Multimodal Reasoning Classifiers Approach the Frontier at Lower Cost and Suggest Two Axes of Intelligence in Language Models

Max Zhang
max.zhang@mail.mcgill.ca

November 6, 2026

Abstract

A one-pass option scorer returns a distribution over an arbitrary option set at the cost of one prefill [47], but it cannot perform intermediate inference. We let a post-trained reasoning model complete its chain of thought over a text or an image and then score every option in one packed pass conditioned on the reasoning: a multimodal reasoning classifier whose output is a distribution over the option set. Two results follow. First, reasoning improves the classifier where the answer must be inferred. On seven held-out tasks whose answer must be inferred, reasoning raises Gemma 4's task mean by 0.14 to 0.16 at every size from E2B to 31B, over three reasoning seeds and, at 31B, on the complete test sets. After the same reasoning the scored decision exceeds the model's own reply by 0.014 to 0.038, in Qwen3-VL as in Gemma 4, and needs no parsing; a readout trained on the model's own reasoning lowers its expected calibration error from about 0.2 to 0.07 at a cost of 0.015 in text score. At 31B the scored decision is at or above the median of thirteen hosted frontier models on one of the seven tasks and within 0.02 of it on three more, and no reasoning setting of these models reaches its accuracy at a lower cost. Second, because the same examples give a one-pass decision and a reasoned one, the classifier separates what reasoning adds from what scale adds. Reasoning raises the inference axis by 0.09 to 0.19 and the classification axis (eighteen tasks) by at most 0.03, and the paired difference of the two shifts excludes zero at every size; scale moves both axes, in Gemma 4 and in Qwen3-VL. Across the thirteen frontier models the two axes correlate at 0.90 on text and 0.39 on images and one factor dominates their scores: the axes are not separate kinds of intelligence across models but separable effects within one. Reinforcement learning of the reasoning, rewarded by the scored decision, moves E4B by 0.01 to 0.03 on the inference axis and by about 0.01 on the classification axis in three runs, while the reasoning lengthens two- to five-fold, and the same recipe at 26B-A4B gains nothing at 31 updates; rewarded by the log score, the policy stops reasoning.

1 Introduction

Selecting from a known option set requires no decoding: the options can be scored by teacher forcing, and packing them behind the prompt with a block attention mask scores all of them in one forward pass. Trained with cross-entropy over the option set on a mixture of public tasks, such a scorer transfers to unseen label sets and is calibrated on them [47]. Its limitation is the one-pass structure. With thinking disabled, Gemma 4 31B scores 0.510 on GPQA Diamond [32], against 0.93 for the best of thirteen hosted frontier models. Chain-of-thought prompting and reinforcement-trained reasoning address this class of task [12, 40], but a reasoning model returns a sampled answer without a usable probability.

We combine the two, for text and images. A post-trained model writes its chain of thought until it closes the thinking block; we then append the task's cue line and score every option in one packed pass. The output is a distribution over the option set conditioned on the prompt, the image if any, and the reasoning. With thinking disabled, the method is the one-pass classifier. We ask whether the classifier reaches the frontier at

equal cost, what reasoning does to calibration, whether the reasoning can be trained with the scored decision as the reward, and, since the same examples give a one-pass decision and a reasoned one, which of a model's capabilities each of reasoning, scale and training moves.

Contributions.

1. **A multimodal reasoning classifier.** Reasoning followed by one packed scoring pass, with the option terminator adapted to post-trained models (Section 3). On seven tasks whose answer must be inferred, reasoning raises the task mean by 0.14 to 0.16 at every Gemma 4 size from E2B to 31B (Section 5.1). After the same reasoning the scored decision exceeds the model's reply on the inference tasks at every size and in a second model family, Qwen3-VL, and needs no parsing (Sections 5.3 and 5.4); a readout trained on the model's own reasoning lowers ECE to 0.07, at a cost in text score (Section 5.5). At 31B the scored decision is at or above the median hosted frontier model (thirteen on text, nine on images) on one of the seven tasks and within 0.02 of it on three more, and no reasoning setting of these models reaches its accuracy at a lower cost (Section 5.2).

2. **Two axes of intelligence.** The same examples give a one-pass decision and a reasoned one, so the classifier measures what a model decides in one pass and what it infers with reasoning separately. Reasoning moves the inference axis and leaves the classification axis within 0.03 at every size; scale moves both, more inference on text and more classification on images; across the thirteen frontier models the axes correlate at 0.90 on text and 0.39 on images (Section 5.7).

3. **Reinforcement learning of the reasoning with the scored decision as the reward.** The accuracy reward moves E4B by 0.01 to 0.03 on the text inference axis and by about 0.01 on the classification axis in three runs, with two to five times longer reasoning; the log-score reward makes the policy stop reasoning (Section 5.8). Planned experiments and their decision rules are listed in Section 5.10.

2 Background and Related Work

One-pass option scoring. Zhang [47] score options with a block attention mask and per-option position ids, so that one forward pass gives each option's exact log-likelihood; cross-entropy over the option set on 18 tasks transfers to 12 held-out tasks, with an LM-head or a learned scalar readout. We reuse the scorer, its tasks, its prompts and its metrics.

Reasoning and test-time compute. Chain of thought [40] and reinforcement learning with verifiable rewards [12, 21] improve accuracy on multi-step tasks. Budget forcing [27] controls test-time compute by ending or extending the thinking block; we do not constrain the reasoning length. Self-consistency [38] takes the majority answer over several sampled chains; a scored decision allows the option distributions to be averaged.

Vision-language models. Gemma 4 [10] reads images and text and reasons before replying through a switch in its chat template. Contrastive models classify images zero-shot by matching them to label names [31]. Generative vision-language models classify images less accurately than CLIP on standard benchmarks; scoring each class name by its likelihood narrows the gap without closing it, and Zhang et al. [48] attribute the gap mainly to how often a class occurs in training. CLIP models also exceed the vision-language models built on their image encoders at fine-grained multiple-choice classification [9], and, scoring each option by its summed log-probabilities, Cooper et al. [7] find contrastive models ahead on most object and scene recognition datasets and models coupled with a language model ahead on tasks that require reasoning and

outside knowledge. Recent models match or exceed CLIP-style models on several datasets [24], and part of the reported gap stems from noisy labels and from evaluation protocols that discard replies outside the class list [18].

Reinforcement learning of reasoning. REINFORCE with a leave-one-out baseline [1, 19, 41] needs no value model; dynamic sampling [44] keeps only prompts whose samples receive different rewards. Accuracy rewards are improper scoring rules and produce over-confident models [8]; Zhang [47] reproduce the effect in a one-pass scorer trained by REINFORCE. RLCR adds a Brier term to the reward of a model that verbalizes its confidence [8]. For language models fine-tuned with human feedback, verbalized confidence is typically better calibrated than the conditional probabilities [36], and reasoning models verbalize better-calibrated confidence than their non-reasoning counterparts [43]. CoT-decoding selects the decoding path with the largest mean margin between the two most probable tokens at the answer positions, a margin that is higher on paths containing a chain of thought [37]. DEER induces a trial answer at transition points of a reasoning model's chain of thought and stops the reasoning once the answer's token probabilities exceed a threshold [42]. VeriFree rewards a sampled reasoning trace with the likelihood of the reference answer conditioned on it [49], and RLPR with the mean token probability of that answer [45]; our reward normalizes the likelihoods of all options after the reasoning.

3 Method

Thinking. The task prompt of Zhang [47], without its cue line and followed by an instruction to reply with one listed answer, is the user message of the model's chat template, after the image for image tasks, with thinking enabled. The model samples its reasoning with the settings its developers ship for the mode until it closes its thinking block. A hosted deployment of the same open weights, with the providers pinned, generates the reasoning of Gemma 4 31B; vLLM on our hardware [20] generates it for the smaller models (Table 1). Gemma 4's model card recommends no output length; a cap of 32,768 tokens stops runaway chains and was reached by 23 of E2B's 18,396 chains, 9 of 26B-A4B's 12,022 and none of E4B's or 31B's. Qwen3-VL generates up to the output length its model card recommends, 32,768 tokens on text, 40,960 on images and 81,920 on GPQA Diamond, which 47 of the 2B model's 12,022 chains, 10 of the 4B model's and 15 of the 32B model's reached.

Scoring. After the closed block we append the cue line ("Answer:" or "Label:") and score each option o_i as a continuation, in one packed pass:

$$s_i = \sum_t \log P_\theta\left(o_{i,t} \mid x, r, c, o_{i,<t}\right), \qquad p_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)}, \tag{1}$$

where x is the chat-formatted prompt, r the reasoning and c the cue. With thinking disabled the template emits an empty thinking block, and the decision is the one-pass classifier on the post-trained model. For an image task, x contains the image's tokens. The reasoning is generated once; the scoring pass costs one prefill of the options, as in Zhang [47].

The option terminator. Zhang [47] render each option as a leading space, the option text and a scored newline. A post-trained model ends a one-option reply with its end-of-turn token: in the Qwen3 pilot (Appendix C), the newline terminator leaves less than 10^{-4} of the probability on the option set at every size from 0.6B to 32B, and the end-of-turn token at least 0.98. We score with the end-of-turn token.

Table 1: Models: parameters, the source of the reasoning, and their use.

model	parameters	reasoning generated by	use
Gemma 4 E2B, E4B	2B, 4B effective	vLLM	evaluation; RL (E4B)
Gemma 4 26B-A4B	26B, 4B active	vLLM	evaluation
Gemma 4 31B	31B	hosted (Friendli, Novita)	evaluation; trained readout
Qwen3-VL (Thinking)	2B, 4B, 32B	vLLM	evaluation, reasoning on only
Qwen3	0.6B to 32B	vLLM	pilot

Readouts. The LM-head readout (Eq. 1) needs no training. The scalar readout $s_i = w^\top h_i$ of Zhang [47] reads the final hidden state at option i's terminator; Section 5.5 trains it on the model's own reasoning with the option-set cross-entropy, the reasoning frozen.

The model's own reply. The baseline is the model used as its developers recommend: the generation continues past the end of the thinking block to the end of the turn, and the reply is matched to an option with the protocol of Zhang [47]; it returns no probability. The scored decision and the reply condition on the same reasoning.

Reinforcement learning of the reasoning. The policy is the model with a LoRA adapter [15] of rank 16. For each prompt, vLLM samples $G = 8$ reasonings from the current adapter; each is scored with Eq. 1, and sample j receives the reward

$$R_j = u(p^{(j)}, y) - w\left(L_j/s\right)^k, \tag{2}$$

where u is the scored accuracy $[\arg\max_i p_i^{(j)} = y]$, the gold option's probability $p_y^{(j)}$ or its logarithm, and L_j is the number of reasoning tokens. The length term is a cost per token, relative among a prompt's samples; generation has no budget beyond the safety cap. The gradient is REINFORCE on the reasoning tokens with the leave-one-out baseline $R_j - \frac{1}{G-1}\sum_{l\neq j} R_l$ [1, 19]: one update per batch of 64 prompts, no KL term, AdamW with a learning rate of 2×10^{-5} and the gradient clipped to norm 1. With dynamic sampling [44], 3×64 prompts are drawn and the first 64 whose samples' rewards differ are trained on. The readout shares the policy's weights, so an update also changes how options are scored after the reasoning. Every sample, its option scores and its reward are recorded at every step, and a fixed probe of unseen prompts is scored every 25 updates.

4 Experimental Setup

Models. Gemma 4 [10] E2B, E4B, 26B-A4B and 31B (Table 1); the Qwen3 text models from 0.6B to 32B [30] form the pilot. The Thinking checkpoints of Qwen3-VL [2] at 2B, 4B and 32B form a second model family; they have no reasoning-off mode and are evaluated with reasoning only. Scoring runs in bf16 on one NVIDIA RTX PRO 6000 (96 GB) or on a DGX Spark; Qwen3-VL 32B's reasoning on GPQA Diamond, up to 81,920 tokens, is scored on one H200 NVL (141 GB), the packed pass's attention mask at that length exceeding the Spark's memory.

Tasks. Seven held-out tasks whose answer must be inferred, with options that differ per example: BIG-Bench Hard [35] (its 23 subtasks with an option set, pooled), MMLU-Pro [39], MuSR [34] and GPQA Diamond [32] (the correct and three incorrect answers in a seeded order per question) as text, and MMMU [46] (the validation set, questions with one image), MathVista [26] (testmini, its multiple-choice questions) and

AI2D [17] with an image. As classification tasks, the twelve held-out text tasks of Zhang [47] and six held-out image tasks with the label names as options: EuroSAT [13], DTD [4], STL-10 [6], Oxford-IIIT Pets [29], Flowers-102 [28] and Rendered SST-2 [31]. Each task is evaluated on the first 500 examples of the seeded test permutation of Zhang [47], and GPQA Diamond on all its questions; at 31B, the seven inference tasks are in addition evaluated on their complete test sets (22,894 questions). An example whose options are not distinct is excluded (2 of GPQA Diamond's 198; 1 to 5 of the first 500 of MuSR, MMMU, MathVista and AI2D; 1 to 17 per complete test set), as are 3 of the 5,511 BIG-Bench Hard questions, in which the dataset divides the correct answer at a comma into several lettered options, so that no listed option is correct. One reasoning sample per example; E2B, E4B and 31B have two further reasoning seeds on the inference tasks, and E2B and E4B on the classification tasks.

Training mixtures. The trained readout and the classification RL runs use the training mixture of Zhang [47] (18 text tasks) and nine image tasks. The inference RL mixture is AQuA-RAT [22], LogiQA 2.0 [23] and ScienceQA [25] (with images), 5,000 or 6,000 training examples each. None of the seven inference tasks is in a training mixture, and no GPQA data is trained on. Two near neighbours are declared: MMLU-Pro extends MMLU, one of paper 1's held-out tasks, and ScienceQA (training) and AI2D (held out) both ask grade-school science questions over images.

References. Thirteen hosted frontier models, the eleven of Zhang [47], Claude Opus 5.5 and GPT-6.1 Sol, at their default reasoning settings, one free reply per example matched with the same protocol, on the first 100 examples per task (GPQA Diamond: all). The set is fixed as of 29 September 2026; models released later are not added. The cost per decision is the gateway's charge per request.

Metrics and statistics. As in Zhang [47]: macro-F1 where examples share a label set and accuracy where options differ per example (every inference task), with the task mean weighting tasks equally; NLL, Brier score, top-1 ECE with 15 bins and the AUROC of the confidence as an error detector. Differences between two predictions of the same examples carry a paired bootstrap 95% interval (1,000 draws, each resampling every task's examples) on the task mean or per task.

Compute. Hosted generation cost $31.30 for the 31B reasoning and replies ($4.02 for one sample per example on the subsets, $5.29 for the further samples, $4.06 for the cases and verification prompts, $14.39 for the complete test sets, $3.34 for the reasoning over the training mixture of the trained readout and $0.19 for pilots), $122.34 for the frontier models at their default settings, $45.49 for their settings sweep and $54.48 for the comparisons on 500 examples, and $10.75 for the reasoning of Qwen3-VL 235B-A22B on three text tasks, an evaluation stopped for the cost of its scoring. The job ledgers record 156 GPU-hours on the DGX Spark, 80 on Colab GPUs (RTX PRO 6000 or A100) and 16 on H100 80 GB, a lower bound, since some Colab runtimes ended before their ledgers were copied. Rented vast.ai GPUs cost $92.66: $6.06 for two RTX PRO 6000 instances (a pilot and the readout along the reasoning), $82.42 for the two A100s of the 26B-A4B reinforcement-learning run and $4.18 for the H200 that scored Qwen3-VL 32B on GPQA Diamond. By their logged step times the E4B reinforcement-learning runs alone took 36 GPU-hours.

Changes to the protocol. The following deviate from the conventions shared with Zhang [47] or were decided after results were seen. (i) The seven inference tasks were added after the first results showed a small gain from reasoning on the classification tasks; those results are kept and reported. (ii) The conventions ask for complete test sets. These are used for the 31B results on the inference tasks (Table 3); all other results use the 500-example subsets, and the frontier models the first 100 of them. (iii) The conventions ask for three seeds for every headline configuration. E2B, E4B and 31B have three reasoning seeds on the inference tasks,

and E2B and E4B on the classification tasks (Tables 4 and 11), and the inference-mixture reinforcement-learning run has three training seeds; 26B-A4B, 31B on the classification tasks, Qwen3-VL and the other reinforcement-learning runs have one. (iv) The reinforcement-learning runs were stopped or changed while running: the classification run and the inference run under the accuracy reward received a linear length cost ($w = 0.05$, $s = 1000$, $k = 1$) from steps 131 and 91, and both were stopped when a step took 35 minutes; the evaluated snapshots (steps 125 and 75) precede the change. (v) Images whose PNG encoding exceeds 10 MB are sent to the Anthropic models as JPEG at quality 95, the provider's size limit. (vi) Under the cases prompt (Section 5.3), the reply is also matched on its final line, a deviation from the reply protocol of Zhang [47] used with that prompt only. (vii) The frontier models' reasoning settings were swept on the inference tasks after their default-settings results were seen (Section 5.2). The gateway's versions of Qwen3-235B and DeepSeek V3.2 do not reason; Qwen3-235B's reasoning version is its family's "on" point, and DeepSeek V3.2's gave no reasoning either, so that family is one point. (viii) Claude Opus 5.5 and then GPT-6.1 Sol were added after all other results were seen, on the same examples and protocol. In a probe Opus 5.5's named reasoning settings did not change its reasoning (about 75 tokens per question on MMLU-Pro at every setting), so its family is one point; GPT-6.1 Sol's did, so its settings below the default were swept as in (vii). (ix) In the 26B-A4B reinforcement-learning run, each sample's gradient is taken on its first 16,000 tokens (the training pass of a 20,000-token sample exceeds an 80 GB GPU); every sampled token still counts in the reward, and the log records the samples cut (one by step 9).

5 Results

5.1 Reasoning helps where the answer must be inferred

On the seven inference tasks, reasoning raises the task mean of the scored decision by 0.150 [0.129, 0.172] at E2B, 0.152 [0.130, 0.171] at E4B, 0.158 [0.142, 0.176] at 26B-A4B and 0.144 [0.129, 0.160] at 31B on the 500-example subsets, and every task gains at every size (Table 2, Figure 1). Over three reasoning seeds the gain is 0.153, 0.151 and 0.146 at E2B, E4B and 31B, with a seed spread of at most 0.005 (Table 4). On the complete test sets at 31B, the gain is 0.140 [0.126, 0.153] (Table 3). It is largest on GPQA Diamond (0.510 to 0.832) and BIG-Bench Hard (0.765 to 0.930), and its interval includes zero on AI2D (+0.003 [−0.008, +0.013]). The accuracy on every task, with reasoning disabled and enabled, lies within the bootstrap interval of its subset. On the eighteen classification tasks the gain is 0.010 [0.000, 0.019], 0.024 [0.015, 0.031], 0.021 [0.014, 0.028] and 0.013 [0.007, 0.020], and six of them lose at 31B (Table 10 in Appendix A); over three reasoning seeds it is 0.006 at E2B and 0.020 at E4B (Table 11).

Averaged over the seven inference tasks, accuracy is 0.90 to 0.94 in the three shortest fifths of each task's reasoning and 0.658 [0.620, 0.693] in the longest, while the mean confidence of the scored decision is at least 0.99 in every fifth (Figure 2). This is consistent with longer reasoning marking harder questions; the length is not controlled. Per task, the most accurate length differs (Figure 3): about 1,600 tokens on GPQA Diamond and 640 on MuSR, and 170 to 450 tokens on the five other tasks. On MuSR the shortest reasoning is less accurate than the peak beyond the intervals; on GPQA Diamond the intervals overlap. Holding the question fixed removes these differences: with three reasoning samples per question (eight on GPQA Diamond), a question's longest sample is as accurate as its shortest on every task, −0.001 [−0.012, +0.010] on the task mean, while the median length of the longest sample is 1.5 to 2.8 times that of the shortest. The most accurate length of each task is therefore a property of its questions: harder questions draw longer reasoning, and a longer sample of the same question is not more accurate.

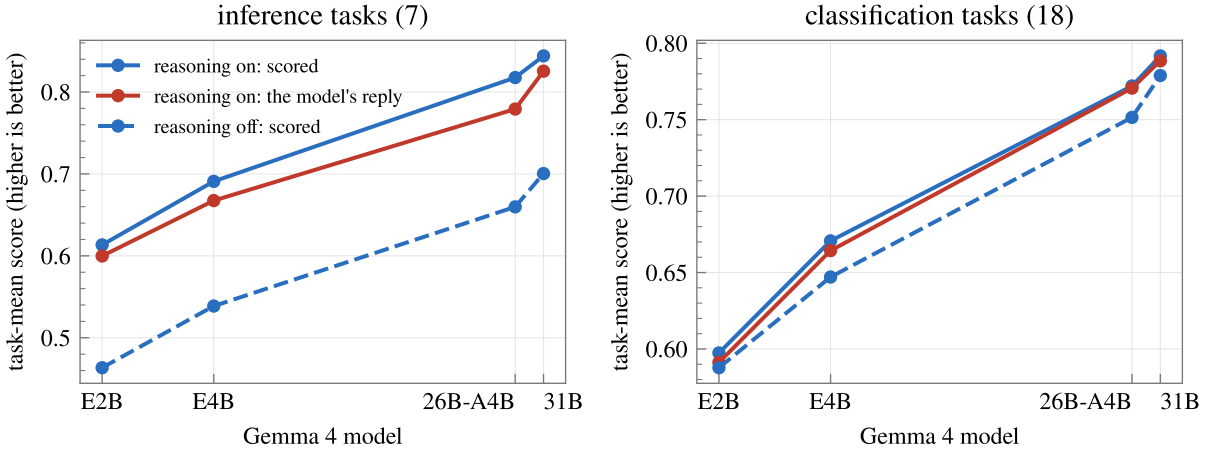


Figure 1: Reasoning raises the scored decision's task mean by about 0.15 on the inference tasks and by 0.01 to 0.02 on the classification tasks, at every Gemma 4 size; after the same reasoning the scored decision is above the model's reply. Primary metric (accuracy on the inference tasks; macro-F1 or accuracy on the classification tasks), 500 test examples per task (GPQA Diamond: 196), one reasoning sample; untrained LM head, end-of-turn terminator.

Table 2: Inference tasks: reasoning raises the scored decision's accuracy on every task at every size, and the scored decision exceeds the reply after the same reasoning on the task mean. Off, on: the scored decision with thinking disabled and enabled; reply: the model's reply after the same reasoning. Bottom rows: paired differences of the task mean with 95% bootstrap intervals. 500 test examples per task, GPQA Diamond 196; untrained LM head.

	Gemma 4 E2B			Gemma 4 E4B			Gemma 4 26B-A4B			Gemma 4 31B		
task	off	on	reply	off	on	reply	off	on	reply	off	on	reply
BIG-Bench Hard	0.532	0.816	0.814	0.584	0.854	0.856	0.676	0.922	0.920	0.746	0.938	0.936
MMLU-Pro	0.350	0.524	0.510	0.476	0.684	0.642	0.686	0.832	0.812	0.734	0.864	0.846
MuSR	0.549	0.613	0.613	0.521	0.609	0.611	0.585	0.695	0.695	0.595	0.679	0.679
GPQA Diamond	0.276	0.383	0.352	0.357	0.510	0.474	0.474	0.745	0.699	0.510	0.832	0.786
MMMU	0.438	0.570	0.552	0.521	0.653	0.620	0.640	0.800	0.733	0.703	0.816	0.796
MathVista	0.484	0.707	0.685	0.588	0.761	0.711	0.737	0.892	0.781	0.763	0.896	0.857
AI2D	0.615	0.681	0.673	0.723	0.766	0.758	0.820	0.838	0.814	0.852	0.884	0.878
mean	0.463	0.613	0.600	0.539	0.691	0.667	0.660	0.818	0.779	0.700	0.844	0.825
on − off	+0.150 [+0.129, +0.172]			+0.152 [+0.130, +0.171]			+0.158 [+0.142, +0.176]			+0.144 [+0.129, +0.160]		
on − reply	+0.014 [+0.008, +0.019]			+0.023 [+0.016, +0.031]			+0.038 [+0.030, +0.047]			+0.019 [+0.011, +0.026]		

Table 3: Gemma 4 31B on the complete test sets of the inference tasks: reasoning raises the task mean by 0.140, and after the same reasoning the scored decision exceeds the reply by 0.015. Off, on: the scored decision with thinking disabled and enabled; reply: the model's reply after the same reasoning. Paired differences with 95% bootstrap intervals; untrained LM head, one reasoning sample.

task	examples	off	on	reply	on − off	on − reply
BIG-Bench Hard	5,507	0.765	0.930	0.929	+0.165 [+0.154, +0.176]	+0.001 [+0.000, +0.001]
MMLU-Pro	12,031	0.720	0.867	0.850	+0.148 [+0.140, +0.155]	+0.017 [+0.015, +0.019]
MuSR	752	0.604	0.689	0.688	+0.085 [+0.057, +0.117]	+0.001 [+0.000, +0.004]
GPQA Diamond	196	0.510	0.832	0.786	+0.321 [+0.255, +0.393]	+0.046 [+0.000, +0.087]
MMMU	800	0.695	0.821	0.814	+0.126 [+0.097, +0.155]	+0.008 [-0.003, +0.019]
MathVista	537	0.764	0.894	0.866	+0.130 [+0.095, +0.168]	+0.028 [+0.013, +0.045]
AI2D	3,071	0.877	0.880	0.876	+0.003 [-0.008, +0.013]	+0.003 [+0.001, +0.006]
mean		0.705	0.845	0.830	+0.140 [+0.126, +0.153]	+0.015 [+0.008, +0.022]

Table 4: Three reasoning seeds: the gain from reasoning and the difference from the reply hold over seeds at E2B, E4B and 31B. Task means over the seven inference tasks, mean ± population standard deviation over three reasoning seeds; the decision with reasoning off is one pass and has no seed. 500 test examples per task, GPQA Diamond 196; untrained LM head.

size	off	on	reply	on − off	on − reply
E2B	0.463	0.616 ± 0.005	0.605 ± 0.007	0.153 ± 0.005	0.012 ± 0.002
E4B	0.539	0.690 ± 0.003	0.666 ± 0.004	0.151 ± 0.003	0.024 ± 0.001
31B	0.700	0.846 ± 0.005	0.826 ± 0.004	0.146 ± 0.005	0.021 ± 0.002

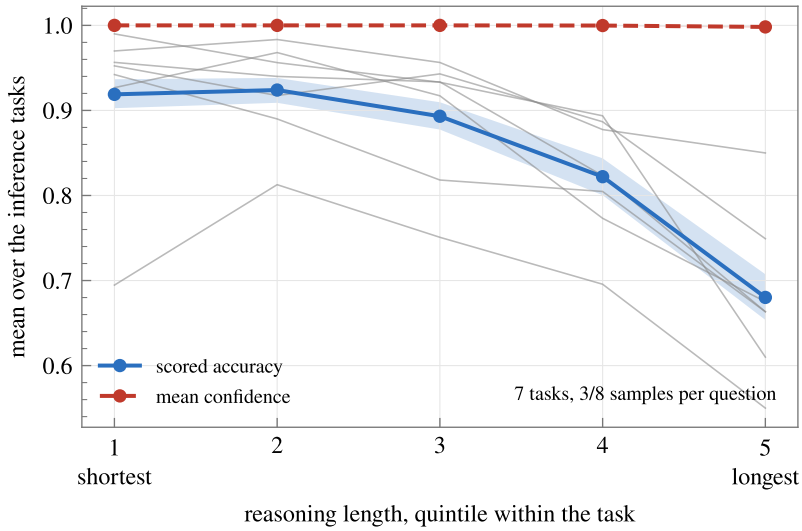


Figure 2: Gemma 4 31B, inference tasks: accuracy falls from about 0.9 to 0.66 in the longest fifth of each task's reasoning, while the scored decision's confidence stays at or above 0.99. Every reasoning sample of a task binned by length quintile within the task (median lengths from 146 to 1,100 tokens on BIG-Bench Hard and from 654 to 5,729 on GPQA Diamond); blue and red: the mean over the seven tasks with its 95% bootstrap interval, resampling questions; grey: each task's accuracy; untrained LM head.

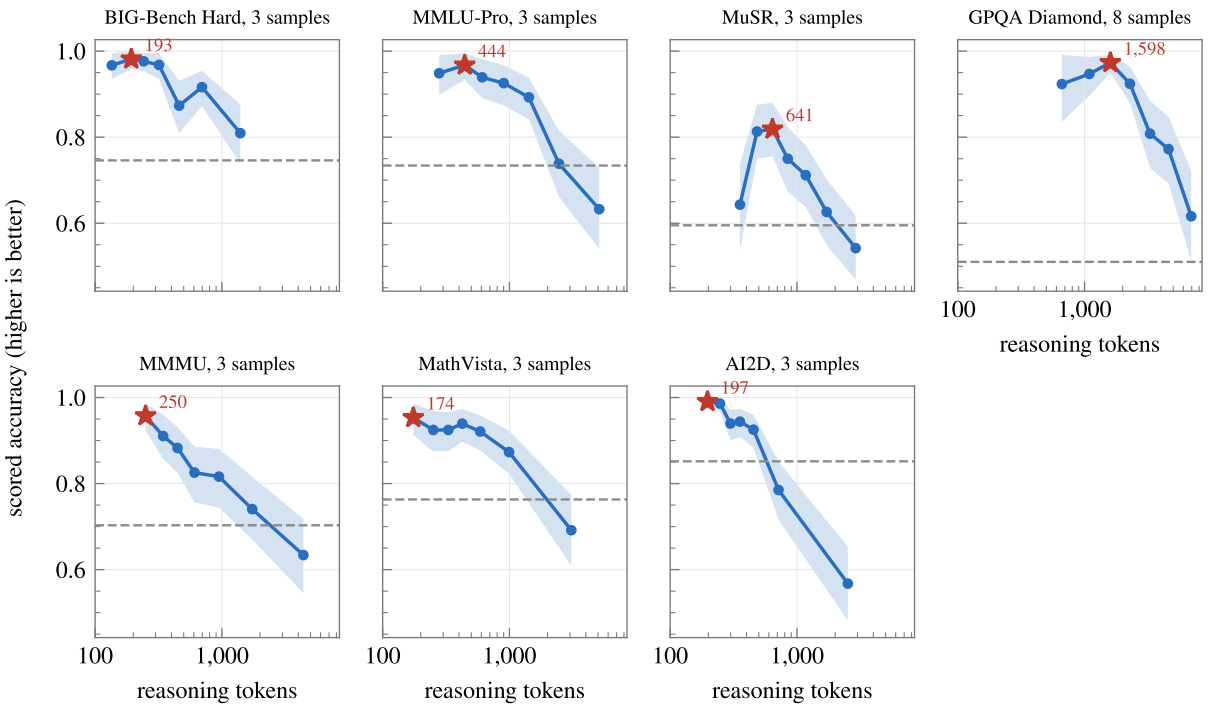


Figure 3: Gemma 4 31B: the reasoning length at which the scored decision is most accurate differs by task, about 1,600 tokens on GPQA Diamond and 640 on MuSR against 170 to 450 on the five other tasks. Every reasoning sample binned by length (about 200 samples per bin, at least four bins; three samples per question, eight on GPQA Diamond), each bin at its median length with its 95% bootstrap interval, resampling questions; star: the most accurate bin; dashed: the scored accuracy with reasoning off; untrained LM head. The bins compare different questions; within a question, longer samples are not more accurate (Section 5.1).

Table 5: Gemma 4 31B against the frontier settings nearest its Pareto curve, on the same 500 test examples per task (GPQA Diamond: all 196): task-mean accuracy of the matched replies, the paired difference of our scored decision with its 95% bootstrap interval, and the mean hosted charge per decision (ours: the reasoning request; the scoring pass is not charged).

	text (4 tasks)			images (3 tasks)		
model, setting	accuracy	ours − theirs	USD	accuracy	ours − theirs	USD
Gemma 4 31B, scored (ours)	0.828		0.00081	0.865		0.00054
Kimi K3, minimal	0.847	−0.018 [-0.037, +0.001]	0.00532	0.899	−0.034 [-0.050, -0.018]	0.00729
GPT-5, low	0.819	+0.009 [-0.009, +0.028]	0.00641	0.861	+0.004 [-0.013, +0.020]	0.00453
GPT-6 Astra, default	0.909	−0.081 [-0.097, -0.064]	0.00767	0.938	−0.072 [-0.088, -0.056]	0.00830

5.2 Against the frontier at equal cost

Per task, the interval of the 31B scored decision overlaps that of the best frontier model on BIG-Bench Hard, MMMU, MathVista and AI2D, and lies below it on MMLU-Pro, MuSR, GPQA Diamond and both task means (Figure 4). At their default settings the frontier models differ in how much they reason: Qwen3-235B, DeepSeek V3.2 and Claude Sonnet 4.6 do not reason on these tasks, GPT-5 and Kimi K3 reason for about 1,500 tokens per question and Muse Spark 1.3 for about 3,000. Figure 5 therefore sweeps each model family over its named reasoning settings, and our model over test-time compute. On the text inference tasks the scored decision reaches 0.840 at $0.00077 per decision. No setting of any frontier family reaches that accuracy at a lower cost: at about the same cost, GPT-5 at its minimal setting scores 0.582 and GLM-5.3 without reasoning 0.776; the cheapest settings above it, GPT-6.1 Sol at its minimal and low settings, score 0.906 and 0.911 at 1.3 times the cost. On the image inference tasks the scored decision reaches 0.857 at $0.00045; the cheapest settings above it, GPT-6.1 Sol at its low and minimal settings, score 0.903 and 0.910 at 2.2 times the cost. The most accurate point on both is Claude Opus 5.5 at its default setting, 0.932 on text and 0.933 on images, at 7.6 and 15 times our cost. Averaging our option distributions over three reasoning samples raises the task means to 0.852 and 0.863 at three times the cost (Section 5.6), below GPT-6.1 Sol's cheapest points at lower cost. The points carry no intervals; at 100 examples per task, a task mean's 95% interval is several points wide (Figure 4), so points within 0.03 of each other are not ordered. On 500 examples per task, the frontier settings nearest the curve when the comparison was planned (Kimi K3 at its minimal setting, GPT-5 at low and GPT-6 Astra at its default) cost 6.6 to 9.5 times as much per decision on text and 8.4 to 15 times on images (Table 5). The 31B scored decision is level with GPT-5 at low (text +0.009 [−0.009, +0.028], images +0.004 [−0.013, +0.020]) and with Kimi K3 on text (−0.018 [−0.037, +0.001]), and below Kimi K3 on images (−0.034 [−0.050, −0.018]) and GPT-6 Astra on both (−0.081 and −0.072). The two Muse Spark 1.3 settings planned were refused by the provider after the first task. None of these settings is cheaper than the scored decision, so the comparison leaves the cost claim of Figure 5 unchanged.

5.3 The scored decision against the model's reply

After the same reasoning, the scored decision exceeds the reply on the inference-task mean by 0.014 [0.008, 0.019] at E2B, 0.023 [0.016, 0.031] at E4B, 0.038 [0.030, 0.047] at 26B-A4B and 0.019 [0.011, 0.026] at 31B on the 500-example subsets (Table 2; over three seeds 0.012, 0.024 and 0.021 at E2B, E4B and 31B, Table 4), by 0.015 [0.008, 0.022] at 31B on the complete test sets (Table 3), and on the classification tasks by 0.001 to 0.006. At 26B-A4B the difference is largest on MathVista (0.892 against 0.781) and MMMU (0.800 against 0.733), where replies that name no option account for 0.014 and 0.018 of it. On the complete test sets, the per-task interval excludes zero on MMLU-Pro (+0.017), MathVista (+0.028) and AI2D (+0.003), reaches it on GPQA Diamond (+0.046 [+0.000, +0.087]) and includes it on MMMU (+0.008), whose subset interval excluded it; on BIG-Bench Hard and MuSR, the two decisions differ on 4 of 5,507 and 1 of 752 questions.

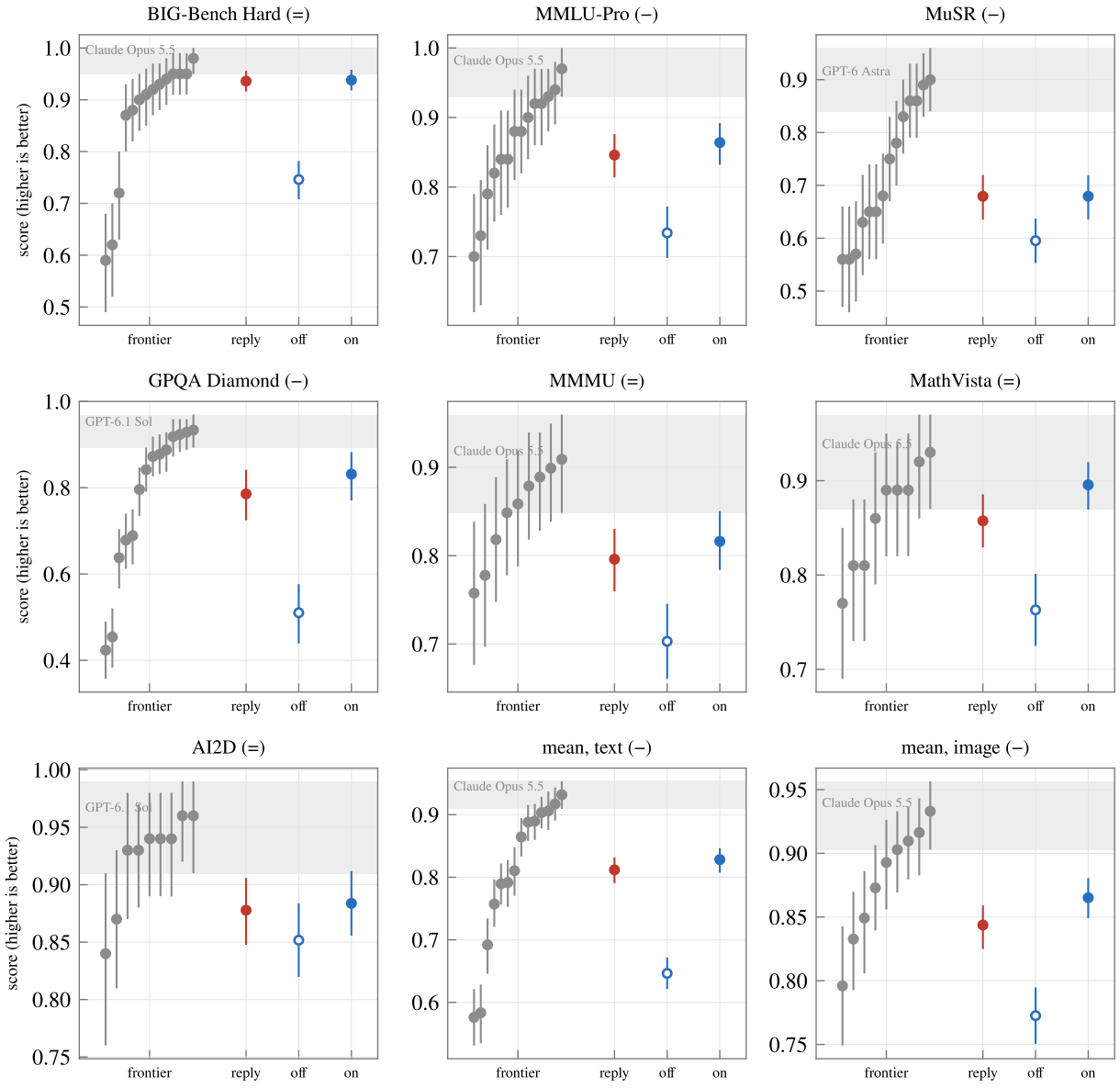


Figure 4: Gemma 4 31B on each inference task against thirteen hosted frontier models (grey, in rank order; the best named, its interval shaded): its reply after reasoning (red), the scored decision without (open) and with (filled) reasoning. Title marks: our reasoning-on interval lies above (+), overlaps (=) or lies below (−) the best model's. Ours on 500 test examples per task, the frontier models on the first 100 (GPQA Diamond: all 196 for both); 95% bootstrap intervals.

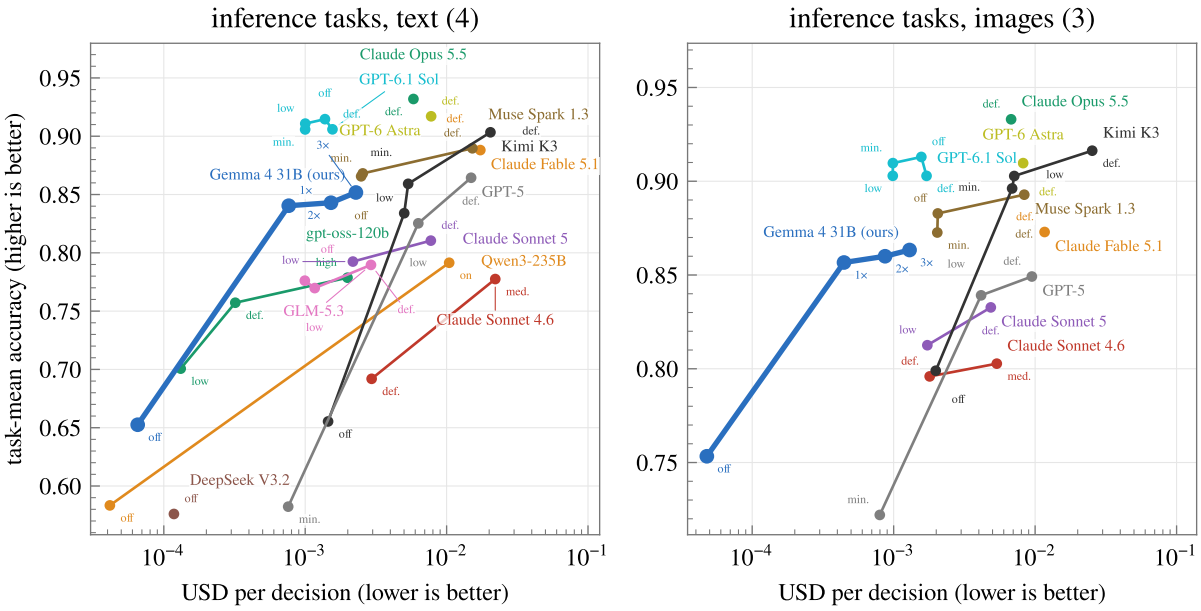


Figure 5: Pareto curves on the inference tasks. Each frontier family is one line through its named reasoning settings (point labels: def. default, min. minimal, low, med. medium, high, off, on; the gateway maps each to the provider's control, with no token budget); Gemma 4 31B (ours, blue) is one line through test-time compute: reasoning off, then the mean of the option distributions of k reasoning samples ($k\times$), each sample charged. Task-mean accuracy on the first 100 test examples (GPQA Diamond: 196) against the mean hosted charge per decision (September 2026 prices); our scoring pass, one prefill of the options on our hardware, is not charged.

Table 6: Qwen3-VL (Thinking checkpoints, reasoning on), inference tasks: the scored decision and the reply after the same reasoning; the share of replies that name no listed option. 500 test examples per task, GPQA Diamond 196; untrained LM head.

	Qwen3-VL 2B		Qwen3-VL 4B		Qwen3-VL 32B	
task	scored	reply	scored	reply	scored	reply
BIG-Bench Hard	0.804	0.784	0.894	0.868	0.918	0.918
MMLU-Pro	0.596	0.464	0.712	0.630	0.824	0.794
MuSR	0.583	0.567	0.637	0.635	0.673	0.671
GPQA Diamond	0.474	0.342	0.638	0.541	0.755	0.724
MMMU	0.632	0.572	0.717	0.675	0.796	0.778
MathVista	0.821	0.759	0.861	0.797	0.914	0.900
AI2D	0.752	0.752	0.806	0.804	0.860	0.858
mean	0.666	0.606	0.752	0.707	0.820	0.806
scored − reply	+0.061 [+0.049, +0.072]		+0.045 [+0.036, +0.055]		+0.014 [+0.009, +0.018]	
no option named	0.084		0.042		0.013	

On MMLU-Pro, 127 of the 12,031 replies (1.1%) name no listed option, against 8 of GPQA Diamond's 196 (4.1%) and at most 0.5% on the other tasks. The two decisions agree on 88% (GPQA Diamond) to 99.9% (BIG-Bench Hard and MuSR) of the examples. On GPQA Diamond the difference holds on each of eight reasoning samples per question (0.046 to 0.087), whose single-sample scored accuracy ranges from 0.832 to 0.872 (mean 0.852); the reported sample is the lowest of the eight.

A prompt that asks for the strongest case for each option before the decision (the cases prompt) lowered the scored accuracy at both sizes on the four classification tasks furthest from the frontier: the mean is 0.557 with the plain prompt, 0.540 with the cases prompt and 0.540 with the cases of the five options the one-pass readout ranks highest at E4B, and 0.626, 0.611 and 0.603 at 31B, with two to three times the reasoning (Table 13 in Appendix B). Under this prompt E4B repeats its cases in the visible reply, and the matcher of Zhang [47], which takes the longest option the reply names, falls up to 0.112 below the scored decision; matching the reply's final line brings it within 0.028 on every task. At 31B at most 7% of replies name a second option. The scored decision needs no parsing.

5.4 A second model family

The Thinking checkpoints of Qwen3-VL always reason, so the comparison is between the scored decision and the reply after the same reasoning (Table 6; classification tasks in Table 12). On the seven inference tasks the scored decision exceeds the reply by 0.061 [0.049, 0.072] at 2B, 0.045 [0.036, 0.055] at 4B and 0.014 [0.009, 0.018] at 32B. The difference shrinks with size as replies that name no listed option become rarer: 8.4, 4.2 and 1.3% of the replies (MMLU-Pro at 2B: 17.8%; GPQA Diamond at 2B: 24.5%); at most 2.2% name two or more options. On GPQA Diamond the scored decision reaches 0.474, 0.638 and 0.755, against 0.342, 0.541 and 0.724 for the reply. On the eighteen classification tasks the difference is 0.010 [0.008, 0.013], 0.003 [0.002, 0.005] and 0.001 [0.000, 0.003]. From 2B to 32B the scored decision rises from 0.666 to 0.752 and 0.820 on the seven inference tasks and from 0.651 to 0.708 and 0.750 on the classification tasks.

5.5 Calibration after reasoning

Reasoning raises the NLL of the untrained LM-head readout at every size on both task groups (Table 7): on the inference tasks from 4.23 to 9.33 at E2B, 2.35 to 5.58 at E4B, 2.59 to 4.77 at 26B-A4B and 2.67 to 3.88 at 31B. It lowers AUROC in every cell. ECE falls with reasoning on the inference tasks at E2B, 26B-A4B (0.295 to 0.182) and 31B (0.267 to 0.156) and rises or stays in the five other cells.

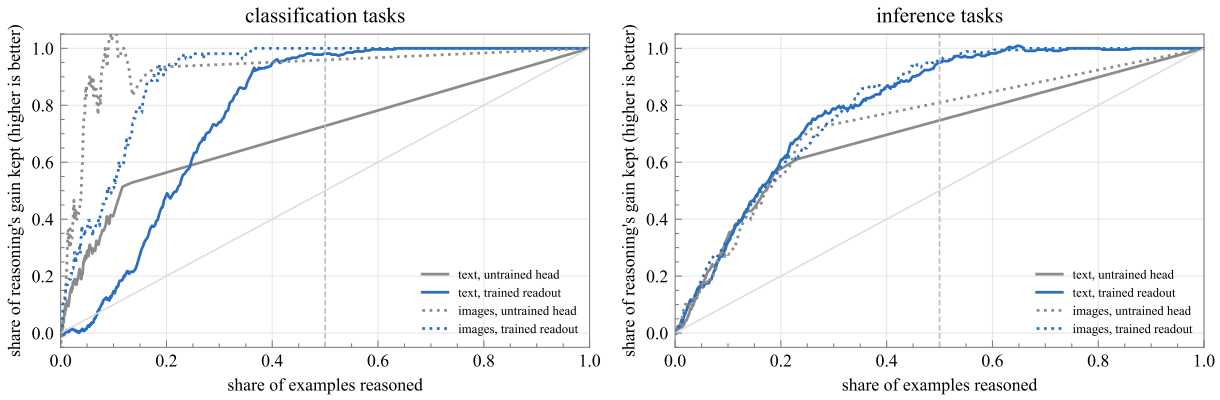


Figure 6: Reasoning when unsure. Gemma 4 31B: the share of reasoning's gain kept against the share of examples reasoned as the threshold on the one-pass top probability sweeps from 0 to 1, for the untrained head and the trained readout, on the classification tasks (left) and the inference tasks (right), text (solid) and images (dotted); the dashed line marks half of the examples, the rule fixed in advance.

The scalar readout trained on the 31B model's own reasoning over the classification training mixture (one seed) lowers ECE with reasoning from 0.195 to 0.069 on the twelve text tasks and from 0.162 to 0.067 on the six image tasks, and raises AUROC from 0.704 to 0.768 and from 0.812 to 0.881 (Table 8). It costs text score: −0.044 [−0.055, −0.030] without reasoning and −0.015 [−0.024, −0.005] with it; on the image tasks the score rises by 0.010 [0.002, 0.020] with reasoning. Calibration after reasoning therefore comes from the trained readout, at a cost in text score. On the six inference tasks the hypothesis that this readout lowers ECE without lowering accuracy beyond the interval holds with reasoning on and fails without it. With reasoning on, its accuracy equals the untrained head's (0.846 against 0.846; paired difference −0.000 [−0.006, 0.006], no task beyond its interval) while ECE falls from 0.06–0.32 to 0.03–0.19, NLL from 1.5–9.0 to 0.2–0.9, and AUROC rises on every task (MuSR 0.573 to 0.708, MMLU-Pro 0.773 to 0.897). With reasoning off it costs 0.020 [0.006, 0.035] of accuracy (MuSR gains 0.064, the other five lose 0.034 to 0.064) for ECE of 0.04–0.15 against 0.14–0.39. A readout trained on classification reasoning serves the inference tasks once the reasoning is there; the one-pass decision on them would need inference reasoning in training.

Reasoning when unsure. The one-pass decision comes with a confidence, so reasoning can be spent only where that confidence is low: a threshold on the one-pass top probability sends the examples below it to reasoning and keeps the one-pass decision for the rest. Figure 6 sweeps the threshold and plots the share of reasoning's gain the policy keeps against the share of examples it reasons on. The rule fixed in advance reads the curve at half of the examples. Under the trained readout, reasoning on half keeps 0.95 of the gain on the text inference tasks (0.822 against 0.680 with reasoning off and 0.829 always on) and 0.97 on the image inference tasks (0.858 against 0.744 and 0.862), at 80 and 86% of the hosted charge of always reasoning (the one-pass request on every example and the reasoning request on the gated ones); on the classification tasks it keeps 0.98 and 1.00. The untrained head cannot reach half: its top probability exceeds 0.99 on most examples, and at the 18% it does send to reasoning it keeps 0.51 of the gain on text. Under the trained readout, the oracle that reasons exactly where the one-pass decision is wrong would reason on 26 to 32% of the inference examples.

Table 7: Reasoning raises NLL and lowers AUROC of the untrained readout at every size. Task means over the inference tasks (7) and the classification tasks (held-out, 18), thinking disabled (off) and enabled (on); untrained LM head, end-of-turn terminator.

		NLL		ECE		AUROC	
size	tasks	off	on	off	on	off	on
E2B	inference	4.225	9.328	0.427	0.386	0.670	0.612
E2B	held-out	4.566	9.254	0.343	0.370	0.732	0.644
E4B	inference	2.347	5.575	0.301	0.306	0.715	0.599
E4B	held-out	2.076	3.373	0.240	0.273	0.753	0.617
26B-A4B	inference	2.590	4.772	0.295	0.182	0.770	0.654
26B-A4B	held-out	2.249	5.517	0.197	0.201	0.797	0.701
31B	inference	2.672	3.875	0.267	0.156	0.799	0.746
31B	held-out	2.465	4.726	0.185	0.184	0.796	0.740

Table 8: Gemma 4 31B: the scalar readout trained on the model's own reasoning lowers ECE to 0.07 and raises AUROC after reasoning, and costs 0.015 of text score with reasoning and 0.044 without. Task means of the classification tasks; difference: trained minus untrained, paired 95% bootstrap interval; one seed.

		primary metric			ECE		AUROC	
tasks	reasoning	untrained	trained	difference	untrained	trained	untrained	trained
text (12)	off	0.767	0.722	−0.044 [-0.055, -0.030]	0.201	0.071	0.773	0.758
	on	0.779	0.765	−0.015 [-0.024, -0.005]	0.195	0.069	0.704	0.768
images (6)	off	0.803	0.809	+0.006 [-0.006, +0.017]	0.152	0.063	0.844	0.865
	on	0.817	0.827	+0.010 [+0.002, +0.020]	0.162	0.067	0.812	0.881

5.6 Averaging reasoning samples

Averaging the option distributions of several reasoning samples requires no training. On the first ten steps of the E4B accuracy-reward run (policy at or near the base model; 64 prompts of the classification training mixture, eight samples each), one sample is correct on 0.653 of the prompts, the mean of two distributions on 0.664 and the mean of eight on 0.672, against 0.664 for the majority of eight replies; on 256 unseen prompts, two samples raise 0.646 to 0.652. These are training-mixture prompts. On the held-out inference tasks at 31B, with three reasoning samples per example, the result is a null. The task mean of the single samples is 0.846, 0.843 and 0.851, a spread of 0.003; the mean of the three distributions gains 0.005 [−0.001, 0.010] over the first sample, and choosing the sample with the highest confidence under the trained readout gains 0.004 [−0.004, 0.011]. After the full reasoning the untrained readout puts at least 0.99 on its top option for 99.8% of the questions, so the mean of three distributions acts as a vote among three nearly one-hot decisions. Resampling only the 10% of questions on which the trained readout's top probability is lowest reaches the accuracy of three samples (0.851 against 0.846) at 1.2 times the cost of one; that gain lies within its interval. On GPQA Diamond, eight samples per question score 0.832 to 0.872 alone (mean 0.852); the mean of all eight distributions scores 0.847, and the mean of three, averaged over every subset of three seeds, 0.859. Under the trained readout the mean of three keeps ECE between 0.03 and 0.16, so aggregation keeps the distribution calibrated; at this size it does not raise accuracy (Figure 8).

A verifying second pass. The readout's top-2 recall on the inference tasks is far above its accuracy (MuSR 0.940 against 0.679, GPQA Diamond 0.913 against 0.832), so we asked the model to reason once more over the two most probable options of its first pass, with the cases prompt of Appendix B, and scored the options after it. The second pass alone scores 0.850 on the six inference tasks against the first pass's 0.846, at 1.4 to

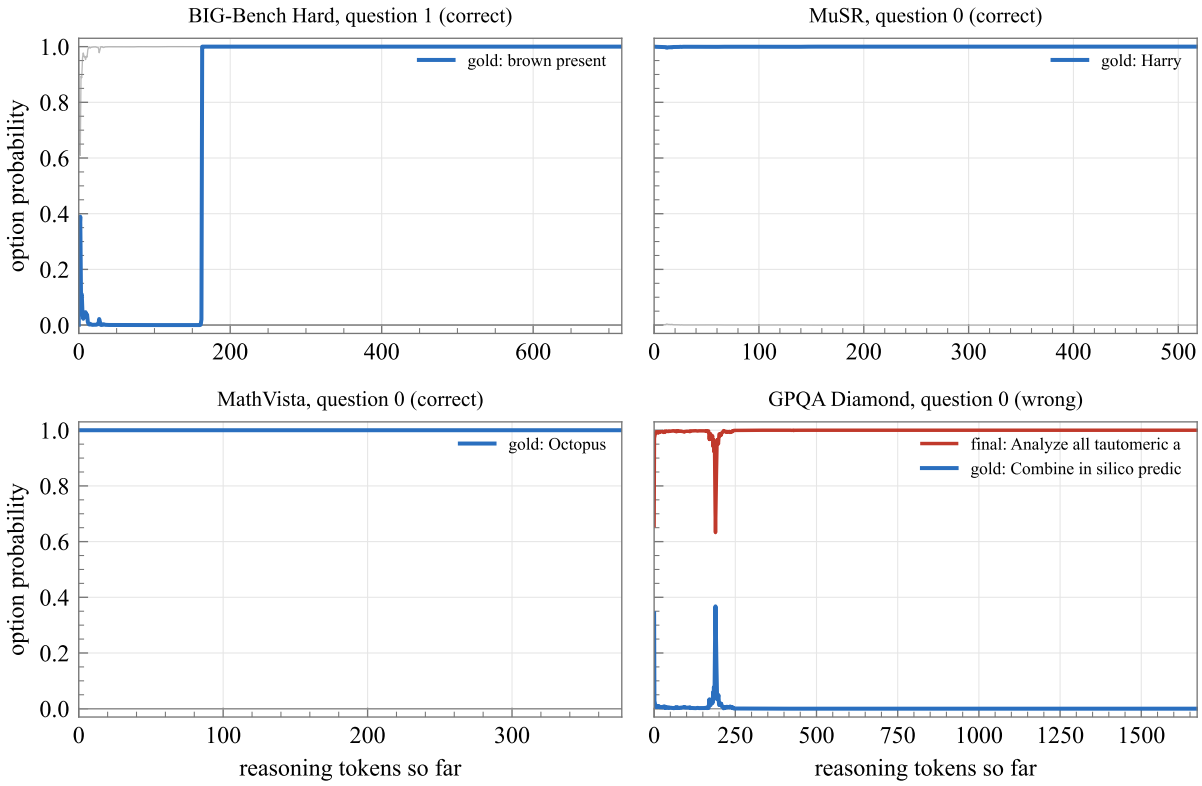


Figure 7: Gemma 4 31B: each option's probability under the untrained readout with the reasoning stopped after its first t tokens, one question per task (the first in test order whose reasoning has 300 to 2,000 tokens); blue: the correct option; red: the decision after the full reasoning where it is incorrect; grey: the other options. The last point is the evaluation's score of the full reasoning.

1.9 times the reasoning tokens, and 0.832 on GPQA Diamond against 0.832 at 1.5 times; the mean of the two distributions 0.849 and 0.827. Gated on the trained readout's confidence, no task gains beyond its interval. The errors that remain after one pass are ones the model repeats when asked to choose between its own two best answers.

The readout along the reasoning. For one stored reasoning per task (the first in test order with 300 to 2,000 tokens), we stopped the reasoning after each of its tokens, closed the thinking block and scored the options as in the evaluation (Figure 7). On MuSR and MathVista the readout is at the correct option from the first token; on BIG-Bench Hard it moves to the correct option at 165 tokens; on GPQA Diamond it holds an incorrect option throughout, the correct one reaching 0.37 at about 190 tokens. In each case the distribution is nearly one-hot within the first 200 tokens. Four questions describe when the readout commits; they do not estimate how often.

5.7 Two axes of capability

Factor analyses of benchmark scores find a general factor and a reasoning factor across models [3, 16, 33]. Our tasks separate the two by construction, and the scored decision measures both on the same examples: Figure 9 places every model by its task-mean score on the classification tasks and on the inference tasks. Across the thirteen frontier models the two correlate at 0.90 on text (12 and 3 tasks; R^2 of the fitted line

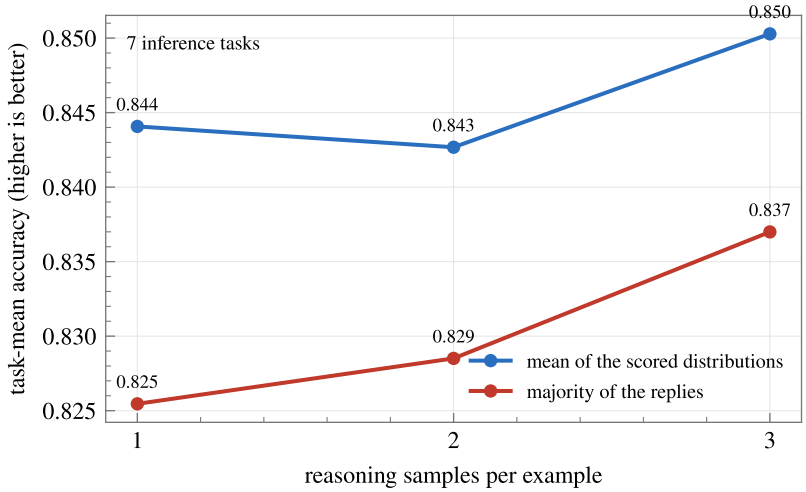


Figure 8: Gemma 4 31B, inference tasks: task-mean accuracy against the number of reasoning samples, for the mean of the scored distributions and the majority of the replies.

0.80) and at 0.39 on images (6 and 3 tasks; R^2 0.15); Kimi K3 and GLM-5.3 sit above the text line and Claude Sonnet 4.6 below it, beyond their bootstrap intervals. A decomposition of the frontier scores (each task standardised over the models) gives one dominant component, 0.56 of the variance on text [0.44, 0.76] over models and 0.43 on images, and a second whose loadings do not follow the inference–classification split (sign agreement 0.67 on text, permutation $p = 0.40$; 0.78 on images, $p = 0.06$): across models the two axes are one factor; they separate within a model, under an intervention. With GPQA Diamond as a fourth text inference task (the RL runs were not evaluated on it), the correlation stays 0.90 and reasoning's difference of shifts 0.13 to 0.17. Within Gemma 4, on the same examples, enabling reasoning raises the inference axis by 0.135 to 0.189 on text and 0.093 to 0.140 on images and the classification axis by at most 0.026, and the paired difference of the two shifts excludes zero at every size. Size moves both axes: from E4B to 31B with reasoning on, inference gains more on text (0.111 against 0.066) and classification more on images (0.231 against 0.139). The three accuracy-reward RL runs of Section 5.8 move the text inference axis by 0.012 to 0.021 and the classification axis by 0.005 to 0.014, the cases-prompt run measured from the untrained model under the same prompt, whose own shifts are within their intervals (0.001 and 0.009); the paired difference of the two shifts is within its interval for every run, and on images every shift is within its interval. The rejection-sampling run of Section 5.8 moves neither text axis (−0.003 and 0.008, within their intervals) and loses 0.023 [0.011, 0.035] on image classification. Over resamples of the tasks themselves (each axis's tasks drawn with replacement, examples fixed) and under every single-task drop, reasoning's difference of shifts stays positive at each size on text and images, and the frontier correlation on text stays within 0.86 and 0.93; the size ladder's difference of shifts and the RL runs' do not resolve at this number of tasks. Qwen3-VL's ladder with reasoning on moves both axes at each step, text inference by 0.087 [0.066, 0.113] and classification by 0.055 [0.043, 0.065] from 2B to 4B, 0.057 and 0.049 from 4B to 32B. The run whose log-score reward made the policy stop reasoning (Section 5.8) moves the other way: with thinking enabled it scores 0.534 on the text inference axis against 0.716 before training, a shift of −0.181 [−0.209, −0.155], and −0.032 on the classification axis; on images −0.126 and −0.014. Reasoning acts on one axis and its removal undoes the same axis; size acts on both, and RL with the accuracy reward at E4B moves both by about a point.

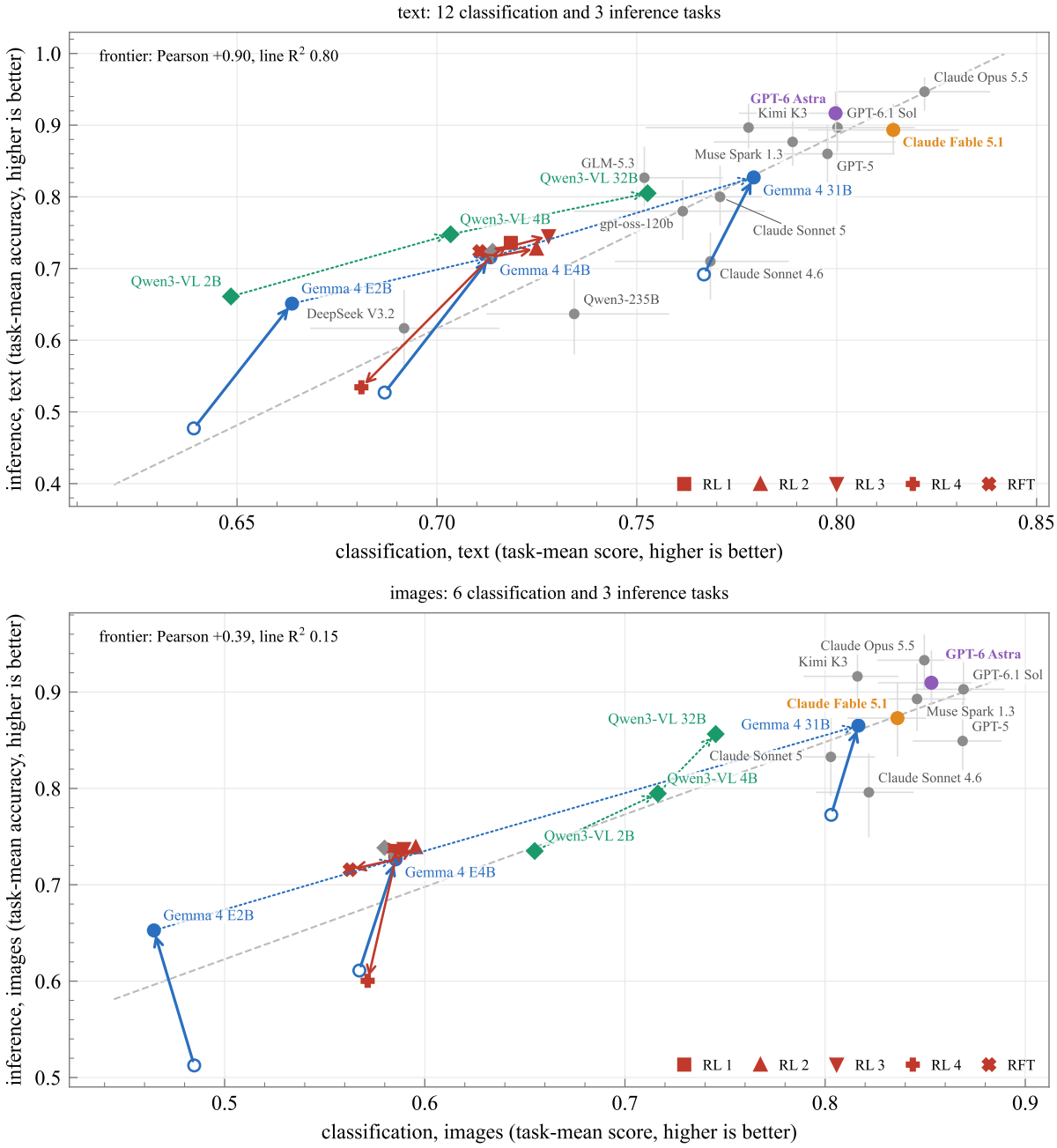


Figure 9: Two axes of capability. Task-mean score on the held-out classification tasks against task-mean accuracy on the held-out inference tasks, text (12 and 3 tasks) and images (6 and 3), 95% bootstrap intervals over examples; the hosted frontier models at their default settings on 100 examples per task, with the least-squares line; Gemma 4 on 500 examples, reasoning off (hollow) to on (filled), the sizes with reasoning on (dotted), the untrained E4B under the cases prompt (grey diamond), and the four E4B RL runs (red; RL 1 the inference mixture at 75 updates, RL 2 the classification mixture at 125, RL 3 the classification mixture under the cases prompt at 200, drawn from the cases-prompt point, RL 4 the log-score reward, whose policy stopped reasoning), rejection-sampling fine-tuning on RL 1's rollouts (red cross, RFT), and Qwen3-VL's sizes with reasoning on (green diamonds, joined). Highlighted: Claude Fable 5.1 (orange) and GPT-6 Astra (purple); on text Fable scores higher on classification (0.814 against 0.800) and GPT-6 Astra on inference (0.917 against 0.893), on either side of the line within their intervals.

5.8 Reinforcement learning of the reasoning

At E4B, 64 prompts and eight samples per step (Eq. 2), one seed each except where stated; one run at 26B-A4B closes this section:

- **Log-score reward.** The policy stops reasoning: the reasoning falls from about 250 tokens per sample to none by step 25, and the probe accuracy falls by 0.03 (Figure 10). After 200 updates the held-out task mean is 0.026 [0.019, 0.033] below the untrained model's with reasoning, and NLL and ECE are 1.78 and 0.209 against 3.37 and 0.273 (Table 9). On the six inference tasks, which it did not train on, it reasons for 51 tokens per example and scores 0.154 [0.136, 0.172] below the untrained model. Since the untrained readout is over-confident after reasoning (Section 5.5), we interpret the collapse as the proper reward preferring the less confident decision without reasoning.
- **Accuracy reward, classification mixture.** After 125 updates the held-out task mean is 0.011 [0.004, 0.019] higher, the reasoning 5.5 times longer (252 to 1,387 tokens), NLL 3.43 and ECE 0.281. The mean over six inference tasks, which the run did not train on, is 0.012 [0.000, 0.026] higher.
- **Accuracy reward, inference mixture.** After 75 updates the task mean over six inference tasks is 0.014 [0.000, 0.028] higher, with reasoning 857 to 1,591 tokens; the classification mean is unchanged (+0.003 [−0.004, 0.010]). Two more training seeds give 0.019 [0.006, 0.032] and 0.021 [0.008, 0.035] on the inference tasks (three-seed mean 0.018, s.d. 0.003) with reasoning of 1,603 and 1,742 tokens, and 0.003 and 0.001 on the classification mean.
- **Accuracy reward, cases prompt.** Rewarding a sample 0 unless it argues every shortlisted option gives the same held-out gain as the plain prompt, 0.011 [0.003, 0.019], and 0.020 [0.006, 0.035] on the inference tasks. This run is evaluated under the cases prompt and the untrained model under the plain prompt, so its difference combines training and prompt. The untrained model under the cases prompt scores 0.724 on the three text inference tasks against 0.716 under the plain prompt and 0.745 for this run, and 0.739 on the three image tasks against 0.727 and 0.737: on text the prompt gives 0.008 and the training 0.021 more; on images the prompt gives the whole difference.

GPQA Diamond was not evaluated for the RL runs. Under the accuracy reward the reasoning grew on both mixtures until a step took 35 minutes (about 3,400 tokens per sample), and both runs were stopped. The per-task held-out gain does not track the base model's reasoning length (rank correlation −0.09 for the cases run and +0.28 for the plain run), so the gain did not come from making the model reason on tasks it answered without reasoning. Each gain is about the size of that from averaging two reasoning samples without training (Section 5.6).

Selection is not the gain. Rejection-sampling fine-tuning on the inference-mixture run's own rollouts, one correct chain per prompt drawn at random from its first 75 steps (4,417 of 4,800 prompts), trained by the same update with every chain at advantage one, the same optimiser, batch and 75 updates and no sampler in the loop, scores 0.720 on the six inference tasks against 0.721 untrained (−0.001 [−0.015, 0.013]) and 0.735 after RL, which exceeds it by 0.016 [0.001, 0.030]. Its reasoning is 1,503 tokens per example, near RL's 1,591, since the chains came from the policy as it lengthened. The gain comes from the on-policy updates, not from the chains they selected.

At 26B-A4B. The accuracy-reward recipe with a cubic length cost ($(L/2{,}000)^3$ for L reasoning tokens), the gradient of each sample taken on its first 16,000 tokens (protocol change (ix)), ran for 31 of a planned 50 updates on the inference mixture. The held-out probe fell from 0.808 to 0.797 by 25 updates while the reasoning grew from 2,401 to 3,063 tokens per sample, and the run was stopped. Its adapter at 25 updates

Table 9: Reinforcement learning at E4B, one seed each: the accuracy reward raises the held-out task mean by about 0.01 with longer reasoning; the log-score reward lowers it and removes the reasoning. After RL: task mean of the primary metric with the run's own reasoning; − untrained: paired difference from the untrained model with reasoning and the plain prompt, 95% bootstrap interval; 500 test examples per task.

run	tasks	after RL	− untrained	reasoning tokens	NLL	ECE	AUROC
accuracy reward, classification mixture, 125 updates	inference (6)	0.733	+0.012 [+0.000, +0.026]	1965	4.518	0.265	0.692
accuracy reward, classification mixture, 125 updates	held-out (18)	0.682	+0.011 [+0.004, +0.019]	1387	3.432	0.281	0.688
accuracy reward, inference mixture, 75 updates	inference (6)	0.735	+0.014 [+0.000, +0.028]	1591	5.362	0.264	0.606
accuracy reward, inference mixture, 75 updates	held-out (18)	0.674	+0.003 [-0.004, +0.010]	509	3.886	0.289	0.591
log-score reward, classification mixture, 200 updates	inference (6)	0.567	−0.154 [-0.172, -0.136]	51	2.314	0.299	0.731
log-score reward, classification mixture, 200 updates	held-out (18)	0.645	−0.026 [-0.033, -0.019]	0	1.775	0.209	0.760
accuracy reward, cases prompt, 200 updates	inference (6)	0.741	+0.020 [+0.006, +0.035]	1596	5.070	0.259	0.658
accuracy reward, cases prompt, 200 updates	held-out (18)	0.682	+0.011 [+0.003, +0.019]	906	4.310	0.294	0.659

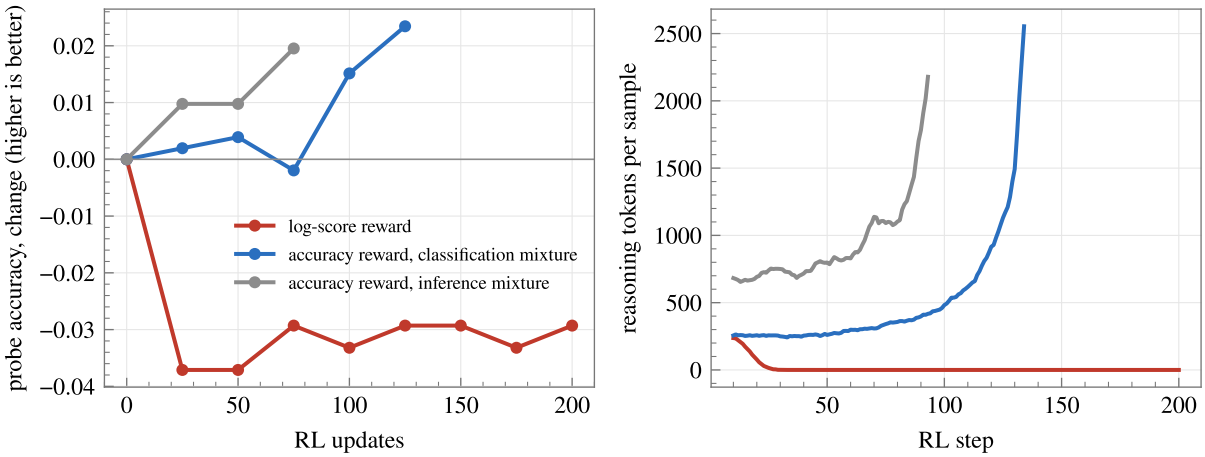


Figure 10: Reinforcement learning at E4B: rewarded by the log score, the policy stops reasoning within 25 steps and its probe accuracy falls; rewarded by accuracy, the probe accuracy rises by about 0.02 while the reasoning grows without bound. Left: change in scored accuracy on each run's fixed probe (256 unseen prompts of its training mixture, two samples each) since the start; right: reasoning tokens per sample, mean over ten steps.

scores 0.834 on the six inference tasks against 0.830 untrained, +0.005 [−0.006, 0.015], at 3,366 reasoning tokens per example against 2,592: the lengthening of the E4B runs at a stronger base model, without their gain.

5.9 Reasoning length under reinforcement learning

Figure 11 follows the reasoning length of every training task through each E4B run. The reward sets the direction. Under the log score, reasoning falls to zero on all 27 tasks of the classification mixture within about 25 steps and stays near zero for the remaining 175. Under the accuracy reward it grows on every task of both mixtures: between the first and the last ten steps, 2.9 (HellaSwag) to 35 times (CLINC150) over the 134 steps of the classification-mixture run, 2.6 to 3.7 times over the 93 steps of the inference-mixture run, and 1.1 to 2.7 times over the 200 steps of the cases-prompt run, whose reasoning argues each shortlisted option. The rate differs more between tasks than between task types: in the classification-mixture run, image tasks grow 6.6 to 17 times, multiple-choice questions 2.9 to 25 times and classification tasks 3.2 to 35 times. The accuracy reward carries no cost for length, and the log score penalizes the over-confident decision that follows reasoning (Section 5.5); the two directions are consistent with these rewards.

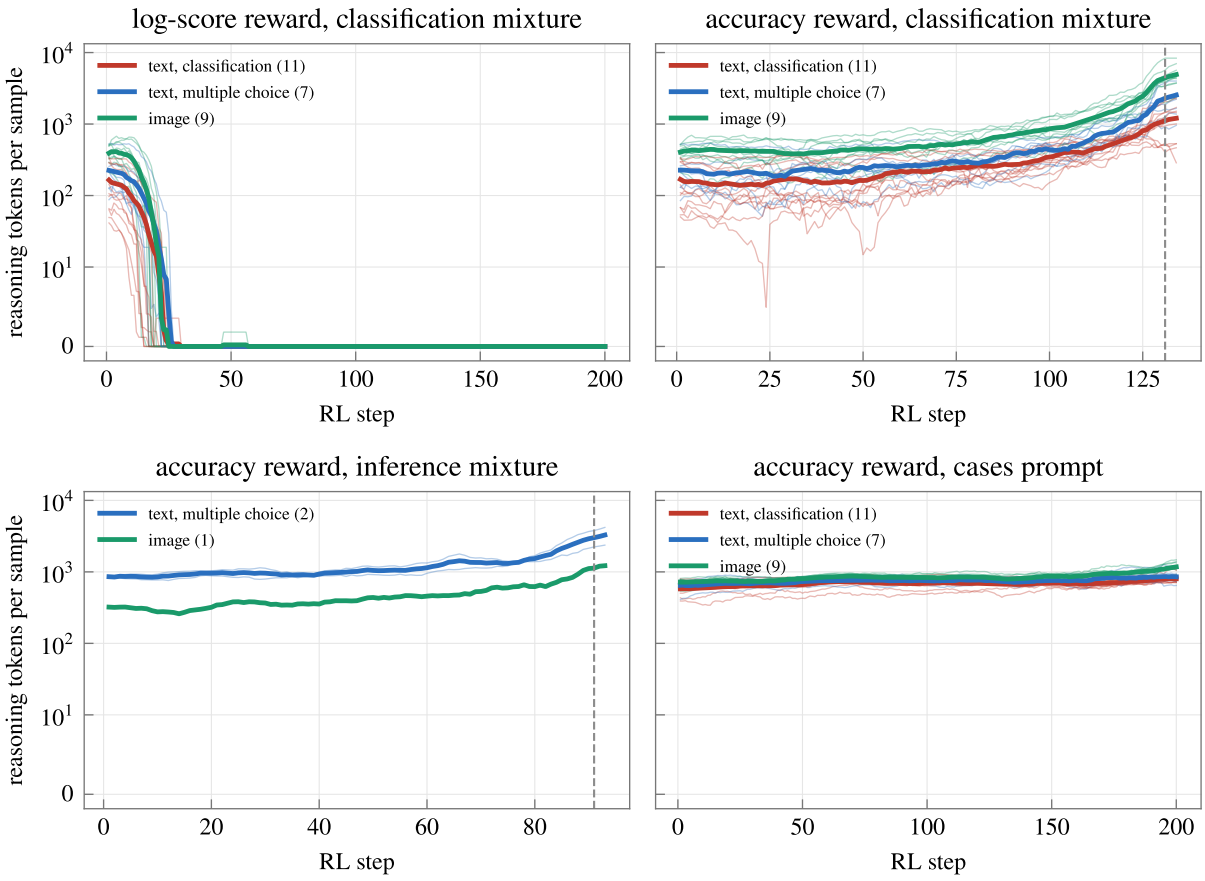


Figure 11: Reasoning length through reinforcement learning at E4B, per training task: the log-score reward drives it to zero on every task, and the accuracy reward makes it grow on every task, at rates that differ by task. Mean reasoning tokens per sample at each step, mean over ten steps (symmetric log scale, linear below 10 tokens); thin lines: tasks, coloured by type; bold: the type means; dashed: the step from which a linear length penalty applied.

5.10 Planned experiments

Each is judged by the inference-task mean and the image tasks; the classification mean must not fall beyond its interval, and the reasoning length on the classification tasks is a secondary measure.

A. **The log-score reward on a calibrated readout**: the LM head divided by the cross-fitted temperature of the thinking-on reports, and the trained scalar readout as the reward model. Confirmed if the policy keeps reasoning (more than half the samples at step 50) with rising reward.

B. **Rejection-sampling fine-tuning**: refuted (Section 5.8).

C. **Test-time aggregation** on the held-out tasks (Section 5.6).

D. **Where the gain comes from**: refuted (Section 5.8).

E. **The inference mixture alone**, evaluated on all tasks: done at E4B (Table 9).

F. **Three seeds** of the accuracy-reward recipe at E4B: done for the inference mixture (above).

6 Limitations

Apart from the 31B results on the inference tasks, which use the complete test sets, the evaluation uses 500-example subsets and the frontier models 100 examples per task; one reasoning sample and one seed per configuration, except three reasoning seeds on the inference tasks at E2B, E4B and 31B and on the classification tasks at E2B and E4B, and three training seeds of the inference-mixture reinforcement-learning run. Reinforcement learning ran at E4B; the one run at 26B-A4B was stopped at 31 updates without a gain (Section 5.8). The 31B reasoning comes from hosted deployments at the providers' undisclosed precision and is re-tokenized from text. The gain from reasoning is measured in Gemma 4 only, since Qwen3-VL, the second family, has no reasoning-off mode; a Qwen3 text pilot precedes both. The cost comparison uses the gateway's prices at one date and excludes our scoring pass; latency was not measured under a common protocol. The benchmarks may have been seen during pretraining by any of the models compared.

7 Conclusion

Completing the model's reasoning before scoring the options raises the one-pass scorer's accuracy by about 0.15 on tasks whose answer must be inferred, at every Gemma 4 size tested, and little on classification tasks. After the same reasoning, the scored decision is at least as accurate as the model's own reply, in Gemma 4 and Qwen3-VL, and returns a distribution; at 31B no reasoning setting of the hosted frontier models reaches its accuracy on the inference tasks at a lower cost. The distribution after reasoning is over-confident under the untrained readout; a trained readout calibrates it at a cost in text score. Reinforcement learning with the scored decision as the reward has so far given gains of about 0.01.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *ACL*, 2024.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

[7] Avi Cooper, Keizo Kato, Chia-Hsien Shih, Hiroaki Yamane, Kasper Vinken, Kentaro Takemoto, Taro Sunagawa, Hao-Wei Yeh, Jin Yamanaka, Ian Mason, and Xavier Boix. Rethinking VLMs and LLMs for image classification. *Scientific Reports*, 15:19692, 2025.

[8] Mehul Damani et al. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

[9] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or European swallow? benchmarking large vision-language models for fine-grained object classification. In *EMNLP*, 2024.

[10] Gemma Team. Gemma 4. https://huggingface.co/google/gemma-4-31B-it, 2026.

[11] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *TACL*, 2021.

[12] Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

[13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[15] Edward J Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[16] David Ilić and Gilles E. Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024.

[17] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[18] Nikita Kisel, Illia Volkov, Klára Janoušková, and Jiří Matas. Multimodal large language models as image classifiers. *arXiv preprint arXiv:2603.06578*, 2026.

[19] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

[20] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 2023.

[21] Nathan Lambert et al. Tülu 3: Pushing frontiers in open language model post-training. In *COLM*, 2025.

[22] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *ACL*, 2017.

[23] Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. LogiQA 2.0: An improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[24] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. Revisiting MLLMs: An in-depth analysis of image classification abilities. *arXiv preprint arXiv:2412.16418*, 2024.

[25] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

[26] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

[27] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[29] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[30] Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[32] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *COLM*, 2024.

[33] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[34] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *ICLR*, 2024.

[35] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.

[36] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*, pages 5433–5442, 2023.

[37] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In *NeurIPS*, 2024.

[38] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[39] Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks*, 2024.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[41] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[42] Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *ICLR*, 2026.

[43] Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *NeurIPS*, 2025.

[44] Qiying Yu et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv:2503.14476*, 2025.

[45] Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. RLPR: Extrapolating RLVR to general domains without verifiers. *arXiv preprint arXiv:2506.18254*, 2025.

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

[47] Max Zhang. Any language model is a classifier: Calibrated one-pass decisions over arbitrary label sets from pretrained LLMs, 2026. Preprint.

[48] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *NeurIPS*, 2024.

[49] Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. In *ICLR*, 2026.

A Classification tasks

Table 10: Classification tasks: reasoning changes the task mean by 0.01 to 0.02. Columns as in Table 2; primary metric per task; 500 test examples per task; untrained LM head.

	Gemma 4 E2B			Gemma 4 E4B			Gemma 4 26B-A4B			Gemma 4 31B		
task	off	on	reply	off	on	reply	off	on	reply	off	on	reply
Banking77	0.597	0.639	0.631	0.657	0.698	0.698	0.721	0.782	0.783	0.776	0.817	0.817
MASSIVE	0.711	0.689	0.679	0.790	0.776	0.768	0.818	0.806	0.804	0.869	0.868	0.868
SST-5	0.461	0.445	0.397	0.515	0.475	0.490	0.556	0.561	0.561	0.524	0.542	0.542
TweetEval emotion	0.740	0.718	0.718	0.715	0.731	0.755	0.742	0.780	0.780	0.773	0.754	0.754
TweetEval hate	0.661	0.662	0.659	0.733	0.690	0.692	0.586	0.701	0.701	0.728	0.696	0.696
Rotten Tomatoes	0.880	0.845	0.841	0.910	0.902	0.902	0.934	0.936	0.936	0.936	0.938	0.938
SMS Spam	0.629	0.725	0.720	0.683	0.753	0.697	0.856	0.903	0.895	0.909	0.935	0.935
CounterFactual	0.809	0.815	0.815	0.722	0.827	0.852	0.771	0.813	0.813	0.737	0.791	0.791
GoEmotions	0.269	0.303	0.294	0.239	0.265	0.259	0.253	0.310	0.310	0.327	0.292	0.292
ARC-Challenge	0.782	0.826	0.822	0.918	0.914	0.902	0.964	0.968	0.962	0.968	0.972	0.970
MMLU	0.630	0.698	0.684	0.756	0.812	0.784	0.858	0.884	0.876	0.890	0.918	0.918
StrategyQA	0.502	0.601	0.600	0.605	0.719	0.714	0.747	0.778	0.779	0.765	0.830	0.825
EuroSAT	0.348	0.337	0.330	0.339	0.398	0.374	0.460	0.441	0.438	0.577	0.577	0.525
DTD	0.467	0.336	0.336	0.613	0.588	0.555	0.680	0.671	0.671	0.679	0.716	0.716
STL-10	0.967	0.956	0.954	0.966	0.973	0.973	0.986	0.987	0.987	0.988	0.987	0.987
Oxford-IIIT Pets	0.195	0.193	0.195	0.395	0.442	0.443	0.856	0.817	0.817	0.873	0.858	0.858
Flowers-102	0.089	0.086	0.085	0.219	0.211	0.211	0.788	0.789	0.790	0.755	0.791	0.791
Rendered SST-2	0.843	0.880	0.880	0.872	0.900	0.888	0.950	0.970	0.970	0.946	0.972	0.972
mean	0.588	0.597	0.591	0.647	0.671	0.664	0.751	0.772	0.771	0.779	0.792	0.788
on − off	+0.010 [+0.000, +0.019]			+0.024 [+0.015, +0.031]			+0.021 [+0.014, +0.028]			+0.013 [+0.007, +0.020]		
on − reply	+0.006 [+0.003, +0.009]			+0.006 [+0.002, +0.011]			+0.001 [+0.001, +0.002]			+0.003 [+0.002, +0.004]		

Table 11: Classification tasks over three reasoning seeds at E2B and E4B: columns as in Table 4; primary metric per task; 500 test examples per task; untrained LM head.

size	off	on	reply	on − off	on − reply
E2B	0.588	0.593 ± 0.003	0.588 ± 0.003	0.006 ± 0.003	0.006 ± 0.001
E4B	0.647	0.667 ± 0.003	0.661 ± 0.003	0.020 ± 0.003	0.006 ± 0.000

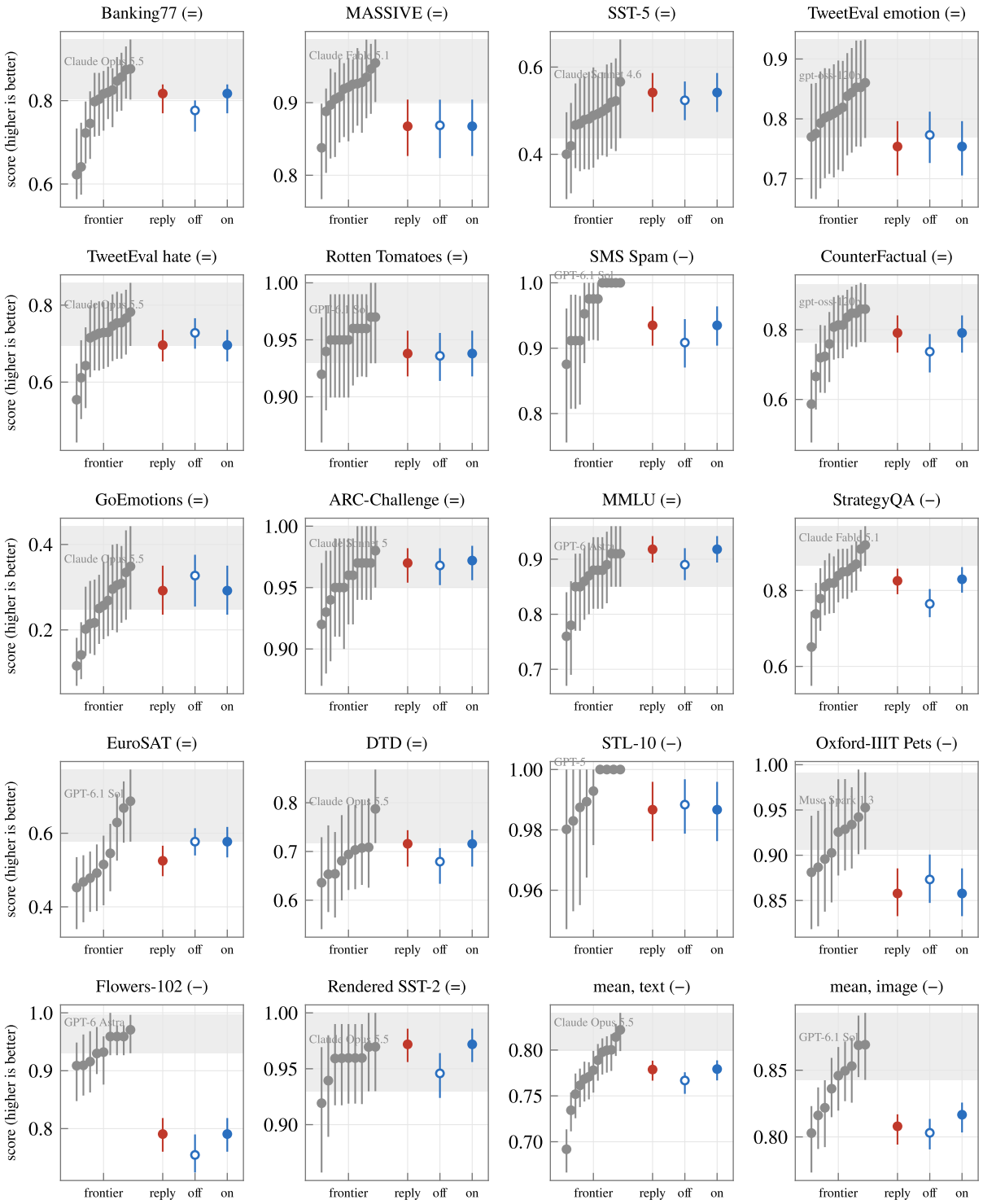


Figure 12: Gemma 4 31B on each classification task against the thirteen hosted frontier models, as in Figure 4.

Table 12: Qwen3-VL (Thinking checkpoints, reasoning on), classification tasks: columns as in Table 6; primary metric per task; 500 test examples per task; untrained LM head.

	Qwen3-VL 2B		Qwen3-VL 4B		Qwen3-VL 32B	
task	scored	reply	scored	reply	scored	reply
Banking77	0.621	0.618	0.732	0.725	0.782	0.781
MASSIVE	0.728	0.727	0.818	0.818	0.868	0.868
SST-5	0.380	0.380	0.432	0.434	0.475	0.475
TweetEval emotion	0.765	0.765	0.780	0.780	0.789	0.789
TweetEval hate	0.525	0.525	0.554	0.554	0.722	0.722
Rotten Tomatoes	0.887	0.886	0.912	0.912	0.924	0.924
SMS Spam	0.472	0.472	0.653	0.653	0.746	0.744
CounterFactual	0.799	0.799	0.821	0.821	0.796	0.796
GoEmotions	0.280	0.285	0.279	0.279	0.292	0.292
ARC-Challenge	0.920	0.850	0.938	0.932	0.958	0.956
MMLU	0.746	0.658	0.816	0.784	0.892	0.884
StrategyQA	0.658	0.655	0.706	0.701	0.788	0.788
EuroSAT	0.408	0.402	0.455	0.457	0.347	0.334
DTD	0.599	0.599	0.603	0.593	0.654	0.654
STL-10	0.990	0.989	0.988	0.987	0.993	0.993
Oxford-IIIT Pets	0.632	0.614	0.706	0.700	0.813	0.813
Flowers-102	0.409	0.410	0.620	0.621	0.720	0.724
Rendered SST-2	0.891	0.891	0.926	0.926	0.946	0.946
mean	0.651	0.640	0.708	0.704	0.750	0.749
scored − reply	+0.010 [+0.008, +0.013]		+0.003 [+0.002, +0.005]		+0.001 [+0.000, +0.003]	
no option named	0.009		0.003		0.002	

B The cases prompt

Table 13: The cases prompt at E4B (top) and 31B (bottom), four classification tasks: the scored decision needs no parsing, while the reply matched as in Zhang [47] falls below it when the reply names several options; the reply's final line restores agreement. Last column: share of replies naming two or more options. Cases, top 5: the cases of the five options the one-pass readout ranks highest.

task	prompt	scored	reply, paper 1	reply, final line	≥ 2 options named
Banking77	plain	0.718	0.718	0.718	0.00
	cases, all options	0.726	0.656	0.724	0.16
	cases, top 5	0.728	0.740	0.740	0.00
GoEmotions	plain	0.310	0.282	0.282	0.00
	cases, all options	0.278	0.230	0.286	0.28
	cases, top 5	0.282	0.268	0.282	0.11
SST-5	plain	0.498	0.518	0.518	0.00
	cases, all options	0.456	0.348	0.484	0.36
	cases, top 5	0.446	0.334	0.472	0.38
TweetEval hate	plain	0.704	0.708	0.708	0.00
	cases, all options	0.702	0.680	0.696	0.30
	cases, top 5	0.704	0.700	0.698	0.30

task	prompt	scored	reply, paper 1	reply, final line	≥ 2 options named
Banking77	plain	0.832	0.832	0.832	0.00
	cases, all options	0.826	0.818	0.826	0.02
	cases, top 5	0.822	0.764	0.822	0.07
GoEmotions	plain	0.340	0.340	0.340	0.00
	cases, all options	0.332	0.332	0.332	0.00
	cases, top 5	0.296	0.294	0.294	0.00
SST-5	plain	0.604	0.604	0.604	0.00
	cases, all options	0.578	0.576	0.580	0.01
	cases, top 5	0.596	0.598	0.598	0.00
TweetEval hate	plain	0.730	0.730	0.730	0.00
	cases, all options	0.710	0.708	0.710	0.03
	cases, top 5	0.698	0.690	0.696	0.03

C Pilot: Qwen3

The pilot used Qwen3 from 0.6B to 32B on StrategyQA [11], MMLU [14] and ARC-Challenge [5], 200 test examples each. Reasoning raises the task mean of the scored decision by 0.04 to 0.10 at every size (Figure 13). The newline terminator leaves less than 10^{-4} of the probability on the option set at every size and the end-of-turn token at least 0.98, and the end-of-turn token scores at least as high in every cell (Table 14). At 4B, reasoning lowers ECE and NLL on every task, but the mean confidence stays above 0.99 against an accuracy of 0.72 to 0.90, a cross-fitted temperature of 7 to 17 brings ECE to at most 0.052, and AUROC falls on every task (Table 15).

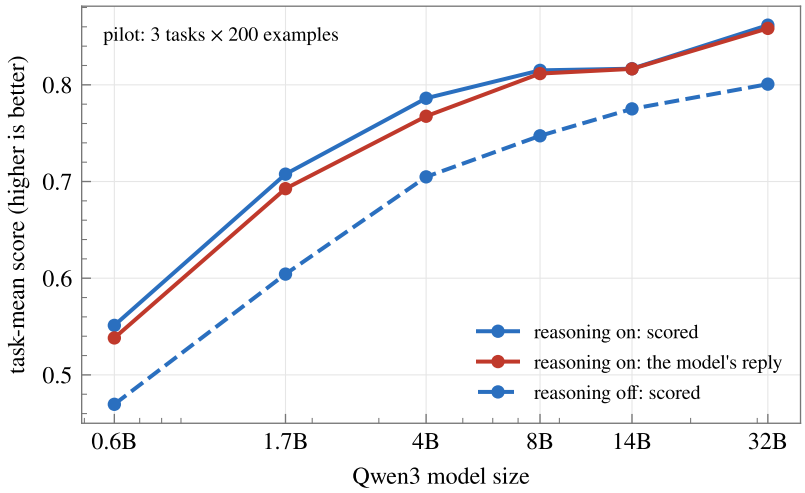


Figure 13: Pilot: reasoning raises the score of the scored decision at every Qwen3 size. Task-mean score (accuracy on ARC-Challenge and MMLU, macro-F1 on StrategyQA), 200 test examples per task, one reasoning sample; untrained LM head, end-of-turn terminator.

Table 14: Pilot: the end-of-turn terminator places nearly all probability on the option set and scores at least as high as the newline terminator at every Qwen3 size, with reasoning off and on. Task-mean score (as in Figure 13) and the median over the 600 examples of the probability on the option set; untrained LM head.

	reasoning off				reasoning on			
	score		option mass		score		option mass	
Qwen3	newline	end of turn	newline	end of turn	newline	end of turn	newline	end of turn
0.6B	0.437	0.470	8×10^{-8}	0.983	0.542	0.551	7×10^{-8}	0.994
1.7B	0.595	0.604	5×10^{-8}	1.000	0.693	0.708	3×10^{-8}	1.000
4B	0.700	0.705	6×10^{-8}	1.000	0.783	0.786	5×10^{-8}	1.000
8B	0.746	0.747	2×10^{-8}	1.000	0.807	0.815	8×10^{-8}	1.000
14B	0.774	0.775	1×10^{-9}	1.000	0.789	0.817	1×10^{-8}	1.000
32B	0.757	0.801	2×10^{-8}	0.999	0.814	0.862	2×10^{-5}	0.994

Table 15: Pilot: after reasoning, Qwen3-4B is more accurate and less miscalibrated but remains over-confident, and its confidence separates correct from wrong decisions less well. ECE (T): after a cross-fitted temperature. Untrained LM head, end-of-turn terminator, 200 test examples per task; †accuracy, otherwise macro-F1.

	reasoning off					reasoning on				
task	score	ECE	ECE (T)	NLL	AUROC	score	ECE	ECE (T)	NLL	AUROC
ARC-Challenge†	0.845	0.154	0.080	2.091	0.769	0.900	0.099	0.052	1.972	0.734
MMLU†	0.640	0.328	0.078	4.687	0.772	0.760	0.239	0.044	4.592	0.694
StrategyQA	0.629	0.342	0.079	4.470	0.661	0.698	0.279	0.052	4.200	0.649
mean	0.705	0.274	0.079	3.749	0.734	0.786	0.206	0.049	3.588	0.693