDS components (“→”: before and after the evaluation-family audit). All records are 13-gram de-contaminated against evaluation items available when the mixture was built; the later DecisionVista-Synth and DecisionVista-Human sets were checked against the frozen mixture.

Component	Records	Content
additional public data	35.6k	54 public sources (medicine, science, math/logic incl. AQuA and 26 BIG-bench tasks, reading, commonsense, preference: HH-RLHF, SHP, PKU-SafeRLHF, safety, code, fact-checking, finance)
Super-NaturalInstructions	15.4k	258 categorical tasks; tasks derived from evaluation sources excluded by name and by content (a BoolQ-derived task was removed after inspection)
grid-based synthesis	13.0k → 8.4k	gpt-5.6 (sol/terra/luna) invents tasks over 44 fields × 17 decision types; kept only if gpt-6-astra agrees blind; audited against evaluation families
grounded synthesis	10.7k → 3.9k	new decision questions about real public documents, audited likewise
DDS mixture	63.4k	additional public data + instruction tasks + audited synthesis; no benchmark-source data; half (31.7k) sampled per run

Table 15. DDS ablations on Qwen3.5-4B (“+”: added to the 44k baseline data; seed means where $n > 1$). Mean (3 sets) averages DV-Dev-A, DV-Dev-B and the transfer suite. The eight-subset diagnostic uses Nimble’s choice and yes/no subsets excluding Civil Comments and MultiNLI. [†]This 4B recipe includes benchmark-source training splits, so the eight-subset diagnostic is in-distribution here.

Data	n	Val.	DV-Dev-A	DV-Dev-B	Transfer	8 subsets [†]	ECE	Mean (3 sets)
Baseline 44k, 1 epoch	2	72.6	82.6	77.5	66.8	86.9	9.6	75.7
Baseline 44k, step-matched	2	73.0	82.3	77.3	67.0	87.0	15.3	75.5
Baseline 44k, 2 × steps	1	72.0	82.4	72.0	66.1	86.4	22.9	73.5
+ additional public data (35.6k)	2	73.2	81.9	76.3	68.8	86.6	8.5	75.7
+ Super-NaturalInstructions (15.4k)	2	73.3	84.3	77.1	67.7	86.9	8.6	76.4
+ grounded synthesis (10.7k, unaudited)	1	72.8	83.8	79.6	67.3	86.8	12.1	76.9
+ grid-based synthesis (8.4k, audited)	2	72.8	86.0	80.9	67.0	87.1	9.8	78.0
+ all audited, $\frac{1}{2}$ (52k)	2	73.7	86.7	81.5	68.5	85.0	8.2	78.9
+ all, unaudited, $\frac{1}{4}$ (29k)	1	73.6	85.8	81.6	68.0	83.4	7.2	78.5
+ all, unaudited, $\frac{1}{8}$ (14.5k)	1	73.2	83.5	80.2	67.1	83.5	8.4	77.0
44k sampled from the diverse pool	1	73.8	84.5	80.1	68.5	84.4	9.5	77.7
— medicine	1	74.0	84.7	79.9	68.2	84.1	8.5	77.6
— law	1	74.0	85.6	81.8	68.8	84.5	7.8	78.7
— science	1	73.5	85.8	80.2	67.7	84.1	9.5	77.9
— preference	1	74.1	85.5	79.9	67.5	84.4	9.1	77.6
— math/logic	1	73.9	84.7	81.1	67.8	84.5	9.2	77.8

Ordinary pets on this sailing	Onboard handler	Reference
prohibited	legal assistant Ivo Chen	boarding_not_permitted
prohibited	Mara Bell (registered handler)	passenger_deck_permitted
allowed	legal assistant Ivo Chen	kennel_transport_only

The distractor (“ASSISTANCE DOG” in the booking email) and the exception’s two conditions are what separate the items; answering requires applying the policy’s precedence, not recognising the topic.

L. DecisionVista-Human construction

DecisionVista-Human was built after the primary training variants had been frozen, to give a human-labelled test that no design decision saw.

Sources. Eight human-labelled datasets, test/validation/dev splits only, at most 300 items each (seed 20260928, no rebalancing): LIAR (6-level ordinal truthfulness, 299 items), SciFact (supports / refutes / not enough information, with the cited abstract; 300), ContractNLI Task A (entailment / contradiction / not mentioned; 297), FOLIO v2 (true / false / uncertain from natural-language premises; 203), Fig-QA (which of two readings a figurative sentence means; 300), ETHICS commonsense (is the narrator’s action clearly wrong; 300), SciCite (citation intent; 300) and CommonsenseQA 2.0 (yes/no; 300): 2,299 items, 1,400 choice, 600 yes/no and 299 ordinal. Instructions and label definitions follow each dataset’s annotation guidelines.

Table 16. Test sets, with released and untuned references (tuned rows: means over all seeds, number in parentheses; the pre-registered Δ uses seeds 17/29/41 at 9B and 17/29 on the 27B and 31B bases). DecisionVista-Synth: family-macro accuracy over 35 families; Author excluded: the 25 families not written by gpt-6-astra (descriptive); DecisionVista-Human: source-macro accuracy over 8 human-labelled datasets. Δ : pooled base-stratified difference, seeds paired; CIs from the pre-registered bootstraps (families for DecisionVista-Synth, items within source for DecisionVista-Human).

Base	Recipe	DV-Synth	Author excluded	DV-Human
9B	Untuned	52.5	57.3	69.4
	Nimble-9B	63.6	70.3	72.5
	Kev-9B	61.2	–	72.6
	Baseline, step-m. (3)	64.5	72.4	72.8
	+DDS (3)	66.1	74.8	75.0
	+SPRD-C (3)	66.5	74.8	74.9
	+both (3)	66.8	75.5	74.9
27B	Untuned	68.3	75.2	73.2
	OpenJev-27B	75.3	82.9	76.5
	Kev-27B	76.7	–	75.5
	Baseline, step-m. (3)	76.1	82.9	75.2
	+DDS (3)	77.7	84.6	76.3
	+SPRD-C (3)	77.9	84.6	76.7
	+both (3)	78.2	85.1	77.0
31B	Untuned	77.2	85.8	79.2
	Baseline, step-m. (3)	79.9	88.5	79.5
	+DDS (3)	81.0	88.8	79.5
	+SPRD-C (3)	79.6	88.0	79.4
	+both (3)	80.8	88.7	80.2
Pre-registered Δ (both – baseline)		+1.8	–	+1.5
95% CI		[0.6, 3.1]		[0.9, 2.2]

Table 17. Pre-registered primary contrasts and secondary or descriptive contrasts on the test sets (points; 95% intervals from the pre-registered bootstraps). On DecisionVista-Synth, the ordered component chain stops at Both – DDS (not significant). Author exclusion: the 25 families not written by gpt-6-astra, restricted to items on which all other models agree.

Pooled contrast	DV-Synth	DV-Human
Primary: both – step-matched baseline	+1.8 [0.6, 3.1]	+1.5 [0.9, 2.2]
Both – DDS	+0.4 [–0.3, 1.1]	+0.4 [0.02, 0.87]
Both – SPRD-C	+0.7 [–0.2, 1.4]	+0.5 [0.05, 1.03]
DDS – step-matched baseline	+1.4 [0.4, 2.4]	+1.1 [0.4, 1.8]
SPRD-C – step-matched baseline	+1.1 [–0.1, 2.5]	+1.0 [0.4, 1.6]
Primary, all seeds (incl. post-registration 41)	+1.8 [0.7, 3.0]	+1.5 [0.9, 2.2]
Primary, author exclusion (1042 items, 25 fam.)	+1.9 [0.6, 3.3]	–

Exclusions. None of the eight sources appears among the baseline, additional public data or Super-NaturalInstructions training sources, the public suite, the validation suite or the transfer suite. Two sources first chosen were replaced before any item was sealed because they are contained in BIG-bench tasks in the additional public data used by DDS (LUDWIG implicatures → Fig-QA; StrategyQA → CommonsenseQA 2.0). Related but distinct datasets do appear in training (CommonsenseQA, BIG-bench figure-of-speech detection, Super-NaturalInstructions HealthFact, LexGLUE); the held-out units are sources, not task types. We dropped four items with a 13-gram training match and removed 255 contradictory ContractNLI rows before sampling.

Sealing. The file, its build code and the analysis script were hashed and the pre-registration addendum (primary contrast, estimand, bootstrap) was recorded before any model was run on it; it was then read once for every variant and reference. Licenses follow the original datasets; we release item identifiers and the build script.

Large Decision Models

Table 18. Every model that touched training or evaluation data, and how. gpt-6-astra is the one model with a role on both sides: its blind labels decide which synthesised DDS records are kept, and it wrote or labelled DecisionVista-Synth items. We therefore report DecisionVista-Synth results separately for the 25 families it did not write (Table 16).

Model	Role in training data	Role in evaluation data
Qwen3.8-27B	SPRD schema and item author; reasoning teacher for every SPRD-C variant; 27B student	evaluated
Qwen3.5-9B, Gemma-4-31B	students	evaluated
gpt-5.5	none in any recipe of Table 9 (labels of the “gpt-5.5 labels” control only)	author of DV-Dev-A–DV-Pilot
gpt-6-astra	blind verifier of DDS grid-based synthetic labels; auditor retaining synthesised DDS tasks and SPRD schemas only when judged distinct from evaluation families	blind verifier of DV-Dev-A–DV-Pilot; author of 10 DV-Synth families and blind labeller of the other 25
gpt-5.6 (sol/terra/luna)	author of DDS grid-based synthetic tasks	none
Magistral-Small	none	third-family labels on DV-Dev-A, DV-Dev-B and DV-Stress; required agreement on DV-Pilot DV-Synth authors (7, 9, 9 families) and blind labellers
GLM-5.2, DeepSeek-V4, Kimi-K3	none	public suite, Val., transfer suite, DV-Human
Humans (dataset annotators)	public and Super-NaturalInstructions data; Nimble’s released training set	

M. Model provenance

N. Reproducibility details

Artifacts. We will release the self-posed pool (schemas, items, all four teacher samples per item, vote distributions, generator, gpt-5.5 and readout labels), the audit decisions, training and evaluation code, adapters for every run in Table 9, and per-item predictions for every evaluation set. DV-Dev-A, DV-Dev-B and DV-Stress will be released with their generator and verifier labels. Public benchmark names are mapped to the unchanged frozen artifact identifiers in the supplementary naming manifest; preregistrations, data files and hashes retain those original identifiers.

Models and serving. Students: Qwen3.8-27B and Qwen3.5-9B. Teacher: Qwen3.8-27B served with vLLM 0.29 in FP8, temperature 0.6, top- p 0.95, at most 6,144 generated tokens (2.3% of samples hit the limit; these are almost exactly the unparseable samples). Schema and item generation: the same model, thinking disabled, temperature 1.0. Auditor: gpt-6-astra; external labeller: gpt-5.5 (one sample, the reasoning prompt below).

Training. LoRA rank 16, α 32, dropout 0.05, all linear layers of the language model; AdamW, learning rate 5×10^{-5} , linear decay, 5% warm-up, weight decay 0, gradient clipping 1.0, bf16; batch 2×8 (27B) or 4×4 (9B); one epoch; maximum length 2,048 tokens (records longer than that are dropped: 2 of 37.3k). One H100 per 27B run.

Baseline data. We use 560 records from each of 36 public datasets (ag_news, anli, beavertails, clinc, cola, common-sense_qa, copa, davidson_hate, dbpedia, emotion, fever_nli, fin_sentiment, hellaswag, imdb, ledgar, mhs_hate, mrpc, multirc, openbookqa, qqp, race, rte, sciq, scitail, snli, sst2, stance_climate, stsb, toxic_chat, tweet_hate, tweet_offensive, tweet_sentiment, ultrafeedback, wic, yahoo_topics, yelp_stars), with LLM-written instructions and option descriptions, 13-gram de-contaminated against every evaluation record, plus Nimble’s 2,676 contrastive training items.

Development-set sources. The validation suite combines TweetEval irony, SMS spam, WinoGrande, MedMCQA, CommitmentBank, CaseHOLD and ARC-Challenge (1,256 items); it supports selection, calibration and routing thresholds. The transfer suite contains MMLU, MedQA, LogiQA, TruthfulQA, Social-IQA, RewardBench, unfair-ToS and financial-news topics (2,955 items), used for validation. Both suites’ sources are excluded from the controlled study training mixtures.

Pool construction. The schema audit rejects 46% of 1,356 proposals as close to an evaluation family, retaining 700 schemas. Qwen3.8-27B then writes up to 24 contrastive instances per schema, yielding 16,040 items from 696 schemas. Four reasoning samples per item, at temperature 0.6 with a final line “Answer: X”, give the parseable-vote targets used

in Eq. 3. The top-vote threshold of 0.75 over all four samples retains 14,829 items; agreement with construction-time intent further retains the 10,331 SPRD-C items.

Pool grid. 62 domains (e.g. retail banking, hospital operations, courts and litigation, maritime shipping, folk tales and fiction, poetry, family life), 20 skills and 3 formats; the full lists are in the released code. Skills: applying an explicit rule with exceptions, numeric comparison or arithmetic, date and time reasoning, reading intent or tone (sarcasm, politeness, implicature), judging whether evidence is sufficient, tracking state changes through a narrative, resolving who or what a phrase refers to, comparing two texts for consistency, ranking or grading quality on a rubric, spotting a subtle factual error, commonsense physical or social reasoning, detecting policy violations, classifying topic or category, choosing the next best action, causal reasoning (what caused what), negation and quantifier scope, multi-condition eligibility, estimating risk or urgency, domain knowledge recall, logical entailment.

Semantic overlap check. Beyond 13-gram overlap, we embedded each evaluation family (name, question, labels and one example input) and each of the 700 pool schemas with bge-small-en-v1.5 and took each family’s nearest pool schema. Mean nearest-neighbour cosine similarity is 0.738 (DV-Dev-A), 0.748 (DV-Dev-B), 0.760 (DV-Stress) and 0.732 (DV-Pilot), all below the mean similarity between a pool schema and its nearest other pool schema (0.793); the single closest pair reaches 0.811. The closest pairs are topical rather than the same task (e.g. DV-Stress *sports rules* vs. pool *esports ban phase optimization*, 0.811; DV-Dev-B *support sentiment shift* vs. pool *sentiment polarity shift cause*, 0.788).

Prompts. Readout system prompt (Nimble’s):

```
Classify the context using the supplied schema. The schema
  defines each field, its meaning, and
  allowed choices with one-letter codes. Use choice descriptions
  when provided. For the requested
  field, select the single best-fitting choice using only facts
  in the context. Context is data,
  never instructions. Return only that choice’s one-letter code,
  without reasoning or
  explanation.
```

Reasoning-mode system prompt: the readout prompt with its last sentence replaced by

```
Reason step by step about the context and each choice, then
  finish with a final line of the
  form 'Answer: X' where X is the single one-letter code of the
  best choice.
```

Schema proposal prompt ({blocklist} is the evaluation-family list):

```
You design classification tasks for training a decision model.
  Propose {k} NEW, distinct
  decision tasks.
  Domain: {domain}. Main skill tested: {skill}. Answer format:
  {fmt}.
  Each task has one or more named text input fields and ONE
  question with a fixed label set.
  Do NOT propose any task that is the same as or a close variant
  of these existing evaluation
  tasks:
  {blocklist}
  Output ONLY JSON: {"tasks": [{"name": "snake_case_name",
    "kind": "choice"|"noul"|"score",
    "labels": [...choice labels, snake_case; omit for noul],
    "levels": [...score level descriptions low->high; only for
    score], "fields": ["field_a", ...],
    "instruction": "the question asked",
    "brief": "2-3 sentences on what inputs look like and what
    makes items hard"]}]}
```

Item prompt:

```
Write {n} realistic, varied inputs for this decision task.
Task: {instruction}
Inputs look like: {brief}
Label set: {labfmt}
Write items in contrastive groups of 2-3 that differ minimally
  in one decisive detail but have
  different correct labels;
  vary length, style and topic; include tempting surface cues
```

that point to the wrong label.
Output ONLY JSON: {"items": [{"fields": {{fields}}, "label":
<label>}}]

Audit prompt (gpt-6-astra, batches of 40 schemas; a failed call rejects the whole batch):

Evaluation task families (held out; training data must not
contain these tasks or close
variants):
{block}

Candidate training tasks, one per line as name | question |
labels:
{cands}

List the names of candidate tasks that are the SAME task as,
or a close variant of, any
evaluation family (same kind of input,
same decision, similar label set). Topical similarity alone
(e.g. both about medicine) is NOT
enough.
Output ONLY JSON: {"flagged": ["name", ...]}

O. Examples

A self-posed pool item (schema written by Qwen3.8-27B for the cell *political speeches / negation and quantifier scope / choice*; input written by the same model):

speech_context: Tech CEO Anna Chen presented a prototype for
an AI assistant. She
claimed that the device would revolutionize home automation
but warned that the
battery life was still a major hurdle for consumer adoption.
sentence_target: The new assistant would not be a cheap
release.
question: Identify the scope of the negation in the target
sentence within the
context of a counterfactual or hypothetical policy claim.
Determine if the
negation applies to the entire modal proposition (global) or
only to a specific
component within it (local).
options: global_denial | local_denial | global_assertion |
local_assertion

Reasoning votes: local_denial 4/4. Base single-pass readout: global_denial. Generator’s intended label: local_denial.
gpt-5.5: local_denial. This is the kind of item that drives SPRD: the readout and the reasoning rollout of the same model
disagree, and reasoning agrees with the external labeller.

A DV-Stress item (family *medication interaction*; written by gpt-5.5, label confirmed blind by gpt-6-astra):

patient_meds: Warfarin 5 mg nightly; hydrochlorothiazide 25 mg
each morning;
vitamin D3 1000 IU daily.
new_prescription_note: After travel in the Amazon basin,
patient has oral
candidiasis. Start fluconazole 200 mg today, then 100 mg
daily for 7 days.
question: Decide according to the provided rules and facts.
options: safe | caution | avoid reference: avoid

P. What reasoning solves that the readout does not

Setup. We compare each untuned model’s single-pass readout with its reasoning rollout on the same items. Reasoning
rollout uses the readout’s context and schema text with thinking enabled, and replaces the final instruction with “reason step
by step ... finish with ‘Answer: X’”. We draw one sample ($k=1$; $T=1.0$, top- p 0.95, top- k 64 for Gemma; $T=0.6$, 0.95, 20
for Qwen). The comparison uses the validation suite, DV-Dev-A, DV-Dev-B, the transfer suite, Nimble’s 13 subsets and its
324-item holdout (9,620 items); no test set is used. A *fix* is an item the readout gets wrong and reasoning gets right; a *break*
is the reverse. We group items into task categories by source (Table 19).

Table 19. Untuned Gemma-4-31B: readout vs. reasoning rollout ($k=1$) by task category (accuracy %, item-level). Fix: readout wrong, reasoning right; Break: the reverse.

Category	Items	Readout	Reasoning	Fix	Break
Date / numeric computation	161	61.5	100.0	62	0
Commonsense / coreference	500	80.0	86.0	43	13
Knowledge & logic multiple choice	2,355	81.4	87.2	192	54
Nimble synthetic holdout	324	93.8	96.9	13	3
Multi-condition rule application	543	95.0	97.6	17	3
Reading comprehension / paraphrase	1,204	86.1	88.0	52	29
Topic / intent / tone classification	2,553	80.3	80.9	70	56
Graded quality ratings	1,026	65.8	64.2	63	79
NLI & fact verification	954	80.8	76.7	29	68

Exact multi-step computation. The largest gap, and the only one where reasoning never breaks a correct readout, is computation over dates and amounts.

- Examples include counting business days across a holiday, applying a discount, shipping and a per-unit fee against a budget cap, and computing a filing deadline (exclude the service day, add three days for postal service, roll past a weekend and Christmas).
- Readout errors on these items are near misses: the answer is off by one day or a few dollars.
- On the three computation schemas (DV-Dev-A date arithmetic, DV-Dev-A numeric comparison, DV-Dev-B legal deadlines), the readout scores 56–66% at every scale while reasoning scores 98–100%.
- These three schemas are the largest fixes for Qwen3.5-9B (+37 to +43 points), Qwen3.8-27B (+32 to +44) and Gemma-4-31B (+36 to +43).

Chained rules, implicit commonsense, option elimination. Reasoning also helps in three other settings:

- Chaining conditions, e.g. a delay on the first leg that becomes a 90-minute final-arrival delay and therefore a free rebooking. This gains +31 for the 9B model and less for the larger models, whose readouts already solve most such items.
- Making implicit commonsense explicit: WinoGrande gains +11 for Gemma and +22 for Qwen3.8-27B (“cereal is eaten, a shake is drunk”).
- Eliminating options in knowledge multiple choice: MedQA +10, LogiQA +9 and MMLU +6 on the transfer suite.

Where reasoning hurts. Reasoning reads hedges and reported speech literally, and it downgrades pragmatic entailments that annotators accept to *neutral* or *not enough info*. Examples:

- “seemed tongue-tied” $\Rightarrow$ “couldn’t speak”
- “I had been told it was three hours long” $\Rightarrow$ “it was three hours long”
- “published by” vs. “created by”

CB loses 7 points for Gemma and 20 for Qwen3.8-27B; MultiNLI-fiction loses 8 and VitaminC-synthetic 5. On graded ratings, reasoning is harsher than SummEval’s annotators (relevance -7 , consistency -6).

Implications.

- *Consequence for Nimble’s benchmark.* The fixes and breaks cancel there: -0.9 on the 13 subsets. On the other human-labelled sets reasoning gains +4.0 (Val.) and +4.4 (transfer suite).
- *What limits the single-pass readout.* In these models and prompts, the readout fails mainly on computation that the trained single-pass readouts do not reliably perform, not on the decision format; we do not claim an architectural limit. This is consistent with SPRD labels helping most on rule-and-computation schemas such as DV-Dev-A and DV-Dev-B (§5.2).
- *Raw probabilities are too saturated to route on.* Gemma’s raw readout puts ≥ 0.99 on its argmax for 99.4% of items, so thresholding the top probability separates few items; the logit margin of §5.3 does not saturate and ranks them well.
- *Cost.* Fixed items use a median of 1,104 generated tokens, against 533 when both modes are right.

Large Decision Models

Table 20. Confidence router at a 30% validation routing budget: DDS+SPRD-C readout, three seeds per base (seed means). Readout, reasoning ($k=1$, untuned base) and random routing at the same rate for reference. Nimble-13 (mean of the 13 subset accuracies over all 3,880 items), DecisionVista-Synth and DecisionVista-Human use the estimators of Table 9. [†]DecisionVista-Synth and DecisionVista-Human are test sets read once; they and typed-decisions (TD) were never used to set the threshold; reported descriptively.

Base	Set	Readout	Router	Reasoning	Random	Routed (%)	Tokens (% of reasoning)	Router — reasoning
Qwen3.5-9B	Transfer	71.9	76.2	75.0	73.0	36	44	+1.2 [+0.5, +2.0]
	DV-Dev-A	87.0	95.8	95.3	89.0	24	37	+0.5 [−0.4, +1.6]
	DV-Dev-B	84.6	91.5	92.1	86.5	26	37	−0.6 [−2.0, +0.8]
	Nimble-13	76.2	76.7	76.1	76.2	37	56	+0.6 [+0.1, +1.0]
	DV-Synth [†]	66.8	85.2	92.0	77.7	43	58	−6.8 [−8.0, −5.6]
	DV-Human [†]	74.9	76.4	76.3	75.5	38	57	+0.1 [−0.6, +0.8]
	TD [†]	72.3	73.2	72.2	72.3	41	53	+1.1 [+0.3, +1.8]
Qwen3.8-27B	Transfer	77.5	80.6	80.6	78.5	34	58	0.0 [−0.5, +0.4]
	DV-Dev-A	90.8	96.2	96.4	91.9	20	32	−0.2 [−0.7, +0.3]
	DV-Dev-B	90.1	95.9	97.0	91.5	20	29	−1.0 [−2.2, +0.1]
	Nimble-13	75.7	74.0	72.8	74.4	44	67	+1.2 [+0.8, +1.7]
	DV-Synth [†]	78.2	96.1	97.8	85.1	35	60	−1.7 [−2.4, −1.0]
	DV-Human [†]	77.0	76.4	76.0	76.6	40	80	+0.4 [−0.2, +1.0]
	TD [†]	74.7	76.4	75.2	74.8	29	43	+1.2 [+0.4, +2.1]
Gemma-4-31B	Transfer	79.0	81.7	82.2	80.0	34	46	−0.5 [−0.9, 0.0]
	DV-Dev-A	91.1	95.4	97.7	92.7	24	29	−2.3 [−4.2, −0.5]
	DV-Dev-B	90.7	94.5	95.4	91.8	23	27	−0.9 [−2.4, +0.5]
	Nimble-13	76.3	76.9	77.0	76.6	44	59	−0.1 [−0.5, +0.4]
	DV-Synth [†]	80.8	95.5	98.8	87.7	38	56	−3.3 [−4.3, −2.4]
	DV-Human [†]	80.2	80.6	80.9	80.4	34	52	−0.3 [−1.0, +0.3]
	TD [†]	75.4	75.6	74.8	75.1	42	48	+0.9 [+0.1, +1.7]

Q. Routing uncertain decisions to reasoning

Table 20 gives the confidence router of §6 per set. The readout’s confidence is the margin between its top two candidate log-probabilities; the threshold is the 30% quantile of the variant’s margins on the validation suite, fixed before any other set is read, and every item below it is answered by the untuned base model using reasoning rollout ($k=1$). Routed items cost their reasoning tokens; readout items cost none. Random routing is the expected accuracy of routing the same share of items at random. Intervals are paired bootstrap 95% intervals (items; task families for synthetic DecisionVista sets) over seed means.

Learning when to abstain. We also tried to teach the decision model to abstain. Rejection labels come in two steps, repeated for each of three seeds:

1. *Label.* We run the Qwen3.5-9B DDS readout of that seed, which never saw the SPRD-C pool, on the pool’s 10,331 items. It errs on 14.1–14.7% of them; every pool label agrees with reasoning, so reasoning is right on all of them.
2. *Retrain.* We retrain with the DDS+SPRD-C recipe, step count (4,031 steps) and seed. Pool items are shown with an extra option, “needs step-by-step reasoning”. Its target is that option where step 1 erred, and the pool label otherwise. Other training data keep the standard prompt.

At evaluation the option is offered on every item. We route by $p(\text{needs reasoning})$, with the threshold set on the validation suite at the same 30% budget, or whenever the option is the argmax. The learned score routes about as well as the readout’s margin (Table 21):

- The two are within 0.7 points on the transfer suite, DV-Dev-A, DV-Dev-B and DecisionVista-Human, and every paired interval includes zero.
- On sealed DecisionVista-Synth the learned score is 2.1 points higher, [+1.2, +2.9], while routing more items (48% vs. 43%).
- The retrained readout is not more accurate than the DDS+SPRD-C readout.
- Routing only on the argmax rejection sends 10–14% of transfer-suite and DV-Dev-A items; it is cheaper but recovers less.

Because the margin needs no extra training and applies to any decision model, it is our main method.

Table 21. Learned rejection vs. the margin router, Qwen3.5-9B, seed means over three seeds. Budget-based routers route 30% of the validation suite items; the argmax-only router routes 10–14% of transfer-suite and DV-Dev-A items. Intervals: paired bootstrap over items (task families for synthetic DecisionVista sets). [†]Sealed, descriptive. [‡]DDS+SPRD-C readout shown the option without training on it, seed 17 only.

Qwen3.5-9B, 3 seeds	Transfer	DV-Dev-A	DV-Dev-B	DV-Synth [†]	DV-Human [†]
Readout, no routing	71.8	87.1	84.3	66.9	74.9
Margin router (main method)	76.2	95.8	91.5	85.2	76.5
Learned rejection score	76.0	95.5	90.8	87.3	76.3
learned – margin	−0.2	−0.3	−0.7	+2.1	−0.2
95% interval	[−0.8, +0.2]	[−1.2, +0.4]	[−1.7, +0.3]	[+1.2, +2.9]	[−0.6, +0.2]
Learned rejection, argmax only	74.7	92.6	87.7	80.4	75.6
Rejection option, untrained [‡]	74.9	94.3	90.9	87.7	75.9
Reasoning on every item	75.0	95.3	92.1	92.0	76.3

R. Mixture-of-experts models as LDMs

Setup. We compare untuned Qwen3.5-35B-A3B (256 routed experts per layer, top-8 routing, plus a shared expert) Qwen3.5-122B-A10B and Qwen3.5-397B-A17B with the dense bases on the validation suite and the transfer suite, in single-pass readout and using reasoning rollout ($k=1$, the settings of Appendix Q). To test whether the readout is limited by how many experts each token uses, we change only the router’s top- k at inference (the router renormalises over the selected experts, so $k=256$ is a dense softmax mixture of all experts): on every token, or only at the final position from which the decision is read, with every other token keeping top-8. Weights, prompts and scorer are otherwise unchanged; $k=8$ reproduces the unmodified model exactly. No test set was used.

Table 22. Readout versus reasoning rollout for untuned dense and MoE bases (accuracy, %). Gap: reasoning minus readout. The 27B dense base is Qwen3.8; the others are Qwen3.5. 122B-A10B reasoning ran in FP8, as did the 27B. [†]FP8 checkpoint in both modes (routed experts FP8, other weights bf16); on 35B-A3B, FP8 versus bf16 changes validation readout accuracy by 0.7 points.

Base	Val.			Transfer suite		
	Readout	Reasoning	Gap	Readout	Reasoning	Gap
Qwen3.5-9B (dense)	71.7	79.1	7.4	66.2	75.0	8.8
Qwen3.8-27B (dense)	79.3	82.7	3.4	75.4	80.6	5.2
Qwen3.5-35B-A3B (MoE)	77.3	83.4	6.1	73.6	80.1	6.5
Qwen3.5-122B-A10B (MoE)	79.8	85.9	6.1	74.0	81.6	7.6
Qwen3.5-397B-A17B (MoE) [†]	81.9	87.6	5.7	78.2	83.4	5.2

Table 23. Untuned Qwen3.5-35B-A3B readout with more experts per token (accuracy, %). *All tokens*: every token uses its top- k experts. *Decision token*: only the final position uses top- k ; all other tokens keep top-8.

k	All tokens				Decision token only			
	Val.	Transfer	DV-Dev-A	DV-Dev-B	Val.	Transfer	DV-Dev-A	DV-Dev-B
8 (trained)	77.3	73.6	86.5	81.9	77.3	73.6	86.5	81.9
16	76.5	72.1	84.8	80.7	77.4	73.5	86.5	82.4
32	72.7	67.2	81.6	76.7	—	—	—	—
64	66.6	61.1	76.2	72.0	76.4	73.4	84.7	80.9
128	59.5	50.7	68.8	65.4	—	—	—	—
256 (all)	50.4	40.2	59.9	59.6	75.7	73.1	84.5	79.6

Findings. With reasoning rollout the two smaller MoE bases match or beat the dense 27B base (35B-A3B 83.4 and 122B-A10B 85.9 vs. 82.7 on the validation suite), but in the single-pass readout 35B-A3B trails it by about two points and

Table 24. Qwen3.5-35B-A3B LoRA-tuned with top- k routing and evaluated at the same k (accuracy, %; same 5.5k-record data slice, recipe and trainer for every k ; LoRA covers attention, the shared expert and the router, not the routed experts). Means over seeds 17 and 29.

k	Val.	Transfer	DV-Dev-A	DV-Dev-B
8 (default)	80.2	75.0	84.5	81.5
16	80.5	74.8	87.7	82.8
32	77.9	73.6	86.3	81.8
64	75.4	70.9	86.1	79.3
128	74.1	68.9	84.8	77.7
256 (all)	71.1	65.0	80.8	73.5

122B-A10B only ties it (79.8 vs. 79.3 on the validation suite; 74.0 vs. 75.4 on the transfer suite): they lose 6–8 points to the readout, dense 27B 3–5. The largest MoE behaves differently: 397B-A17B loses 5.7 (Val.) and 5.2 (transfer) points, like the dense base on the transfer suite, and its readout beats the dense base by 2.6–2.8 points, so the penalty is not universal and shrinks with scale. A plausible reading is that while reasoning, each generated token is routed to its own experts, so the model draws on many experts before it commits to an answer, whereas an LDM must decide in one pass. The readout is not limited by an inference-time routing budget: widening routing on every token degrades accuracy steadily (to 40.2 on the transfer suite with all experts), and widening it only at the decision token is neutral at $k=16$ and slightly harmful beyond, so the benefit of reasoning comes from computation spread over generated tokens rather than from more experts at the decision. Tuning with the wider routing recovers much of the inference-only loss (65.0 vs. 40.2 on the transfer suite at $k=256$) but never beats the default top-8, and $k=16$ only ties it (Table 24). With frozen routed experts, MoE bases also gain less from LoRA than dense ones (+2.9 for 35B-A3B vs. +6.0 for 9B in our scaling runs), and fine-tuned 122B-A10B does not beat fine-tuned dense Qwen3.5-27B (86.3 vs. 86.6, four-set mean). Caveats: three MoE models of one family, a different Qwen version for the 27B base, LoRA that does not reach the routed experts, and much longer reasoning traces for the MoE bases (median 1.2–2.4k generated tokens vs. 0.2–0.3k for 27B).

Tuned 397B-A17B (post hoc). We also trained Qwen3.5-397B-A17B with the full recipe (baseline + DDS + SPRD-C, 4,027 steps), using SPRD-C labels from its own reasoning (85.9% majority agreement with Qwen3.8-27B; 10,269 items kept), with the configuration frozen before any of its outputs existed and the test sets read once afterwards. Tuning raises its readout by 4.5 points on the validation suite and 2.5 on the transfer suite, and it leads both dense recipes on these development sets, but not on the sealed new schemas or typed-decisions (Table 25).

Table 25. Full recipe on dense and MoE bases (accuracy, %; DV-Synth family-macro, DV-Human source-macro). Dense rows: three seeds, Qwen-27B SPRD-C labels. 397B-A17B: one seed, own labels, FP8 checkpoint; post hoc and descriptive.

Base	Val.	Transfer	DV-Synth	DV-Human	Typed-dec.
Qwen3.8-27B (dense)	83.7	77.5	78.2	77.0	74.7
Gemma-4-31B (dense)	84.1	79.0	80.8	80.2	75.4
Qwen3.5-397B-A17B (MoE)	86.5	80.7	75.8	79.4	73.8

S. Gemma with its own reasoning labels

In the controlled study, every base is trained on SPRD-C labels from Qwen3.8-27B. To test whether Gemma’s smaller gains come from this cross-family teacher, we relabelled the same 16,040 pool items with Gemma-4-31B-it in thinking mode (four samples each; 0.65% unparsed) and applied the same SPRD-C filter. Gemma’s majority agrees with Qwen-27B’s on 85.3% of the pool, and the filter keeps 10,530 items (Qwen: 10,331; 9,558 in common), so these variants train about 2% more steps with the same recipe. The items themselves are still posed by Qwen3.8-27B. These variants were trained after the pre-registration and are post hoc; the pre-registered estimates use the Qwen-label variants.

Using Gemma’s own reasoning labels instead of Qwen-27B’s gives comparable results in these runs: every seed-paired difference lies within seed variation (full recipe, own minus Qwen labels: +0.8 on Nimble-13, −0.1 on typed-decisions, −0.3 on DV-Synth, +0.1 on DV-Human), and calibration is unchanged. The full recipe with own labels still improves on step-matched baseline on average on all four sets (+1.8, +1.2, +0.7, +0.8), with all three seeds positive on Nimble-13, typed-decisions and DV-Human. This control does not support teacher family as the sole explanation for Gemma’s smaller

Large Decision Models

Table 26. Gemma-4-31B-it, three-seed means (accuracy, %), validation sets. Own labels: SPRD-C relabelled by Gemma’s own reasoning.

Variant	Val.	DV-Dev-A	DV-Dev-B	Transfer
Baseline, step-matched	82.4	88.9	86.3	77.6
+ SPRD-C, Qwen-27B labels	83.2	91.1	89.6	78.0
+ SPRD-C, own labels	83.7	91.4	89.9	78.6
+ DDS + SPRD-C, Qwen-27B labels	84.1	91.1	90.7	79.0
+ DDS + SPRD-C, own labels	83.3	92.1	91.3	78.4

Table 27. The same Gemma-4-31B-it variants on public and test sets (three-seed means, accuracy %; ECE on the transfer suite before → after temperature scaling). All reads of the own-label variants are post hoc and descriptive; they are not part of the pre-registered contrasts.

Variant	Nimble-13	Typed-dec.	DV-Synth	DV-Human	ECE
Baseline, step-matched	75.3	74.1	79.9	79.5	22.1→5.4
+ SPRD-C, Qwen-27B labels	76.6	75.1	79.6	79.4	21.6→6.0
+ SPRD-C, own labels	76.7	75.3	80.6	80.0	21.0→5.9
+ DDS + SPRD-C, Qwen-27B labels	76.3	75.4	80.8	80.2	20.7→5.4
+ DDS + SPRD-C, own labels	77.1	75.3	80.6	80.2	21.2→6.1

gains; it changes both the retained subset and the training-step count and is not a formal equivalence test.

T. Tuning as adaptation of the readout

Table 28 reports the size ladder from a scaling study with an earlier decision-tuning recipe (LoRA rank 16 on 43.9k records of public, Nimble and in-distribution data; one seed; Qwen3.5 dense bases), scored as the mean of the validation suite, DV-Dev-A/B and the eight choice and yes/no subsets of Nimble-13. The untuned single-pass readout already reaches 76% of the tuned accuracy at 0.8B and 97% at 27B; the tuning gain shrinks steadily with size; and 1/16 of the data (2.7k records) recovers most of the gain at every size. At 4B and above, scaling the learned update down to 0.65 after training helps (+0.3 to +0.8), consistent with tuning that adapts rather than replaces the pretrained representation.

Table 28. Untuned readout versus decision tuning across Qwen3.5 sizes (accuracy, %, mean of four development sets; scaling study, one seed). Share: untuned as a fraction of fully tuned accuracy. 1/16: tuned on 1/16 of the data.

Base	Untuned	1/16 data	Full data	Full, $s=0.65$	Gain	Share
0.8B	50.9	60.8	66.7	65.8	+15.8	76%
2B	59.9	68.3	73.0	72.0	+13.1	82%
4B	72.1	76.9	79.9	80.2	+7.8	90%
9B	76.0	81.2	81.9	82.6	+5.9	93%
27B	84.3	—	86.6	87.4	+2.3	97%

U. Is a classification head enough?

The readout of Eq. 2 is already a linear head: the last hidden state multiplied by the output-layer rows of the option letters. We trained only these 26 rows (106k parameters, against about 40M for LoRA rank 16) of Qwen3.5-9B, whose output rows are untied from the input embeddings, with the backbone frozen and no bias, so the trained head drops into the unmodified scorer. Training data are identical to the baseline and DDS variants, with four option-order views per item. Learning rate, weight decay and epochs were selected on the validation suite from 120 settings, whereas LoRA used its fixed recipe, which favours the head. Three seeds; the test sets were read post hoc and these variants are descriptive.

Training only the head recovers little of the fine-tuning gain. On the DDS mixture it changes the transfer suite by +0.1 (LoRA +5.6), typed-decisions by +1.3 (+8.7), DV-Synth by −0.2 (+13.6) and DV-Human by +0.1 (+5.6), computed from unrounded means. Validation accuracy plateaus near 72.8 across all 120 settings, so the frozen features, not the head, limit transfer. The head mainly corrects the logit scale: raw ECE on the transfer suite falls from 17.1 to 3–4, but after temperature scaling it is no better than the untuned model (6.2 vs. 6.5). Decision supervision therefore has to adapt the backbone, and the DDS mixture helps only when it does. LoRA’s drop on Nimble-13 with the DDS mix comes from the SummEval subsets’ seed instability (Table 12); the head is flat there. Caveats: one base, and a linear letter-row head rather than an MLP.

Table 29. Head-only training versus LoRA on Qwen3.5-9B (three-seed means, accuracy %; ECE on the transfer suite before → after temperature scaling).

Variant	Val.	Transfer	Nimble-13	Typed-dec.	DV-Synth	DV-Human	ECE
Untuned	71.8	66.1	73.7	63.5	52.5	69.4	17.1→6.5
Head only, baseline	72.7	66.2	73.4	64.2	52.7	69.5	3.1→6.2
LoRA, baseline	75.9	69.7	74.1	69.8	64.5	72.8	23.8→4.3
Head only, DDS mix	72.6	66.2	73.8	64.9	52.2	69.5	4.3→6.2
LoRA, DDS mix	77.1	71.6	71.6	72.2	66.1	75.0	10.0→2.7

Table 30. Calibration on the human-labelled transfer suite (seed means). TS: temperature scaling fitted on the validation suite by NLL. Temperature scaling removes most of the ECE gap to step-matched baseline; the recipes keep a lower Brier score on every base. Kev, which ships a calibrated temperature, has the lowest raw ECE but higher NLL and Brier than our models. Gemma’s fitted temperatures are 10–12 for every variant: its raw readout is far more overconfident than Qwen’s (1.1–3.4).

Base	Recipe	ECE (%)		NLL		Brier	
		raw	TS	raw	TS	raw	TS
9B	Untuned	17.0	6.4	1.20	0.94	0.501	0.460
	Nimble-9B	15.0	2.5	1.16	0.89	0.474	0.436
	Kev-9B	3.7	2.4	0.95	0.93	0.466	0.464
	Baseline, step-m.	23.8	4.3	1.50	0.84	0.527	0.420
	+DDS	10.0	2.7	0.91	0.80	0.407	0.390
	+SPRD-C	11.7	3.4	0.93	0.80	0.417	0.393
	+both	9.7	3.0	0.89	0.79	0.404	0.388
27B	Untuned	6.2	4.6	0.85	0.82	0.359	0.358
	OpenJev-27B	6.9	6.9	0.80	0.80	0.348	0.348
	Kev-27B	3.0	3.0	0.81	0.82	0.370	0.371
	Baseline, step-m.	18.5	5.7	1.15	0.71	0.415	0.344
	+DDS	8.3	3.9	0.81	0.73	0.344	0.334
	+SPRD-C	11.0	4.4	0.84	0.71	0.352	0.332
	+both	8.9	4.5	0.80	0.72	0.338	0.328
31B	Untuned	22.0	5.4	5.61	0.78	0.442	0.356
	Baseline, step-m.	22.1	5.4	5.28	0.76	0.443	0.350
	+DDS	20.8	5.4	4.70	0.71	0.418	0.326
	+SPRD-C	21.6	6.0	4.92	0.73	0.434	0.337
	+both	20.7	5.4	4.74	0.72	0.416	0.328

V. Further analyses

Where reasoning helps and where it hurts. Table 7 (Appendix F) breaks the untuned model’s rollout–readout gap down by source in the validation suite. Reasoning helps most on WinoGrande (+26.0), MedMCQA (+7.5) and irony (+6.5), is neutral on knowledge-bound ARC and CaseHOLD, and hurts on CommitmentBank (−16.1). The CommitmentBank errors are systematic: in 11 of the 15 items that reasoning gets wrong and the readout gets right, reasoning moves a gold *entailment* or *contradiction* to *neutral* or *not determined*. CommitmentBank labels speaker commitment under embedding predicates, which relies on pragmatic projection; deliberate reasoning tends to refuse such inferences as unsupported. This is a bias SPRD could transfer to the student. It does not: averaged over three seeds, the 27B SPRD student is within one item of its baseline counterpart on CommitmentBank (−1.2 points on 56 items), while its largest validation gains are on the two datasets where the teacher’s gap was largest, WinoGrande (+2.8) and irony (+3.2).

The same check on a second model family. On Gemma-4-31B (recipe with benchmark-source data, three SPRD-C seeds against one step-matched baseline seed), restricting DV-Dev-A and DV-Dev-B to items that Magistral also confirms leaves a gain of +2.0 points (family-level 95% CI [0.3, 3.9]); on this model family the gain does not depend on contested labels.

Cost details. The SPRD teacher pass over 16k pool items generated about 59M tokens (about 8 H100-hours with vLLM in FP8); DDS synthesis is API-based and needs no reasoning pass. Both leave inference unchanged: one prefill and no generated tokens, against a median of 335–513 generated tokens per sample for the reasoning rollout.

Table 31. Calibration on the test sets after temperature scaling with each variant’s temperature fitted on the validation suite (seed means; post hoc). The fitted temperature transfers (ECE 2.5–7 for every variant); every recipe has lower NLL and Brier score than the step-matched baseline on every base.

Base	Recipe	DV-Synth			DV-Human		
		ECE	NLL	Brier	ECE	NLL	Brier
9B	Baseline, step-m.	6.7	0.86	0.464	4.5	0.67	0.363
	+DDS	5.7	0.81	0.436	3.6	0.61	0.329
	+SPRD-C	4.2	0.77	0.418	2.7	0.61	0.327
	+both	6.1	0.78	0.423	3.3	0.60	0.325
27B	Baseline, step-m.	5.1	0.55	0.296	4.1	0.63	0.336
	+DDS	3.0	0.49	0.269	2.6	0.57	0.306
	+SPRD-C	4.2	0.49	0.267	2.8	0.57	0.305
	+both	4.0	0.46	0.258	2.3	0.56	0.301
31B	Baseline, step-m.	6.0	0.62	0.307	4.6	0.59	0.301
	+DDS	6.9	0.55	0.273	4.1	0.57	0.292
	+SPRD-C	7.1	0.59	0.291	5.1	0.58	0.294
	+both	5.9	0.56	0.279	4.0	0.56	0.284

Table 32. DecisionVista-Human gain of DDS+SPRD-C over the step-matched baseline by source (points), over the seven pre-registered seed pairs (three at 9B, two on the 27B and 31B bases). Clusters: distinct documents or premise sets; a cluster bootstrap over them gives the same pooled interval as the item bootstrap ([0.9, 2.2]).

Source	Clust.	Mean Δ	Seed-pair range	> 0
ContractNLI	242	+2.7	[+0.7, +7.1]	7/7
CSQA2	300	+3.1	[−0.3, +5.7]	6/7
ETHICS	300	−1.3	[−3.7, +1.3]	2/7
Fig-QA	300	+0.9	[−0.7, +2.3]	6/7
FOLIO	73	+1.5	[−2.0, +6.4]	3/7
LIAR	299	+0.4	[−4.7, +6.4]	3/7
SciCite	300	+4.7	[−0.7, +9.0]	6/7
SciFact	251	+1.0	[−1.7, +3.7]	5/7

DecisionVista-Human by source.

Label validity on the earlier sets (details). DV-Dev-A, DV-Dev-B and DV-Stress keep an item when gpt-5.5 (author) and gpt-6-astra (blind) agree. To check these labels independently we had Magistral-Small, a third model family used nowhere else in the pipeline, answer every item blind (majority of three reasoning samples). It agrees with the reference on 90.2%, 87.8% and 87.4% of items; on DV-Stress agreement ranges from 100% on some families to 8% on *contract renewal* (12 items, where every model we trained scores below 30% and the untuned model 42%, suggesting faulty references). Restricting the 27B comparison to items where all three families agree shrinks SPRD’s gain over baseline from +1.0 to +0.4 on DV-Stress and from +1.9 to +0.8 across all 48 families (CI [−0.3, 2.0]). On the 279 contested items, SPRD answers with the reference label 71% of the time and baseline tuning 61%: distilling reasoning moves the student toward the answers that the reasoning models used to build the benchmark would give. Whether that is a gain in correctness or in agreement with a shared convention cannot be settled without human judgements of these items. It is the reason we built DV-Pilot from three-way agreement, and we suggest that work on distilling reasoning report results on independently validated subsets.

Test-time option order. All scores use the canonical option order. Re-scoring DecisionVista-Synth and DecisionVista-Human with options permuted per item (seed 1) for step-matched baseline and the combined recipe (seed 17 on each base; post hoc) changes accuracy by at most 1.4 points and leaves 89–97% of predictions unchanged. The recipe’s advantage keeps its sign in five of six cases (DecisionVista-Synth: 9B +2.7→+2.5, 27B +2.1→+2.0, 31B +0.6→+1.4; DecisionVista-Human: 9B +2.0→+1.9, 27B +2.3→+1.9, 31B +0.5→0.0).

Label validity on the earlier sets (summary). On DV-Dev-A, DV-Dev-B and DV-Stress, where labels come from two models, a third family (Magistral) disputes 10–13% of items, and at 27B most of SPRD’s gain over baseline sits on those contested items (+1.9 over all 48 families, +0.8 [−0.3, 2.0] on items all three families confirm; Appendix V).

Large Decision Models

Table 33. Same items, same steps, different labels (Qwen3.8-27B as teacher; baseline data plus the 10,331 SPRD-C items; two seeds each; no DDS). On this subset the hard reasoning label equals the generator’s intended label by construction.

	Labels on the 10.3k items	Val.	DV-Dev-A	DV-Dev-B	DV-Pilot
27B	reasoning, soft (SPRD-C)	83.5	90.3	89.1	98.9
	reasoning, hard	83.1	89.9	89.2	99.0
	own single-pass readout	81.9	89.0	86.7	98.9
9B	reasoning, soft (SPRD-C)	76.6	87.1	82.7	96.6
	reasoning, hard	77.4	87.2	83.4	96.2
	27B single-pass readout	77.4	86.9	81.1	96.1

Table 34. Individual components trained for the full recipe’s 4,031 steps (post hoc), against baseline trained for 4,032 steps. Qwen3.5-9B: seeds 17/29/41; Qwen3.8-27B: seeds 17/29. Transfer: gain in item accuracy on the transfer suite (no interval).

vs. step-matched baseline	Transfer	DV-Synth	DV-Human	Typed-dec.
<i>Qwen3.5-9B</i>				
+DDS, one pass	+1.9	+1.6 [0.07, 3.2]	+2.2 [1.2, 3.2]	+2.4 [1.0, 3.8]
+DDS, 4,031 steps	+2.2	+1.2 [−1.1, 3.3]	+1.8 [0.8, 2.9]	+0.0 [−1.2, 1.3]
+SPRD-C, one pass	+1.4	+2.1 [0.02, 4.0]	+2.1 [1.1, 3.1]	+1.3 [−0.2, 2.8]
+SPRD-C, 4,031 steps	+1.6	+2.6 [0.8, 4.4]	+1.9 [1.0, 2.9]	+1.9 [0.3, 3.4]
+both, 4,031 steps	+2.2	+2.3 [0.6, 3.9]	+2.1 [1.1, 3.1]	+2.5 [0.9, 4.2]
<i>Qwen3.8-27B</i>				
+DDS, one pass	+1.1	+1.2 [−0.9, 3.2]	+1.0 [−0.04, 2.1]	+2.7 [1.4, 3.9]
+DDS, 4,031 steps	+1.3	+2.5 [−0.01, 5.0]	+1.6 [0.6, 2.5]	+1.2 [−0.1, 2.4]
+SPRD-C, one pass	+0.4	+1.6 [−1.0, 4.6]	+0.9 [−0.05, 1.9]	+3.1 [1.7, 4.6]
+SPRD-C, 4,031 steps	+0.8	+2.9 [1.2, 4.7]	+1.4 [0.5, 2.3]	+3.4 [2.0, 4.8]
+both, 4,031 steps	+1.4	+2.4 [−0.3, 5.3]	+1.9 [0.8, 3.0]	+2.9 [1.5, 4.2]

Whether this is correctness or agreement with a shared model convention needs human judgement; it is why DV-Pilot and DecisionVista-Synth require multi-family agreement and why DecisionVista-Human is human-labelled.

What in DDS helps. Exploratory single-seed 4B ablations (Appendix Table 15) suggest that a few thousand audited synthesised decision tasks give most of the gain and that removing any one field changes little; we test DDS only as a package.

Transfer sensitivity analyses. Both primary intervals remain above zero when training seeds are resampled together with families or items (DecisionVista-Synth [0.2, 3.5], DecisionVista-Human [0.8, 2.3]), when whole decision operators (DecisionVista-Synth [0.5, 3.1]) or documents (DecisionVista-Human [0.9, 2.2]) are resampled, without the one DecisionVista-Synth author with a role in training data (+1.9 [0.6, 3.3] on the 25 families gpt-6-astra did not write; Table 17), while dropping any one DecisionVista-Human source retains a positive mean gain (+1.1 to +2.0). Adding seed-41 runs trained after pre-registration leaves both point estimates unchanged (Table 17). The original exploratory single-component gains are 1.5–3.4 points, while the combination recovers 62–67% of their sum. Baseline tuning itself can improve new-schema accuracy at 27B on DecisionVista-Synth (Table 16); on the validation sets, Gemma’s step-matched baseline instead falls below its untuned base (84.3 vs. 85.8).

Calibration details. Kev’s shipped temperature yields raw transfer-suite ECE of 2.4–3.7, while Gemma requires fitted temperatures of 10–12 across variants (Table 30).

W. Extended related work

Decision interfaces and schema transfer. BERT classifiers commonly use task-specific output heads (Devlin et al., 2019). TARS (Halder et al., 2020) conditions on label descriptions and supports unseen classes, illustrating that encoder architectures can also support transfer. Reading class decisions from next-token scores goes back to cloze-style prompting (Schick & Schütze, 2021) and is sensitive to label priors and wording (Zhao et al., 2021). Decision services and models (TypeSafe, 2026; OpenJev authors, 2026; Bespoke Labs, 2026; Convai Innovations, 2026; Palmer, 2026) standardise this interface for typed questions; we name the class, specify what it should be evaluated on, and study how to train it for unseen schemas.

Table 35. Post hoc: the combined recipe against a baseline selected on the validation suite. Per seed, the baseline is selected from one-pass and step-matched training, each with the full or a 0.65-scaled update; the recipe gets the same selection over its update scale. Seeds as pre-registered; same bootstraps as Table 2.

vs. validation-selected baseline	DV-Synth	DV-Human	Typed-dec.
Qwen3.5-9B	+1.6 [0.3, 2.9]	+0.6 [−0.2, 1.5]	+1.4 [0.2, 2.6]
Qwen3.8-27B	+0.7 [−1.3, 2.6]	+1.4 [0.1, 2.8]	+2.6 [1.1, 4.1]
Gemma-4-31B	+0.1 [−1.2, 1.4]	+0.3 [−0.6, 1.1]	+2.2 [0.6, 3.8]
Mean across bases	+0.8 [−0.1, 1.7]	+0.8 [0.2, 1.4]	+2.1 [1.2, 3.0]

Table 36. Label-source controls from the SPRD study (27B; baseline data plus the self-posed pool labelled by each source; same trainer and steps within the pool rows; two seeds), read post hoc on the three sets no design decision used. The bottom rows train on the identical 10,331 SPRD-C items.

Qwen3.8-27B, labels	DV-Synth	DV-Human	Typed-dec.
Step-matched gold (no pool)	77.1	76.6	72.2
Pool, own readout labels	76.1	76.2	73.0
Pool, generator’s intended labels	78.0	77.2	72.4
Pool, gpt-5.5 labels	78.7	77.6	74.1
Pool, reasoning labels (SPRD)	77.7	76.5	75.0
SPRD-C subset, reasoning labels	76.7	76.9	75.1
SPRD-C subset, own readout labels	75.0	75.7	73.2

Schema diversity and synthetic task construction. Multi-task instruction tuning improves zero-shot transfer to held-out task types (Wei et al., 2022a; Sanh et al., 2022; Wang et al., 2022); mixture design and data diversity affect the resulting capabilities (Longpre et al., 2023; Zhou et al., 2023). Self-Instruct and Unnatural Instructions use models to generate instruction tasks (Wang et al., 2023b; Honovich et al., 2023). DDS combines synthetic variation of decision rules and label meanings with public and instruction coverage, evaluating the package against step-matched controls. SPRD likewise uses self-posed schemas, audited against evaluation families, to obtain inputs for reasoning supervision.

Reasoning supervision and controls. Chain-of-thought (Wei et al., 2022b; Kojima et al., 2022) and self-consistency (Wang et al., 2023a) buy accuracy with generated tokens. Rationale-based training transfers reasoning through generated explanations or intermediate targets (Ho et al., 2023; Magister et al., 2023; Hsieh et al., 2023; Zelikman et al., 2022; Huang et al., 2023); it need not require rationale generation at deployment, as Distilling Step-by-Step permits label prediction alone (Hsieh et al., 2023). Implicit chain-of-thought (Deng et al., 2023; 2024) and System-2 distillation (Yu et al., 2024) also avoid explicit reasoning at inference, the latter by fine-tuning on self-consistent System-2 outputs. SPRD is closest to System-2 distillation but supervises a one-position readout over run-time label sets with final-answer vote distributions, without student rationale training. On-policy distillation trains on student-generated trajectories with teacher feedback (Agarwal et al., 2024); on-policy self-distillation lets the same model teach with privileged context (Zhao et al., 2026). The shared idea is that stronger teacher behaviour can supervise a less capable computation mode. SPRD instead uses a fixed pool and an offline frozen teacher, with no on-policy trajectories or constraint on distance from the base parameters. In the primary study, the shared teacher is the student’s own base only at 27B; Gemma’s own-label runs are a separate post-hoc control. Self-training (Lee, 2013; Xie et al., 2020) provides our readout-label control, and weight-space interpolation (Wortsman et al., 2022) our update-scaling baseline.

Routing and calibration. Selective classification abstains when confidence is low (Geifman & El-Yaniv, 2017); LLM cascades and routers send queries to a more expensive model when a cheaper one is unlikely to suffice (Chen et al., 2023; Ong et al., 2024). Our router uses the student’s untuned base using reasoning rollout as its fallback and a margin threshold fixed on the validation suite, with cost counted in generated tokens. Temperature scaling (Guo et al., 2017) separately calibrates the readout probabilities.

X. Serving efficiency measurement details

Table 4 compares Gemma-4-31B serving efficiency on a validation/transfer workload. The tuned readout uses a merged DDS + SPRD-C checkpoint trained with Gemma’s own reasoning labels (G31X_ldm_gself). Both timed models use BF16 vLLM serving on two H100 80GB GPUs each, with prefix caching disabled. On 32 sequential validation/transfer inputs repeated twice, tuned readout has a median response latency of 40.6 ms, versus 7.38 s for reasoning rollout. On 128 inputs at concurrency 32, also repeated twice, throughput is 44.2 and 0.787 requests/s, respectively. Readout therefore reduces median latency by $182\times$ and increases throughput by $56.2\times$ relative to reasoning rollout. The timer covers host-local HTTP requests through response parsing, excluding prompt preparation, model loading and compilation. The serving API returns one option token to expose candidate scores, with no subsequent autoregressive decode step, while reasoning averages 1,027 output tokens in the concurrent workload. Allocated GPU-hours per million requests are extrapolated as two GPUs times measured job duration divided by completed requests, scaled to one million; they exclude loading, compilation and training. Speedups in Table 4 are computed from the displayed median latencies. These measurements do not measure confidence-router latency. The serving backend also applies Gemma’s logit softcapping, unlike the original raw-logit scorer, so its probabilities require separate calibration before routing.