PROBABILISTIC CONSENSUS FOR WEIGHTED MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) offers a promising approach to improving the reasoning capabilities of large language models (LLMs), but reaching a correct consensus requires accounting for uncertainty in both proposed answers and the feedback used to evaluate them. Agents may produce incorrect solutions, and their assessments of other agents' solutions may themselves be mistaken. We introduce Weighted Iterative Society-of-Experts (WISE), a zero-shot framework for weighted multi-agent debate with probabilistic consensus. WISE augments textual critiques with discrete correctness weights assigned to candidate answers, providing structured evidence for modeling uncertainty in agents' evaluations. At its core is WISE-Dawid–Skene, an unsupervised probabilistic model that jointly estimates solver and scorer error patterns and infers answer correctness from weighted responses across debate rounds. This formulation accounts for uncertainty in both answer generation and scoring, enabling consensus to reflect the accumulated evidence and inferred capabilities of participating agents. Iterative feedback guides solution refinement, while the joint probabilistic model integrates the resulting answers and assessments to infer a collective decision. Experiments on standard multimodal reasoning benchmarks demonstrate that WISE consistently improves zero-shot reasoning, outperforming leading models by 6–12% on average. These results highlight weighted debate and joint probabilistic modeling of answer and feedback uncertainty as an effective approach to improving collective reasoning.

1 INTRODUCTION

Reasoning with contemporary foundation models is increasingly shifting away from relying on a single model's prediction toward systems composed of multiple agents that propose, evaluate, and refine solutions. This transition is particularly crucial for complex reasoning problems, where the final answer is only as trustworthy as the underlying chain of reasoning that leads to it. While large language models (LLMs) have shown striking performance across a wide range of everyday tasks—such as multimodal reasoning Team (2026), program synthesis Dong et al. (2025); Assumpção et al. (2025), mathematical problem-solving Luo et al. (2025); Ahn et al. (2024), formal theorem proving Song et al. (2024); Yang et al. (2023), and even algorithm discovery Novikov et al. (2025); Romera-Paredes et al. (2024)—their reasoning processes remain brittle. Minor variations in prompts, sampling temperature, context window, or problem formulation can produce divergent outcomes, and spurious correlations can render otherwise convincing chains of thought incorrect.

A growing line of work has therefore explored mechanisms for improving reliability through repeated sampling, self-reflection, ensembling, and interaction among agents Chen et al. (2023); Wang et al. (2022); Shinn et al. (2023); Liang et al. (2023). Multi-agent debate offers a way to improve language-model reasoning at inference time by allowing models to examine and revise one another's solutions Du et al. (2023); Liang et al. (2023); Wang et al. (2024c); Chen et al. (2023); Gu et al. (2025); Subramaniam et al. (2025); Thind et al. (2025). Its central difficulty, however, is deciding what to believe when both the proposed solutions and the judgments used to assess them can be wrong. Agreement alone does not establish correctness: several agents may repeat the same mistake, while a correct answer may receive little support. Feedback introduces a second source of uncertainty, since an evaluator may endorse an incorrect solution or reject a correct one. Effective collective reasoning therefore requires a mechanism for interpreting evidence from fallible answer generators and fallible evaluators.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

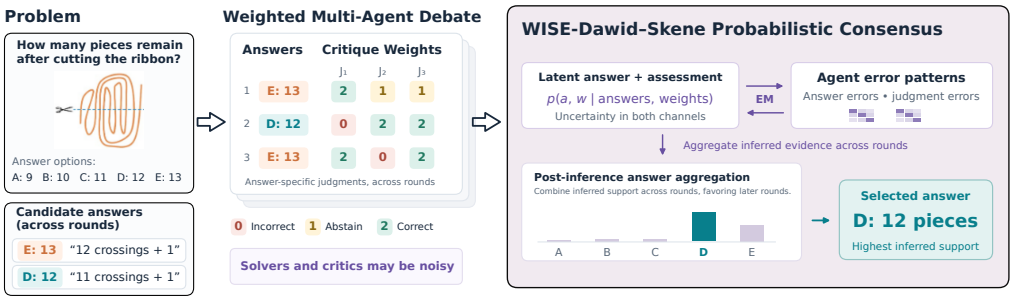


Figure 1: WISE MAD illustration. Agents propose competing answers to a problem across debate rounds, while independent critics evaluate the answers, provide feedback, and assign discretized assessments as *weights*. WISE treats the answers and weights as noisy evidence rather than fixed measures of reliability, and uses WISE-Dawid–Skene probabilistic inference to find the latent answer.

Existing approaches combine agent outputs through majority voting (Du et al., 2023), confidence-weighted voting (Chen et al., 2023), or adjudication by a language-model judge (Liang et al., 2023). RECONCILE, for example, combines discussion among diverse models with confidence-weighted voting Chen et al. (2023). Yet a reported confidence or correctness score is itself a model output whose evidential value must be established. Counting an answer treats its occurrence as support; weighting it by an evaluator's score additionally relies on that evaluator's judgment. Neither operation, by itself, estimates the errors underlying both observations. Empirical findings that debate does not consistently outperform simpler sampling and ensembling methods further motivate separating the benefits of generating more responses from those of interpreting them effectively Smit et al. (2023).

Consider a reasoning problem with two competing answers. Several agents propose one answer, but an evaluator identifies a substantive error in their reasoning and supports the alternative. The collective decision depends on more than the number of votes or the strength of the endorsement: it also depends on how the participating models behave when generating answers and assessing candidate solutions. An agent may be a strong solver but a weak evaluator, and an evaluator's confident approval need not be accurate. This motivates treating consensus as inference about answer correctness from noisy answers and assessments. Here, consensus denotes the inferred collective decision; it does not require every agent to converge to the same response.

We introduce Weighted Iterative Society-of-Experts (WISE), a zero-shot framework that combines weighted debate with probabilistic consensus. WISE pairs natural-language critiques with discrete assessments of candidate-answer correctness. These complementary feedback channels serve different purposes: textual critiques provide guidance for revising solutions, while discrete scores provide structured observations for probabilistic inference. Crucially, the scores remain fallible evidence rather than being accepted as known probabilities of correctness. Discretizing assessments makes it possible to model uncertainty in the feedback alongside uncertainty in the answers being evaluated.

At the core of WISE is WISE–Dawid–Skene, a probabilistic consensus formulation that extends the classical Dawid–Skene (DS) approach to observer-error estimation Dawid & Skene (1979). Classical DS infers a latent label from one-shot annotator responses using a single error model for each annotator. WISE-DS models multi-round debate as a coupled two-channel process where agents propose answers and evaluate candidate-conditioned hypotheses. It assigns agent-specific distributions to generation and assessment errors and jointly infers them with the latent answer. Using expectation–maximization, WISE-DS performs this inference over all responses, scores, and rounds without ground truth, then uses accumulated evidence to make the final collective decision. This enables weighted aggregation while allowing generation and evaluation competence to differ, keeping language models fixed and estimating only their observed response and assessment behavior.

Our formulation is agnostic to the input modality, and we instantiate it for multimodal reasoning. This setting introduces additional opportunities for error: a solution may fail through visual misperception, incorrect grounding, or faulty reasoning, and an evaluator may have access to only part of the evidence needed to identify the failure. For example, an evaluator that receives only a textual solution can assess its logical consistency but cannot independently verify its visual premises. Multimodal tasks therefore provide a demanding setting for studying consensus under uneven access to evidence. Probabilistic aggregation can account for observed error patterns, but its conclusions remain limited by the evidence

available to the participating agents. Figure 1 illustrates this challenge on a representative multimodal reasoning problem Cherian et al. (2024), where solving the task requires not only visual recognition but also targeted reconsideration, task-specific knowledge, and feedback from multiple perspectives.

To evaluate the efficacy of WISE, we conduct experiments on five datasets: (i) SMART-840 Cherian et al. (2024), (ii) VisualPuzzles Song et al. (2025), and (iii) EvoChart-QA Huang et al. (2025), (iv) MMMU, and (v) SMART-840++, a new programmatic extension of SMART-840. Across these datasets and varied MAD configurations, we show that WISE delivers consistent improvements, beating the state-of-the-art models with substantial margins (6-12%).

To summarize, our contribution is a framework that connects structured, weighted debate to probabilistic inference over uncertain answers and uncertain evaluations. Specifically, we present: 1) WISE, a weighted debate formulation that exposes candidate assessments as discrete, potentially erroneous observations alongside textual feedback, and 2) WISE-Dawid-Skene, a joint probabilistic consensus model for answer-generation and assessment errors. 3) Our framework demonstrates state-of-the-art performance on five vision–language benchmarks, including the new SMART-840++ dataset.

2 Related Works

Multi-Agent Debate Methods. MAD extends the notion of self-correction and feedback by enabling agents to debate candidate solutions. For instance, Du et al. (2023) examines debates between identical LLMs, whereas Liang et al. (2023) supplements them with external feedback and a judge model. Heterogeneous multi-agent configurations include collaborative or competitive settings Feng et al. (2024), secretary-style orchestration Wang et al. (2024c), and RECONCILE Chen et al. (2023), where agents exchange uncertainty estimates and iteratively refine their responses. Explain-Analyze-Generate Gu et al. (2025) emphasizes the danger of misleading agents and argues for task decomposition, while Subramaniam et al. (2025) fine-tunes MAD systems to improve convergence. Recent studies go beyond fully symmetric debates by assigning agents complementary, specialized roles. Multiagent Finetuning Subramaniam et al. (2025) demonstrates that varied reasoning specializations can strengthen group reasoning, and MARS Wang et al. (2025) uses distinct author, reviewer, and meta-reviewer roles to enable rapid iteration. In contrast, WISE employs clearly defined agent roles, avoids dependence on self-reported confidence (which frequently needs external calibration Thind et al. (2025); Khan et al. (2024)), and naturally enables abstention via weighted scoring.

Mixture of LLMs. Ensemble-based methods combine multiple LLMs. Mixture-of-Agents (MoA) Wang et al. (2024a) treats ensembles as prompt- and gate-controlled layers, and Self-MoA Li et al. (2025a) forwards only top-performing outputs. Other work studies heterogeneous mixtures with majority voting Li et al. (2024) or orchestrated systems like Magnetic-One Fourney et al. (2024). In contrast, WISE places agents in an iterative reflective loop with critic weights, using feedback to progressively refine solutions. Mixture-of-experts approaches Li et al. (2025b); Shen et al. (2024); Wu et al. (2024) typically require model training, whereas we focus on zero-shot settings.

Aggregation Approaches. Most MAD systems aggregate solutions via simple majority voting Du et al. (2023) or weighted averaging based on self-reported confidence Chen et al. (2023). From a crowdsourcing perspective, agents act as annotators with systematic biases and errors, motivating the use of quality-control methods for statistically principled consensus Dawid & Skene (1979); Hovy et al. (2013); Ma & Olshevsky (2020); Majdi & Rodriguez (2023); Ustalov et al. (2021). However, Dawid–Skene-style models, despite their prevalence in human-annotation tasks, remain underexplored for agent-based feedback loops. WISE addresses this gap by introducing a probabilistic framework that jointly models solver and critic reliabilities, enabling unified aggregated solution inference across all debate rounds.

3 Proposed Method – WISE

Problem Setup. Given a problem $p \in \mathcal{P}$, our goal is to construct a model $M : \mathcal{P} \to \mathcal{A}$ that predicts its correct answer $a \in \mathcal{A}$. We denote this prediction by $\hat{a} = M(p)$ and seek $\hat{a} = a$. We assume access to a collection $\mathcal{L}$ of heterogeneous language and vision-language models (VLMs) and aim to combine their capabilities to solve problems robustly in a zero-shot setting. Our central premise is that models trained with different architectures, modalities, and objectives exhibit complementary

competencies: a model that reliably extracts one aspect of a problem may be less reliable on another. This heterogeneity provides multiple, partially independent sources of evidence that can be aggregated to improve collective reasoning, consistent with the society-of-experts paradigm Minsky (1986).

For example, consider a multimodal mathematical problem comprising a geometric diagram and a textual query. A vision-language model such as Qwen-VL Wang et al. (2024b) may accurately localize visual entities and recover the diagram structure; a text-oriented model such as Phi Abdin et al. (2024) may derive symbolic relations from textual or intermediate representations; and a general-purpose GPT model may provide stronger cross-modal integration, commonsense reasoning, or code-assisted verification. However, the reliability of these capabilities can vary across problems and cannot be known a priori. The central challenge is therefore not merely to solicit and refine multiple solutions, but to infer which candidate answers are most credible from the pattern of support they receive from heterogeneous models. This requires accounting for the differing and initially unknown reliability of the models providing that support.

Motivated by this perspective, we introduce the Weighted Iterative Society-of-Experts (WISE) framework, illustrated in Figure 2. WISE treats candidate solutions and their evaluations as uncertain observations of an unknown correct answer and iteratively estimates a reliability-weighted probabilistic consensus. Although the framework is modality-agnostic, our experiments focus on multimodal reasoning, where differences among agent capabilities are particularly pronounced. Accordingly, we instantiate each problem p as an image–question pair for joint visual and linguistic understanding.

WISE Architecture. Given $|\mathcal{L}|$ heterogeneous agents, conventional multi-agent debate protocols often allow every agent to communicate with every other agent Chen et al. (2023); Du et al. (2023), requiring up to $\mathcal{O}(|\mathcal{L}|^2)$ messages per round. Such all-to-all communication can be redundant, costly for API-based models, computationally demanding for locally hosted models, and difficult to accommodate within finite context windows. It also implicitly assumes that every agent can consume and reason over the same input modalities, which is unnecessarily restrictive in multimodal settings.

Rather than organizing interaction as unrestricted conversation, WISE instantiates a *bipartite weighted debate*: a set of n proposal models $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ generates candidate solutions, while a set of m evaluation models $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ independently assesses each candidate. Their candidate-conditioned assessments form a structured matrix of weighted evidence from which WISE infers a probabilistic consensus. A lightweight text-only operator D compresses the associated diagnostic feedback before the next round. Thus, the defining object in WISE is not a sequence of solver, critic, and judge decisions, but the weighted bipartite interaction between candidate solutions and heterogeneous evaluations. The same underlying model may instantiate more than one function; formally, $\mathcal{S} \cup \mathcal{C} \cup \{D\} = \mathcal{L}$, without requiring these sets to be disjoint.

This organization also permits modality-asymmetric participation. Proposal models are multimodal because they must directly interpret the image–question pair. Evaluation models may be either multimodal or text-only: the former can verify claims against the original visual input, whereas the latter can assess the logical and linguistic consistency of a proposed reasoning trace using its textual representation. The exact prompts used to implement these interactions are provided in the Appendix.

Candidate-Solution Generation. For a problem p, each proposal model $s_i \in \mathcal{S}$ receives $t_1^S = \left(\mathrm{PROMPT}_1^S, p, \mathcal{A}_p, f = \phi\right)$, where PROMPT_1^S specifies the task, $\mathcal{A}_p$ denotes the candidate answer set, and f contains evidence-derived feedback from preceding rounds. In the first round, no such feedback is available, and hence $f = \phi$. At round k, model s_i generates a response r_i^k comprising: (i) a candidate answer $\hat{a}_i^k \in \mathcal{A}_p$, (ii) a reasoning trace supporting that answer, and (iii) a textual description of the relevant visual content. The final component provides text-only evaluation models with an intermediate representation through which they can assess the reasoning, while multimodal evaluators may additionally verify its visual grounding against p. We omit the superscript k when the round index is unimportant.

Weighted Candidate Evaluation. At round k, each evaluation model $c_j \in \mathcal{C}$ independently assesses every candidate response r_i^k. Its input is

$$t_{k,ij}^C = \left(\mathrm{PROMPT}^C, p, r_i^k\right),$$

where PROMPT^C specifies the evaluation task, as illustrated in Figure 1. A text-only evaluator receives the textual question together with the visual description contained in r_i^k, whereas a multimodal evaluator may additionally inspect the original image.

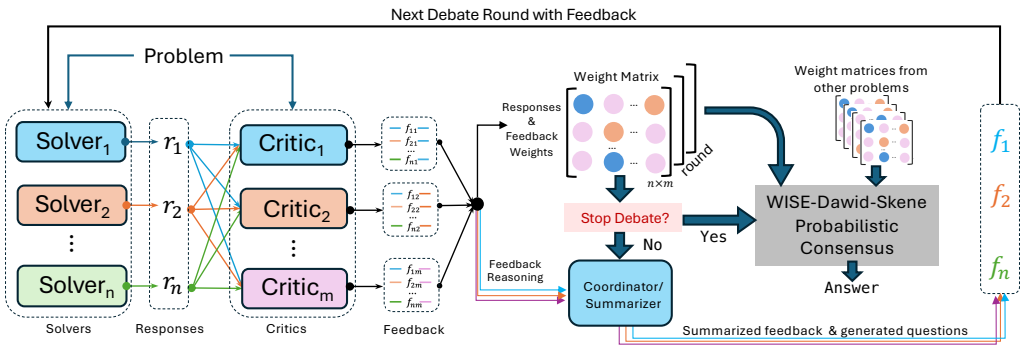


Figure 2: WISE architecture, depicting the control flow, responses, feedback, and the WISE-Dawid-Skene probabilistic consensus scheme to produce the final response.

Free-form critiques Chen et al. (2024); Welleck et al. (2022) provide rich diagnostic information, but they do not offer a standardized observation that can be compared across evaluators, candidates, and debate rounds. Their linguistic expressions of agreement, disagreement, and uncertainty can vary substantially, making them unsuitable for direct probabilistic aggregation. We therefore ask each evaluator to *project* its assessment onto a common three-state evaluation variable $w_{ij}^k \in \{0, 1, 2\}$, where 0 denotes evidence that the candidate is incorrect, 1 denotes uncertainty or abstention, and 2 denotes evidence that the candidate is correct. This discretization converts an otherwise unstructured semantic judgment into a categorical observation on which a likelihood can be defined. The intermediate state is important: without it, an evaluator lacking sufficient visual or logical evidence would be forced to contribute either positive or negative evidence. Conversely, using a more finely graded or continuous confidence score would require stronger and often unjustified assumptions about numerical calibration across heterogeneous models.

Alongside the categorical assessment, evaluator c_j produces a diagnostic explanation f_{ij}^k that justifies its judgment by identifying valid intermediate steps, unsupported claims, and visual or logical errors. Thus, the two outputs serve distinct purposes: w_{ij}^k provides a standardized observation for probabilistic inference, while f_{ij}^k preserves the semantic information needed to refine the candidate in subsequent rounds. Evaluating all n candidates with all m evaluators yields the score matrix

$$W^k = \left[w_{ij}^k\right]_{i\in[n],\, j\in[m]} \in \{0, 1, 2\}^{n\times m}$$

and the corresponding diagnostic-feedback matrix $F^k = \left[f_{ij}^k\right]_{i\in[n],\, j\in[m]}$. Importantly, the entries of W^k are not interpreted as calibrated confidence values or combined through direct averaging. They are noisy, candidate-conditioned observations whose reliability depends on the evaluator that produced them. Consequently, identical scores from different evaluators need not constitute equally strong evidence. This structured representation thus facilitates our probabilistic model in Section 3.

Feedback Compression and Refinement. The textual explanations in F^k can collectively exceed a proposal model's context window. We thus use a text-only compression operator D, implemented by an LLM, to transform the evaluations associated with each candidate into concise feedback for the next round. This operator is also not exposed to the original problem p; its function is limited to preserving the substantive agreements, disagreements, and diagnostic evidence contained in the evaluations while removing redundant text, and distill them into focused challenge questions. These questions push the solver to re-examine specific visual, linguistic, or logical aspects of the problem in the subsequent round. For candidate i, the compressed feedback at round k is $f_i^k = \mathrm{L}^D\left(\mathrm{PROMPT}^D, \bigoplus_{j\in[m]} f_{ij}^k\right), \quad i \in [n]$, where PROMPT^D instructs D to express the collected evaluations as focused questions or corrective suggestions without adjudicating among candidate answers. If another round is required, proposal model s_i receives $t_{k+1,i}^S = \left(\mathrm{PROMPT}_{k+1}^S, p, \mathcal{A}_p, f = r_i^k \oplus f_i^k\right)$. The model may then retain, revise, or replace its candidate solution in response to the diagnostic evidence. This produces an iterative process in which candidate hypotheses are refined through structured feedback, while answer selection remains governed by the subsequent reliability-weighted probabilistic consensus.

WISE-Dawid-Skene Solution Aggregation. As discussed, a key challenge with heterogeneous models is rigorously assessing reliability. A model may produce near-random outputs (e.g., from high sampling temperature), lack sufficient reasoning ability, or have other systematic flaws, making it difficult to judge how much its responses can be trusted. To address this issue, we build on the crowdsourcing literature Majdi & Rodriguez (2023), where annotator reliability is classically calibrated with the Dawid–Skene (DS) model Dawid & Skene (1979), and present our WISE-Dawid-Skene framework. With a set of workers each providing one of a fixed set of categorical labels per item, WISE-DS is an unsupervised probabilistic model that uses expectation–maximization (EM) to jointly infer agent-specific error rates and the latent ground-truth label distribution. We treat each (agent, round) pair as a distinct annotator with its own error profile, allowing us to model agents' ability to incorporate critiques over rounds, their evolving error patterns, and the higher reliability of later-round responses. WISE-DS thus estimates agent-specific error rates and the joint posterior over all debate rounds to produce a single most likely solution. EM in WISE-DS downweights near-random agents and amplifies those whose outputs consistently align with the emerging consensus.

Mathematically, after K debate rounds, let each solver $s_i \in \mathcal{S}$ produce a sequence of responses $\{r_i^k\}_{k=1}^{K_i}$, where $K_i \leq K$ allows for the possibility that solver s_i does not participate in every round. Let $\hat{a}_i^k \in \mathcal{A}_p$ denote the candidate answer extracted from r_i^k, and let $w_{ij}^k \in \{0,1,2\}$ denote the correctness weight assigned by critic c_j to solver s_i at round k. Importantly, these scores are themselves noisy: a critic may assign a high score to an incorrect solver response, or a low score to a correct one. Our goal is to infer a global posterior distribution over the latent true answer $y_p \in \mathcal{A}_p$ using all solver answers and critic weights produced across debate rounds.

Solver responses as noisy annotations. Following the Dawid–Skene perspective, we treat each solver-round pair (i,k) as a noisy annotator of the unknown answer y_p. Each solver-round pair is associated with a confusion matrix $\Theta_i^k(a,b) := \Pr\big(\hat{a}_i^k = b \mid y_p = a\big)\,,\ a,b \in \mathcal{A}_p$, which captures the probability that solver s_i outputs answer b at round k when the true answer is a. This allows the reliability of a solver to vary across debate rounds, reflecting the fact that a solver may improve after receiving feedback or may drift toward an incorrect answer.

Critic weights as noisy correctness observations. Critics do not directly vote for the final answer. Instead, they evaluate solver responses by assigning weights $w_{ij}^k \in \{0,1,2\}$. We model each critic supplied weight as a noisy observation of whether the solver answer is correct. Define the latent correctness indicator $c_i^k(y_p) = \mathbb{1}\big[\hat{a}_i^k = y_p\big]$, where $c_i^k(y_p) = 1$ indicates that solver s_i is correct at round k, and $c_i^k(y_p) = 0$ indicates that it is incorrect. Each critic c_j is associated with a verification matrix: $\Psi_j(c,w) = \Pr\big(w_{ij}^k = w \mid c_i^k(y_p) = c\big)\,,\ c \in \{0,1\},\ w \in \{0,1,2\}$. Thus, $\Psi_j(1,2)$ captures how often critic c_j assigns a high correctness weight to a correct solver response, while $\Psi_j(0,2)$ captures how often it mistakenly endorses an incorrect response.

Consistency coupling across rounds. We define the evidence contributed by solver s_i at round k under the candidate truth y as:

$$E_i^k(y) = \Theta_i^k\big(y, \hat{a}_i^k\big) \prod_{j=1}^{m} \Psi_j\big(\mathbb{1}[\hat{a}_i^k = y], w_{ij}^k\big)\,. \tag{1}$$

Thus, $E_i^k(y)$ combines two sources of evidence: the solver's answer likelihood under candidate truth y, and the critic-weight likelihood induced by whether that answer matches y. Aggregating over all solvers in round k, we define the round-level evidence as $E^k(y) = \prod_{i=1}^{n_k} E_i^k(y)$, where n_k denotes the number of solvers that participate in round k. This allows the model to handle variable participation across rounds.

Global posterior over the true answer. Let $\pi(y) = \Pr(y_p = y)$ denote the prior probability of candidate answer $y \in \mathcal{A}_p$, with $\sum_{y \in \mathcal{A}_p} \pi(y) = 1$. Given all solver answers and critic weights across debate rounds, denoted by $\mathcal{D}_{p,1:K}$, the global posterior over the latent true answer is

$$p_p(y) := \Pr(y_p = y \mid \mathcal{D}_{p,1:K}) = \frac{\pi(y) \prod_{k=1}^{K} E^k(y)}{\sum_{a \in \mathcal{A}_p} \pi(a) \prod_{k=1}^{K} E^k(a)}. \tag{2}$$

The final aggregated answer is obtained by maximum a posteriori inference: $\widehat{y}_p = \arg\max_{y \in \mathcal{A}_p} p_p(y)$. For each candidate answer y, the posterior combines solver and critic evidence from all rounds, then normalizes over $\mathcal{A}_p$. The solver term is the likelihood of the observed

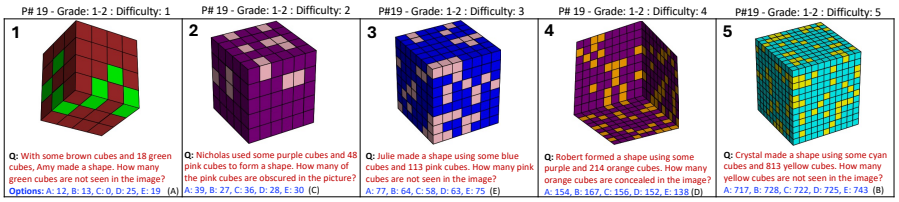


Figure 3: Example problems from our proposed SMART-840++ dataset. Cells 1-5 show the same puzzle but with increasing difficulties. See Appendix H for more examples.

answer $\hat{a}_i^k$ under y, and the critic term is the likelihood of the weight w_{ij}^k given the correctness implied by y. All rounds share the same latent answer y_p, so every solver response and critic weight must agree with a single underlying truth. For each y, the model accumulates round-level evidence $E^k(y)$ over the trajectory, so answers consistently supported by multiple solvers and critics obtain higher posterior mass, while isolated or inconsistent answers have limited impact. We estimate parameters with EM; see Appendix A for details.

4 Experiments

Datasets. We evaluate WISE on four vision–language reasoning datasets: SMART-840 Cherian et al. (2024), Visual Puzzles Song et al. (2025), EvoChart-QA Huang et al. (2025), and MMMU-val Yue et al. (2024). SMART-840 comprises 840 children's math Olympiad problems (grades 1–12) for analyzing age-appropriate VLM performance. Visual Puzzles includes 1,168 problems assessing algorithmic, analogical, deductive, inductive, and spatial multimodal reasoning. Both are multiple-choice datasets. EvoChart-QA is a challenging subset of 320 ChartQA problems (line, bar, scatter, pie); answers are free-form. MMMU is a large-scale benchmark for multidisciplinary, college-level multimodal reasoning; we use its test-mini split with 900 multiple-choice questions. We report multiple-choice accuracy for the performance, and use LLM-as-a-judge on EvoChart-QA.

SMART-840++ Dataset. We introduce SMART-840++, a controlled generalization benchmark of 55 programmatically instantiated puzzle families. Each family has three new instances at five difficulty levels (825 problems total). SMART-840++ evaluates whether VLMs can transfer a known puzzle family's solution procedure to unseen instantiations with different entities, visual attributes, layouts, constraints, and numerical values, rather than requiring new reasoning paradigms. We recreate each puzzle in Python and vary these factors while preserving the underlying reasoning structure. This controlled design supports evaluation across increasing complexity and limits instance-level memorization from public benchmarks, without assuming the puzzle types are unseen. For quality control, five independent annotators verified that each instance is well posed, solvable from the given information, and has a unique answer. Figure 3 shows SMART-840++ examples.

Agents and Debate Notation. We use diverse models in our experiments, listed below with their abbreviations: GPT-4.1 (G41), GPT-5.2 (G52), o4-mini (o4m), Claude-Sonnet-3.7 (CS37), Claude-Sonent-4.6 (CS46) and Gemma3-8B (Ge3). To represent a debate, we use the notation $(s_1 + s_2 + \cdots) \times (c_1 + c_2 + \cdots)|$Distiller, where the first set are the solver models and the second set are the critics. The compact notation $(\mathrm{L}_1 + \mathrm{L}_2 + \cdots)^2|D$ depicts $\mathrm{L}_1 + \mathrm{L}_2 + \cdots$ as both solvers and critics.

Baseline Comparisons: In Table 1, we report performance across the five datasets, comparing WISE with SOTA VLMs. WISE consistently outperforms all baselines. On SMART-840, WISE with (G52+CS46) improves G52 accuracy by 6% (72%→78.4%), and yields about 7% gains over the best G41 model when using older, weaker models. This trend holds for both image-text and text-only SMART-840 subsets, indicating the modality-agnostic nature of our method. Other datasets show similar patterns: on Visual Puzzles, WISE gives an 11% gain over G52 and CS46; on EvoChart (short-answer rather than multiple choice), it yields 7% improvements; and on general VQA (MMMU), gains are similar. Overall, WISE provides 6–12% improvements over singleton baselines. On our new dataset SMART-840++, SOTA models perform markedly worse (around 50% vs. 70%+ on other datasets), suggesting that complex visually grounded mathematical reasoning remains challenging for VLMs. Even here, WISE achieves about a 5% gain. Table 1 also compares WISE with the strong MAD baseline RECONCILE Chen et al. (2023). While both use heterogeneous multi-agent debate, RECONCILE relies on model self-confidence rather than explicit cross-modal

Method	LLMs / VLMs	Image+Text	Text	Overall
Single	GPT-4.1	48.2	88.2	60.6 ± 0.6
	GPT-5.2	65.0	95.9	72.0 ± 1.7
	Gemma3	24.0	73.7	38.6 ± 4.8
	CS37	35.3	71.0	44.3 ± 2.9
	CS46	55.2	92.5	65.8 ± 1.4
RECONCILE	(G52+CS46)	67.2	95.8	72.7±1.2
WISE (ours)	(G41+Ge3+CS37)^2|G4o	55.8	91.4	68.1 ± 1.1
	(G52+CS46)^2|G52	**70.6**	**98.2**	**78.4 ± 0.6**

SMART-840 dataset

Method	LLMs/VLMs	Avg*	Our Run
Single	o4-mini	57.0	53.3 ± 1.1
	CS37	48.3	46.6 ± 2.8
	Gemma3	N/A	31.6 ± 3.9
	CS46	N/A	49.2 ± 1.7
	G52	N/A	50.4 ± 2.1
RECONCILE	(G52+CS46)	N/A	60.3 ± 1.1
WISE (ours)	(o4m+CS37+Ge3)^2|G4o	N/A	59.9 ± 1.4
	(G52+CS46)^2|G52	N/A	**61.2 ± 1.6**

Visual Puzzles Dataset

Method	LLMs/VLMs	Avg
Single	o4-m	45.4 ± 1.1
	CS37	51.9 ± 1.2
	G52	68.7 ± 1.1
	CS46	69.5 ± 1.5
RECONCILE	(G52+CS46)	71.8 ± 3.0
WISE (ours)	(o4m+CS37)^2|G4o	75.1 ± 1.2
	(G52+CS46)^2|G52	**76.2 ± 1.3**

EvoChart Dataset

Method	LLMs/VLMs	Avg
Single	G52	74.1 ± 1.7
	CS46	74.9 ± 1.5
	o4-m	73.2 ± 2.1
RECONCILE	(G52+CS46)	72.8 ± 0.5
WISE (ours)	(o4m+CS46)^2|G4o	76.2 ± 1.3
	(G52+CS46)^2|G52	**79.7 ± 1.0**

MMMU Dataset

Method	LLMs/VLMs	Avg
Single	G52	48.9 ± 3.2
	CS46	40.4 ± 4.1
	o4m	62.2 ± 5.1
RECONCILE	(G52+CS46)	54.4 ± 1.5
WISE (ours)	(o4m+CS46)^2|G41	**62.8 ± 2.8**
	(G52+CS46)^2|G52	59.0 ± 2.9

SMART-840++ Dataset

Table 1: We compare the WISE to singleton and strong MAD baselines Chen et al. (2023), using both recent strong VLMs (GPT-5 and Claude-4) and older / weaker models.

Dataset	Config.	Maj. V	Wt. MV	DS	Wt. DS	WISE-DS
SMART-840	(G52+CS46)^2|G52	76.3 ± 1.2	77.6 ± 1.5	76.9 ± 1.4	76.9 ± 1.1	**78.4 ± 0.6**
SMART-840	(G41+CS37+Ge3)^2|G4o	63.9 ± 1.7	63.3 ± 0.9	66.3 ± 2.1	66.1 ± 2.8	**68.1 ± 1.1**
SMART-840++	(G52+CS46)^2|G52	57.3 ± 2.8	57.6 ± 2.9	53.9 ± 1.5	57.0 ± 1.0	**59.0 ± 2.9**
Visual Puzzles	(G52+CS46)^2|G52	58.1± 2.8	59.5 ± 1.6	53.2 ± 1.5	54.4 ± 1.0	**61.2 ± 1.6**
MMMU	(G52+CS46)^2|G52	74.4 ± 1.6	74.7± 1.6	75.5± 0.7	76.5± 2.1	**79.7 ± 1.0**

Table 2: Comparisons of WISE-DS vs. aggregation methods. DS: Dawid-Skene, Maj. V: majority voting, Wt. Maj.V uses our critic weights, and Wt. DS uses temporal weighting on DS solutions.

feedback and reliability-aware aggregation. All experiments use the same debate setup: WISE uses (G52+CS46)×(G52+CS46)|G52, and RECONCILE uses the same models, with up to 4 debate rounds. Across all five datasets, WISE outperforms RECONCILE by 1–8%.

Comparison to Aggregation Methods. Table 2 compares WISE-DS with alternatives: majority voting as a baseline, the standard Dawid–Skene (DS) model, and a Weighted DS variant where DS is applied per round and responses are linearly weighted by k/K for the k-th of K rounds. Our WISE architecture produces critic weights for each solver response; we also use these as weights for majority voting (Wt. MV).[1] Across aggregation methods, debate configurations, and datasets, the table shows consistent gains. WISE-DS outperforms DS yielding steady improvements.

Compute Matched Performance Analysis. In order to separate the effects of additional sampling in WISE, model heterogeneity, and role-based orchestration, we conducted experiments using locally hosted open-weight models: Molmo2-O-7B (M), Gemma4-31B (G4), and the text-only GPT-OSS-20B (OSS). All models were served with vLLM on A100 GPUs, with each invocation limited to 8,192 generated tokens and 60 seconds, and each method limited to four debate rounds. We evaluate on the complete SMART-840 dataset. Table 3 reports these results.

#	Configuration	Debate	Majority-vote	Dawid-Skene	WISE-DS
1	(G4 × G4)|G4	WISE (homogeneous)	74.2 ± 0.6	74.1 ± 0.5	74.5 ± 0.3
2	(M + G4 + OSS)|OSS	RECONCILE Chen et al. (2023)	65.6 ± 1.6	74.5 ± 3.04	78.7 ± 1.5
3	(M+G4)×(OSS+G4)|OSS	WISE (full model)	79.7 ± 0.4	82.1 ± 0.5	**83.6 ± 0.3**
4	(M+G4)× G4|OSS	WISE	80.6 ± 2.1	82.0 ± 1.2	82.3 ± 0.85
5	(M + G4)|OSS	WISE (w/o critics)	71.8 ± 0.95	79.1 ± 1.1	79.2 ± 1.6
6	(M+G4)×(OSS+G4)	WISE (No distiller)	81.2 ± 0.89	81.9 ± 0.62	82.3 ± 1.56
7	(M × M) + (G4 × G4)|OSS	Self-Feedback	75.7 ± 0.87	79.6 ± 0.86	80.4 ± 1.63

Table 3: Ablation study over various components in the WISE under compute-matched settings.

First, within every row, the three aggregation methods operate on the same agent outputs and therefore use identical models, calls, tokens, and debate traces. WISE-DS achieves the highest mean accuracy in all seven settings, improving over majority voting by an average of 4.6% points and over classical Dawid–Skene by 1.1 points. In the full WISE configuration, accuracy increases from 79.7% with majority voting to 82.1% with classical Dawid–Skene and 83.6% with WISE-DS. This exactly matched comparison shows that the gain cannot be attributed to additional models, rounds, or

[1]This is not a true baseline but an ablation enabled by our architecture.

Config.	Debate	Tokens/LLM	#Tokens/Round	Cost (cents)	Latency (s)	Acc. (%)
$(G52+CS46)^2$|G41	WISE	7215.35	1946.77	7.6	12.23	65.3
$(G41+Ge3+S35)^2$|G4o	WISE	6392.45	1617.94	4.8	10.95	58.2
(G41+Ge3+S35)	RECONCILE	5758.60	1447.93	4.4	5.93	50.0
G41+CS35+Ge3+Q3	Homogeneous	3050.69	771.28	2.6	12.96	41.3

Table 4: Comparison of compute budget versus performance on a subset of the SMART-840 dataset. We report the tokens/round, latency (s) / round and cost/round.

sampling. It arises from reliability-aware aggregation and the joint modeling of solver answers and candidate-conditioned critic judgments.

Second, the architectural ablations show that WISE's improvement is not produced by merely increasing the number of calls. Removing critics (#5) reduces WISE-DS accuracy from 83.6% to 79.2%, while replacing heterogeneous cross-model feedback with self-feedback (#7) reduces it to 80.4%. Removing the Distiller (#6) reduces accuracy to 82.3%, showing that synthesizing critiques into targeted cross-round guidance provides value beyond direct response exchange. Removing the text-only OSS critic (#3 vs. #4) also reduces performance from 83.6% to 82.3% (expt #3 vs. #4). This supports the intended heterogeneous design: although a text-only critic cannot independently verify visual perception, it can identify errors in the solver's argument. Its contribution is thus complementary to that of image-aware agents. The homogeneous G4-only (#1) reaches only 74.5% against 83.6% for heterogeneous WISE. Thus, repeatedly invoking one strong model under different role prompts does not reproduce the benefit. Similarly, RECONCILE reaches 78.7% under the same open-weight serving constraints, indicating that multi-model, multi-round interaction alone is insufficient.

Compute Budget? In Table 4, we report the full token-cost-performance analysis. Relative to RECONCILE, WISE increases per-round cost only slightly while improving accuracy substantially. Likewise, our strongest WISE configuration is far above homogeneous, showing that the extra cost is converted into stronger reasoning than redundant model calls.

Performance versus Debate Rounds?

Figure 4 compares cumulative accuracy—the fraction of problems solved by or before round k—with mean per-round accuracy across aggregation methods. WISE performs strongly under both metrics. Moreover, the small gap between its cumulative accuracy, which represents the performance attainable from the available candidate solutions, and its mean accuracy indicates that the probabilistic consensus reliably identifies correct candidates. This gap is substantially smaller for WISE-DS than for the alternative aggregation methods.

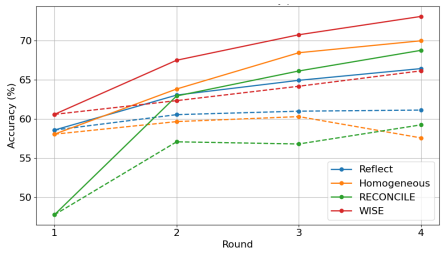


Figure 4: Cumulative and mean per-round accuracy.

5 Conclusions

We introduced WISE, a reliability-aware framework for collective reasoning with heterogeneous VLMs. Instead of fixed solver/critic/coordinator roles, WISE frames multi-agent debate as a structured bipartite exchange between candidate solutions and candidate-conditioned evaluations. Each free-form evaluation is paired with a categorical score, yielding a matrix of noisy evidence for probabilistic aggregation. This cleanly separates channels: discrete scores enable consensus inference, while critiques drive candidate refinement across rounds. The design supports heterogeneous agents with different modalities and reasoning abilities without requiring uniformity or unrestricted all-to-all communication.

Our results also expose a fundamental limitation of conventional voting-based aggregation: it treats judgments from heterogeneous models as equally reliable, despite systematic differences in their visual, linguistic, and logical competencies. WISE-DS instead jointly infers evaluator reliability and latent candidate correctness from the pattern of assessments across candidates and debate rounds. Across our benchmarks, reliability-aware consensus turns complementary but noisy model judgments into correct solutions more effectively than standard voting and debate-based aggregation. These results suggest robust collective reasoning requires treating evaluations as uncertain evidence and explicitly modeling its reliability, not merely adding more agents or specialized roles.

LIMITATIONS

WISE boosts reasoning accuracy but increases inference-time compute due to multi-agent debate, reflection, and iteration, which can hinder use in latency- or resource-limited settings. Costs can be reduced via parallel execution, early stopping once consensus stabilizes, and using smaller distilled models for reflection or feedback. The reliability estimates learned by WISE-DS could also enable adaptive orchestration—routing only the most reliable or complementary agents, suppressing uninformative feedback, and adding rounds only when uncertainty or disagreement is high—cutting compute versus running every agent each round.

Further, our evaluation mainly targets multimodal reasoning with discrete answer spaces. Because WISE-DS assumes a finite hypothesis set, applying it to open-ended generation or code synthesis would require constructing candidates (e.g., clustering equivalent outputs, canonicalization, or grouping programs by execution behavior). Extending WISE-DS to dynamically built hypothesis spaces with reliability-aware, cost-adaptive orchestration for long-form outputs is an important direction for future work.

IMPACT STATEMENT

As language and vision–language models diversify, combining complementary strengths can outperform relying on a single model. WISE advances this by treating heterogeneous model evaluations as uncertain evidence and inferring a reliability-aware probabilistic consensus, rather than weighting all judgments equally. Because it is zero-shot over existing models, it can help reuse and compose models across modalities, scales, and competencies without task-specific training. However, consensus is not correctness. Correlated errors, shared biases, or systematically misleading evaluations can still yield wrong outcomes, and inferred reliabilities are not calibrated safety guarantees. Iterative evaluation also adds inference cost versus a single model, which may reduce the benefits of reuse. High-stakes deployments should assess correlated failure modes, include human oversight, and weigh cost–accuracy trade-offs.

AI USE STATEMENT

In this study, generative AI (GenAI) tools were used to support the drafting, revision, and editing of portions of the manuscript. Specifically, these tools assisted in improving linguistic clarity, coherence, and overall readability, and in refining implementation-related descriptions where appropriate. GenAI tools were also used in parts of the implementation of the method described in the paper. However, the genAI was not used in core design of the mechanisms or conceiving of the main ideas proposed in the paper. All content produced or suggested with AI assistance was critically evaluated, verified, and edited by the authors, who take full responsibility for the accuracy, originality, and integrity of the final manuscript. The authors assume full responsibility for the final content of this work, including all text, claims, and any artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work involved no human participants and did not collect, process, or store any personally identifiable information (PII). The SMART-840++ dataset was curated and packaged from pre-existing source assets without adding or inferring sensitive personal attributes. All included components were reviewed for licensing and redistribution terms. SMART-840++ is distributed only as permitted by those terms, with required notices (e.g., copyright, attribution, license text) preserved in documentation and/or accompanying files. Assets with restricted redistribution are excluded or referenced in accordance with applicable permissions. The release is intended to comply with relevant legal and contractual obligations and is provided for research and development use under those conditions.

REPRODUCIBILITY STATEMENT

We provide full implementation and experimental details in the paper and appendix for reproducibility, including model/training settings, preprocessing, evaluation protocol, and ablations/analyses. We will release the complete source code (training/evaluation scripts and configs) upon acceptance.

REFERENCES

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

Assumpção, H., Ferreira, D., Campos, L., and Murai, F. Codeevolve: An open source evolutionary coding agent for algorithm discovery and optimization. *arXiv preprint arXiv:2510.14150*, 2025.

Chen, A., Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Bowman, S. R., Cho, K., and Perez, E. Learning from natural language feedback. *Transactions on machine learning research*, 2024.

Chen, J. C.-Y., Saha, S., and Bansal, M. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023.

Cherian, A., Peng, K.-C., Lohit, S., Smith, K. A., and Tenenbaum, J. B. Are deep neural networks smarter than second graders? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10834–10844, 2023.

Cherian, A., Peng, K.-C., Lohit, S., Matthiesen, J., Smith, K., and Tenenbaum, J. Evaluating large vision-and-language models on children's mathematical olympiads. *Advances in Neural Information Processing Systems*, 37:15779–15800, 2024.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Dong, Y., Jiang, X., Qian, J., Wang, T., Zhang, K., Jin, Z., and Li, G. A survey on code generation with llm-based agents. *arXiv preprint arXiv:2508.00083*, 2025.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., and Tsvetkov, Y. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J., Alber, J., et al. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

Gu, W., Han, J., Wang, H., Li, X., and Cheng, B. Explain-analyze-generate: A sequential multi-agent collaboration method for complex reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7127–7140, 2025.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130, 2013.

Huang, M., Lai, H., Zhang, X., Wu, W., Ma, J., Zhang, L., and Liu, J. EvoChart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

Li, W., Lin, Y., Xia, M., and Jin, C. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025a.

Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., Ma, L., and Zhang, M. Uni-MOE: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Lin, B. Y., Bras, R. L., Richardson, K., Sabharwal, A., Poovendran, R., Clark, P., and Choi, Y. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

Luo, Z., Yang, Z., Xu, Z., Yang, W., and Du, X. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

Ma, Q. and Olshevsky, A. Adversarial crowdsourcing through robust rank-one matrix completion. *Advances in Neural Information Processing Systems*, 33:21841–21852, 2020.

Majdi, M. S. and Rodriguez, J. J. Crowd-certain: Label aggregation in crowdsourced and ensemble learning classification, 2023. URL https://arxiv.org/abs/2310.16293.

Minsky, M. *Society of mind*. Simon and Schuster, 1986.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J., Mehrabian, A., et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Shen, L., Chen, G., Shao, R., Guan, W., and Nie, L. MoME: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D., and Pretorius, A. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*, 2023.

Song, P., Yang, K., and Anandkumar, A. Lean copilot: Large language models as copilots for theorem proving in lean. *arXiv preprint arXiv:2404.12534*, 2024.

Song, Y., Ou, T., Kong, Y., Li, Z., Neubig, G., and Yue, X. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Subramaniam, V., Du, Y., Tenenbaum, J. B., Torralba, A., Li, S., and Mordatch, I. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.

Team, Q. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Thind, R., Sun, Y., Liang, L., and Yang, H. Optimai: Optimization from natural language using llm-powered ai agents. *arXiv preprint arXiv:2504.16918*, 2025.

Ustalov, D., Pavlichenko, N., and Tseitlin, B. Learning from crowds with crowd-kit. *arXiv preprint arXiv:2109.08584*, 2021.

Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., and Zou, J. Mixture-of-agents enhances large language model capabilities. *URL https://arxiv. org/abs/2406.04692*, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wang, Q., Wang, Z., Su, Y., Tong, H., and Song, Y. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024c.

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. *arXiv preprint arXiv:2203.11171*, 2022.

Wang, X., Wang, J., Wang, Y., Dang, P., Cao, S., and Zhang, C. MARS: toward more efficient multi-agent collaboration for llm reasoning. *arXiv preprint arXiv:2509.20502*, 2025.

Welleck, S., Lu, X., West, P., Brahman, F., Shen, T., Khashabi, D., and Choi, Y. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36:21573–21612, 2023.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

APPENDIX

Appendix Overview. The table below summarizes the contents of the appendix so that specific items can be located quickly. Section numbers and page numbers are clickable.

Sec.	Contents	Page
A	**WISE-DS extensions and parameter estimation.** Recency-weighted posterior, expanded log-evidence, EM parameter estimation, and relation to Dawid–Skene.	14
B	**Qualitative results.**	16
B.1	WISE-DS vs. critic-weighted voting on SMART-840: round-by-round debates, critic weights, error matrices, and posterior evidence (four examples).	16
B.2	The same analysis on SMART-840++, with one example per difficulty level (D1–D5).	21
C	**Implementation details.** Debate configuration, early stopping, evaluation, and implementation.	27
D	**Prompts.** Prompts used by the solvers, critics, and the Feedback Distiller.	27
E	**Additional results.** Per-category comparisons of WISE-DS with alternative aggregation methods on VisualPuzzles, SMART-840++, and SMART-840 (Tables 5, 6, 7).	28
E.1	Comparison to feedback-based methods under matched compute and LLM-call budgets.	29
F	**WISE with strong versus weak models.** LLMs/MLLMs used, their abbreviations, and single-model baselines (Tables 9, 10).	29
F.1	Performance with varied solver/critic LLM configurations (Table 11).	31
F.2	Ablation on the critic feedback weights (Table 12).	31
F.3	Robustness under a maximum-token budget (Figure 5).	31
G	**SMART-840++ dataset.** Dataset statistics (Figure 6).	32
H	**SMART-840++ examples.** Example instances of all problem types across difficulty levels (Figure 8).	34

A WISE-DS EXTENSIONS AND PARAMETER ESTIMATION

Recency-weighted posterior. In multi-round debate, later rounds may be more informative because solvers have received feedback from earlier rounds and may have had an opportunity to revise their reasoning. To account for this, we introduce nonnegative round weights $\alpha_1, \ldots, \alpha_K$, where larger α_k gives greater influence to round k. For a candidate answer $y \in \mathcal{A}_p$, define the round-level evidence

$$E^k(y) = \prod_{i=1}^{n_k} \left[\Theta_i^k(y, \hat{a}_i^k) \prod_{j=1}^{m} \Psi_j \left(\mathbb{1}[\hat{a}_i^k = y], w_{ij}^k \right) \right], \tag{3}$$

where n_k is the number of solvers participating in round k, $\Theta_i^k(y, \hat{a}_i^k)$ is the likelihood that solver L_i^S outputs $\hat{a}_i^k$ when the true answer is y, and $\Psi_j(c, w)$ is the likelihood that critic L_j^R assigns weight w when the solver response has correctness status $c \in \{0, 1\}$.

The recency-weighted posterior is then

$$p_p^{(\alpha)}(y) = \frac{\pi(y) \prod_{k=1}^{K} E^k(y)^{\alpha_k}}{\sum_{a \in \mathcal{A}_p} \pi(a) \prod_{k=1}^{K} E^k(a)^{\alpha_k}}, \tag{4}$$

where $\pi(y)$ is the prior over candidate answers. The final prediction is the MAP estimate

$$\widehat{y}_p = \arg\max_{y \in \mathcal{A}_p} p_p^{(\alpha)}(y). \tag{5}$$

Equivalently, in log-space, define

$$L_p(y) = \log \pi(y) + \sum_{k=1}^{K} \alpha_k \log E^k(y). \tag{6}$$

The posterior can be computed as

$$p_p^{(\alpha)}(y) = \frac{\exp L_p(y)}{\sum_{a \in \mathcal{A}_p} \exp L_p(a)}. \tag{7}$$

This form is numerically stable and makes the fusion rule interpretable as a recency-weighted product of experts over debate rounds.

Expanded log-evidence. For implementation, the round-level log-evidence can be written as

$$\log E^k(y) = \sum_{i=1}^{n_k} \log \Theta_i^k(y, \hat{a}_i^k) + \sum_{i=1}^{n_k} \sum_{j=1}^{m} \log \Psi_j \left(\mathbb{1}[\hat{a}_i^k = y], w_{ij}^k \right). \tag{8}$$

Thus, for each candidate answer y, the posterior accumulates solver evidence and critic evidence across all rounds before normalizing over $\mathcal{A}_p$. The solver term measures how likely the observed answer $\hat{a}_i^k$ is under candidate truth y, while the critic term measures how likely the assigned weight w_{ij}^k is under the correctness status induced by y. Note that the debate trajectory is coupled through the shared latent answer y_p. Rather than estimating a separate answer for each round, we require all solver responses and critic weights across rounds to explain the same underlying truth. For each candidate answer y, the model accumulates round-level evidence $E^k(y)$ over the debate trajectory, so that answers consistently supported by multiple solvers and critics receive higher posterior probability. This coupling captures the intuition that reliable debate should progressively concentrate evidence around the correct answer, while isolated or inconsistent answers should have limited influence on the final posterior.

Parameter estimation. The parameters of the posterior fusion model are

$$\Omega = \left\{ \pi,\ \{\Theta_i^k\}_{i,k},\ \{\Psi_j\}_{j=1}^m \right\},$$

where π is the prior over candidate answers, Θ_i^k is the solver-round confusion matrix for solver L_i^S at debate round k, and Ψ_j is the critic verification matrix for critic L_j^R. We estimate these parameters using EM over a collection of Λ debated problems, denoted by $\{p_\ell\}_{\ell=1}^{\Lambda}$.

In the E-step, for each problem p_ℓ, we compute the posterior distribution over its latent true answer,

$$p_{p_\ell}(y) = \Pr\left(y_{p_\ell} = y \mid \mathcal{D}_{p_\ell,1:K}; \Omega\right), \qquad y \in \mathcal{A}_{p_\ell},$$

using Eq. 2. Here $\mathcal{D}_{p_\ell,1:K}$ denotes all solver answers and critic weights observed across debate rounds for problem p_ℓ.

In the M-step, the solver confusion matrices are updated using posterior-weighted answer counts. For solver L_i^S at round k, we update

$$\Theta_i^k(y, a) \leftarrow \frac{\sum_{\ell=1}^{\Lambda} p_{p_\ell}(y)\, \mathbb{1}\left[\hat{a}_{\ell,i}^k = a\right]}{\sum_{\ell=1}^{\Lambda} p_{p_\ell}(y)}, \qquad y, a \in \mathcal{A}_{p_\ell},$$

where $\hat{a}_{\ell,i}^k$ denotes the answer produced by solver L_i^S at round k for problem p_ℓ. This update estimates how likely solver L_i^S is to output answer a at round k when the latent true answer is y.

The critic verification matrices are updated using the posterior probability that each evaluated solver response is correct. Recall that, under candidate truth y, the correctness status of solver L_i^S at round k is

$$c_{\ell,i}^k(y) = \mathbb{1}\left[\hat{a}_{\ell,i}^k = y\right].$$

For each critic L_j^R, correctness status $c \in \{0, 1\}$, and assigned weight $v \in \{0, 1, 2\}$, we update

$$\Psi_j(c, v) \leftarrow \frac{\sum_{\ell=1}^{\Lambda} \sum_{k=1}^{K_\ell} \sum_{i=1}^{n_{\ell,k}} \sum_{y \in \mathcal{A}_{p_\ell}} p_{p_\ell}(y)\, \mathbb{1}\left[c_{\ell,i}^k(y) = c\right]\ \mathbb{1}\left[w_{\ell,ij}^k = v\right]}{\sum_{\ell=1}^{\Lambda} \sum_{k=1}^{K_\ell} \sum_{i=1}^{n_{\ell,k}} \sum_{y \in \mathcal{A}_{p_\ell}} p_{p_\ell}(y)\, \mathbb{1}\left[c_{\ell,i}^k(y) = c\right]}.$$

This update estimates how likely critic L_j^R is to assign weight v when the evaluated solver response is correct ($c = 1$) or incorrect ($c = 0$).

Finally, the answer prior is updated by averaging the posterior responsibilities:

$$\pi(y) \leftarrow \frac{1}{\Lambda} \sum_{\ell=1}^{\Lambda} p_{p_\ell}(y).$$

Dirichlet smoothing may be added to all multinomial updates to avoid degenerate probabilities, especially when some solver-round or critic-weight events are rare.

Relation to Dawid–Skene. The proposed aggregation can be viewed as a debate-aware extension of the classical Dawid–Skene model. If the critic terms Ψ_j are removed and the consistency factors are set to one, i.e., $\Gamma_i^k(\cdot) = 1$, the model reduces to Dawid–Skene aggregation over solver-round answers. The critic likelihood incorporates external assessments of solver correctness, while the consistency factor captures the sequential structure of debate. Thus, the model preserves the central Dawid–Skene idea of estimating latent answer posteriors and annotator reliabilities, but adapts it to the multi-round, feedback-driven setting of WISE.

WISE-DS Implementation. The approach was implemented in Python. We used a maximum of 200 iterations of the EM algorithm and a tolerance of 1e-7. We used a Dirichlet prior with a concentration parameter of 1.0. We use the debate responses from 100 randomly chosen problems in a dataset to estimate the WISE-DS parameters for the solvers and critics.

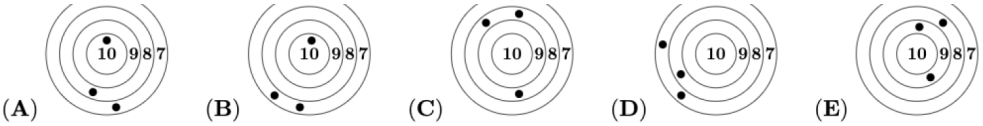
B Qualitative Results

B.1 WISE-DS Qualitative Results on the SMART-840 Dataset

We present four SMART-840 image–text puzzles on which the Critic-weighted vote (Wt. MV) selects an incorrect answer, whereas WISE-DS recovers the correct one. All debates use the configuration (G52+CS46)^2|G52, GPT-5.2 (G52) and Claude-Sonnet-4.6 (CS46) act as both solvers and Critics, and GPT-5.2 serves as the Distiller. Each debate runs for $K{=}4$ rounds. For every round k we report each solver's answer $\hat{a}_i^k$ and the Critic weights $w_{ij}^k \in \{0, 1, 2\}$. We also report the per-response vote weight $v_i^k = \frac{k}{K} \cdot \frac{1}{2m} \sum_j w_{ij}^k$ used by Wt. MV, which selects $\arg\max_a \sum_{i,k:\hat{a}_i^k=a} v_i^k$. WISE-DS is fitted separately for each grade over all puzzles of that grade. It uses the scalar-reliability form of the solver confusion matrices, $\Theta_i^k(a, a) = q_i^k$ and $\Theta_i^k(a, b) = \frac{1-q_i^k}{4}$ for $b \neq a$, which is monotone non-decreasing in k. Each Critic has a verification matrix $\Psi_j(c, w)$. To explain each decision, we decompose the unnormalized log-posterior of Eq. (2) for a candidate y into a prior term $\log \pi(y)$, a solver term $\sum_{i,k} \log \Theta_i^k(y, \hat{a}_i^k)$, and a Critic term $\sum_{i,j,k} \log \Psi_j(\mathbb{1}[\hat{a}_i^k{=}y], w_{ij}^k)$. Correct answers are shown in **green** and incorrect ones in **red**.

Across the four examples, WISE-DS corrects Wt. MV through two complementary mechanisms: (i) *Critic calibration* (Examples 1–2): Ψ learns that CS46-as-Critic frequently endorses wrong answers, e.g., $\Psi_{\text{CS46}}(0, 2) = 0.17$ vs. $\Psi_{\text{G52}}(0, 2) = 0.06$ in grades 3–4, whereas a zero from the strict G52 Critic is strong evidence of a miss ($\Psi_{\text{G52}}(0, 0) \approx 0.9$). Wt. MV treats a zero weight as no vote and so discards this signal; (ii) *solver reliability and cross-round consistency* (Examples 3–4): when neither Critic is informative, Θ favors the solver that is more reliable across the dataset and consistent across rounds.

EXAMPLE 1: SMART-840, GRADES 3–4, MK-2021-3-4-5 (DISTILLER: GPT-5.2)



Question: Five boys competed in a shooting challenge. Ricky scored the most points. Which target was Ricky's? **Options:** (A)–(E) as shown. **Ground truth:** **E** (10+9+8).

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	**E**	2	2	0.250	**D**	0	1	0.063
2	**E**	2	0	0.250	**D**	0	2	0.250
3	**D**	0	0	0.000	**D**	0	2	0.375
4	**E**	1	1	0.500	**D**	0	2	0.500

Wt. MV totals: **D**= 1.188, **E**= 1.000 ⇒ Wt. MV selects **D**. WISE-DS posterior: **E**= 0.9996, **D**= 0.0004 ⇒ **E**.

Rd	Solver G52	Solver CS46	Critics	Distiller → next round
1	Scores A=25, B=25, C=23, D=21, E=27 (10+9+8); picks **E**.	Counts *four* shots on D (≈33 pts); picks **D**.	G52: "D is mis-scored but E is still clearly highest" (2); D has 3 shots, not 4 (0). CS46: E correct (2); D "may be correct if D has 4 shots" (1).	To G52: re-check D's rings; could D exceed 27? To CS46: recount—D appears to have only 3 shots.
2	Re-scores D=7+8+8=23; E=27 still highest; **E**.	Insists D has 4 shots (10+9+8+7=34); **D**.	CS46 on own answer: "D4 is correct" (2); on G52: "missed D's fourth shot, answer should be D" (0). G52: E correct (2), D miscounted (0).	To G52: *"One critic claims Target D clearly displays four shots"*—inspect D. To CS46: are there really 4 shots?
3	Adopts the 4-shot reading, D=31; switches to **D**.	D=32, highest; **D**.	G52: D miscounted for both solvers (0, 0). CS46: own D correct (2); G52's D wrong, "C and E tie at 27" (0).	To G52: recount dots, especially on D. To CS46: G52 says D has 3 shots.
4	Recounts: D has 3 dots (23); E=27 uniquely highest; back to **E**.	Keeps 4-shot D (≈32); **D**.	G52: D wrong (0); E "may be correct" (1). CS46: own D correct (2); E "may be correct" (1).	–

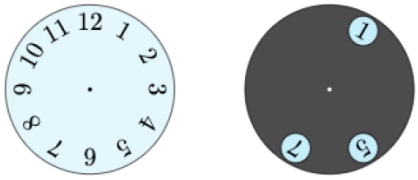
q_i^k (grades 3–4)	k=1	k=2	k=3	k=4
G52	0.694	0.694	0.729	0.755
CS46	0.532	0.532	0.546	0.546

$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.194	0.043	0.763
G52, c=0	0.911	0.032	0.057
CS46, c=1	0.119	0.054	0.827
CS46, c=0	0.533	0.297	0.170

y	prior	solver	Critic	total
E	−1.82	−12.35	**−16.55**	**−30.72**
D	−1.32	**−10.72**	−26.61	−38.66

Analysis. CS46 hallucinates a fourth shot on target D, and CS46-as-Critic endorses this answer with w=2 in rounds 2–4. Through the Distiller, the claim also pulls G52 to D in round 3, before G52 recounts and returns to E in round 4. Because v grows with k, these late self-endorsements give D the larger Wt. MV total. The solver term of WISE-DS also slightly favors D, since D receives 5 of the 8 votes. The Critic term, however, favors E by 10.1 nats. WISE-DS has learned that a w=2 from CS46 is weak evidence of correctness ($\Psi_{\text{CS46}}(1,2)/\Psi_{\text{CS46}}(0,2) = 4.9$), while G52's five zeros on D are strong evidence of a miss ($\Psi_{\text{G52}}(0,0)/\Psi_{\text{G52}}(1,0) = 4.7$ each). The posterior thus concentrates on E (0.9996).

EXAMPLE 2: SMART-840, GRADES 5–6, MK-2023-5-6-9 (DISTILLER: GPT-5.2)



Question: The gray circle with three holes punched in it is placed on top of the clock face. The gray circle is turned around its center. Which three numbers is it possible to see at the same time? **Options:** (A) 2, 4, 9; (B) 1, 5, 10; (C) 4, 6, 12; (D) 3, 6, 9; (E) 5, 7, 12. **Ground truth:** C.

	Solver G52				Solver CS46			
Round	$\hat{a}$	w_{G52}	w_{CS46}	v	$\hat{a}$	w_{G52}	w_{CS46}	v
1	**D**	0	0	0.000	**D**	0	0	0.000
2	**D**	0	0	0.000	**D**	0	1	0.125
3	**C**	0	0	0.000	**D**	0	1	0.188
4	**C**	0	2	0.500	**D**	0	1	0.250

Wt. MV totals: **D**= 0.563, **C**= 0.500 ⇒ Wt. MV selects **D**. WISE-DS posterior: **C**= 0.734, **D**= 0.252 ⇒ **C**.

Rd	Solver G52	Solver CS46	Critics	Distiller → next round
1	Reads the holes as 1, 7, 8; finds that no option matches the gap pattern; makes a forced choice of **D**.	Assumes evenly spaced (90°) holes; picks **D** (3, 6, 9).	All weights 0. Both Critics note holes near 1, 5 and 7 o'clock, so D's symmetric spacing is inconsistent; G52's forced guess is unjustified.	To G52: re-check which numbers are visible. To CS46: holes are near 1, 5, 7 o'clock; compare angular gaps.
2	Re-reads as 1, 2, 3; still no match; forced **D**.	"L-shaped" 3–6–9 pattern; **D**.	CS46: own D "may or may not be correct" (1); all others 0.	To G52: re-examine hole positions/rotations. To CS46: compute hole angles systematically.
3	Re-reads as 1, 4, 6, gaps $\{2, 3, 7\}$; no match; forced **C**.	Keeps **D**.	CS46: own D "methodology correct" (1); others 0; G52's arbitrary choice criticized.	To G52: critics read 1, 7, 3; recompute the gaps. To CS46: use exact clock positions; spacing is irregular.
4	Reads 1, 7, 3 ⇒ gaps $\{2, 4, 6\}$; only C (4, 6, 12) matches; **C**.	Keeps **D**.	CS46: "if the holes show 1, 3, 7, C is correct" (2); own D (1). G52: 0 to both (misread start configuration; D mismatched).	–

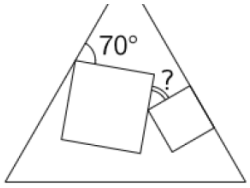
q_i^k (grades 5–6)	k=1	k=2	k=3	k=4
G52	0.732	0.828	0.828	0.828
CS46	0.561	0.594	0.594	0.607

$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.410	0.035	0.554
G52, c=0	0.897	0.026	0.077
CS46, c=1	0.229	0.166	0.605
CS46, c=0	0.485	0.289	0.227

y	prior	solver	Critic	total
C	−1.33	−15.33	**−10.31**	**−26.98**
D	−1.55	**−8.92**	−17.58	−28.05

Analysis. Both solvers start at D. G52 corrects its reading of the holes over three rounds and reaches C in rounds 3–4, while CS46 never revisits its symmetric 3–6–9 assumption. Wt. MV follows CS46-as-Critic's lenient weight of 1 on D in three rounds and narrowly prefers D (0.563 vs. 0.500). The solver term of WISE-DS prefers D by 6.4 nats because D receives 6 of the 8 votes. The Critic term overturns this by 7.3 nats. G52-as-Critic assigns w=0 to all six D responses, and under Ψ_{G52} a zero is about 2.2× more likely for an incorrect answer than for a correct one. CS46's weight of 1 is also more likely under a miss (0.289 vs. 0.166), and CS46's weight of 2 on C in round 4 is 2.7× more likely under a hit.

EXAMPLE 3: SMART-840, GRADES 9–10, MK-2020-9-10-17 (DISTILLER: GPT-5.2)



Question: Two squares of different size are drawn inside an equilateral triangle. One side of one of these squares lies on one of the sides of the triangle, as shown. What is the size of the angle marked by the question mark? **Options:** (A) 25°; (B) 30°; (C) 35°; (D) 45°; (E) 50°. **Ground truth:** **E**.

	Solver G52				Solver CS46			
Round	$\hat{a}$	w_{G52}	w_{CS46}	v	$\hat{a}$	w_{G52}	w_{CS46}	v
1	**E**	0	1	0.063	**B**	0	1	0.063
2	**E**	0	0	0.000	**B**	0	2	0.250
3	**E**	0	0	0.000	**B**	0	0	0.000
4	**E**	0	0	0.000	**B**	0	1	0.250

Wt. MV totals: **B**= 0.563, **E**= 0.063 ⇒ Wt. MV selects **B**. WISE-DS posterior: **E**= 0.607, **B**= 0.371 ⇒ **E**.

Rd	Solver G52	Solver CS46	Critics	Distiller → next round
1	Angle-chases from the 70° mark (80°, then −30°); **E** (50°), but with an unexplained 30° step.	"90° − 60°" complement at the triangle vertex; **B** (30°).	G52: 0 to both (unjustified 30° step; diagram misdescribed). CS46: 1 to both ("answer may be correct, derivation flawed").	To G52: re-locate the 70° angle; justify angle relations. To CS46: squares are rotated; use the 70° angle.
2	Redoes the chase with directed angles (10° offset between the triangle and square sides); **E**.	Assumes the small square sits flat on the base: 90° − 60°; **B**.	CS46: own B "correct despite flawed logic" (2); E "incorrect, standard answer is 30°" (0). G52: 0 to both (B assumes wrong orientation and ignores 70°).	To G52: verify the 80° step. To CS46: the small square is tilted; use the 70° angle.
3	Keeps **E**: "although feedback suggested 30°, the geometry gives 50°".	Same flat-square argument; **B**.	All weights 0 (CS46 now also rejects its own B as a misreading).	To G52: justify the −10° orientation. To CS46: re-evaluate the tilt of the small square.
4	**E** (70° → square side → connector direction).	60° − 30° at the vertex; **B**.	CS46: own B "might be correct" (1); E "lacks rigor" (0). G52: 0 to both.	–

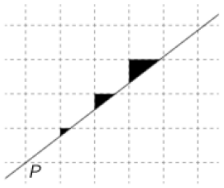
q_i^k (grades 9–10)	k=1	k=2	k=3	k=4
G52	0.848	0.913	0.921	0.921
CS46	0.764	0.769	0.769	0.769

$\Psi_j(c,w)$	w=0	w=1	w=2
G52, c=1	0.294	0.050	0.656
G52, c=0	0.897	0.015	0.089
CS46, c=1	0.129	0.100	0.771
CS46, c=0	0.541	0.299	0.160

y	prior	solver	Critic	total
E	−1.84	**−11.81**	−18.64	**−32.28**
B	−1.47	−16.01	**−15.29**	−32.77

Analysis. The two solvers never agree: G52 answers E and CS46 answers B in every round. G52-as-Critic rejects both, and CS46-as-Critic repeatedly endorses its own answer B, with weight 2 in round 2 and weight 1 in rounds 1 and 4. Wt. MV therefore selects B by a wide margin (0.563 vs. 0.063). In WISE-DS the Critic term still favors B, but only by 3.4 nats, because Ψ discounts CS46's endorsements. The solver term favors E by 4.2 nats: across grades 9–10, G52 is the more reliable solver in later rounds (q=0.92 vs. 0.77), and four consistent votes from it outweigh four from CS46. WISE-DS thus selects the correct answer E (p=0.607). The posterior retains substantial mass on B (0.371), reflecting genuine disagreement.

EXAMPLE 4: SMART-840, GRADES 11–12, MK-2020-11-12-22 (DISTILLER: GPT-5.2)



Question: On a square grid paper, a little kangaroo draws a line passing through the lower left corner P of the grid and colors in three triangles as shown. Which of the following could be the ratio of the areas of the triangles? **Options:** (A) 1:2:3; (B) 1:2:4; (C) 1:3:9; (D) 1:4:8; (E) None of the above. **Ground truth:** **E**.

	Solver G52				Solver CS46			
Round	$\hat{a}$	w_{G52}	w_{CS46}	v	$\hat{a}$	w_{G52}	w_{CS46}	v
1	**E**	0	1	0.063	**C**	0	1	0.063
2	**E**	0	1	0.125	**D**	0	0	0.000
3	**E**	0	0	0.000	**C**	0	1	0.188
4	**E**	0	0	0.000	**C**	0	1	0.250

Wt. MV totals: **C**= 0.500, **E**= 0.188, **D**= 0 ⇒ Wt. MV selects **C**. WISE-DS posterior: **E**= 0.999, **C**= 0.0006 ⇒ **E**.

Rd	Solver G52	Solver CS46	Critics	Distiller → next round
1	Reads legs 1, 2, 3 ⇒ areas 1:4:9, which is not listed; **E**.	Speculates "quadratic growth" with legs 1, $\sqrt{3}$, 3; **C** (1:3:9).	G52: 0 to both (dimensions misread; C is a guess). CS46: 1 to both ("reasonable if the interpretation is correct"; "speculative").	To G52: critics suggest the two smaller triangles both have legs of 1; re-check. To CS46: find the grid intersections and compute bases and heights.
2	Re-measures 1×1, 1×1, 2×2 ⇒ 1:1:4, not listed; **E**.	Infers areas 1, 4, 8; **D**.	G52: 0 to both (1:1:4 contradicts image; D contradicts 1:4:9). CS46: E "internally consistent" (1); own D (0).	To G52: is triangle 2 really the same size as triangle 1? To CS46: how would 1×1, 2×2, 3×3 triangles change 1:4:8?
3	Back to legs 1, 2, 3 ⇒ 1:4:9; **E**.	Legs "1 : $\sqrt{3}$: 3"; **C**.	G52: 0 to both. CS46: E "appears incorrect" (0); own C "might be correct" (1).	To G52: re-derive leg lengths. To CS46: compute the slope and the triangle areas explicitly.
4	1:1:4, not listed; **E**.	"Quadratic growth"; **C**.	G52: 0 to both. CS46: E "likely incorrect, answer is probably C or B" (0); own C (1).	–

q_i^k (grades 11–12)	k=1	k=2	k=3	k=4
G52	0.875	0.944	0.944	0.944
CS46	0.713	0.713	0.713	0.713

$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.386	0.054	0.560
G52, c=0	0.910	0.018	0.071
CS46, c=1	0.172	0.121	0.707
CS46, c=0	0.394	0.270	0.336

y	prior	solver	Critic	total
E	−1.53	**−10.84**	−16.79	**−29.16**
C	−1.63	−19.92	**−15.08**	−36.63

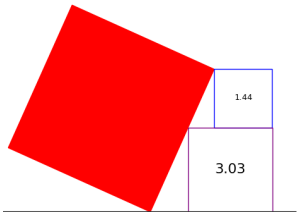
Analysis. G52 answers E in all four rounds, although its intermediate area ratio alternates between 1:4:9 and 1:1:4. CS46 drifts between C and D and justifies C by an unsupported "quadratic growth" argument. CS46-as-Critic gives its own C a hedged weight of 1 in rounds 1, 3 and 4. Because v grows with k, Wt. MV selects C (0.500 vs. 0.188). WISE-DS instead learns that for CS46-as-Critic a weight of 1 is more likely on incorrect responses than on correct ones (0.270 vs. 0.121), so these hedged endorsements carry almost no support. The Critic term differs by only 1.7 nats. The solver term decides the outcome: G52's four consistent votes at reliability q=0.94 outweigh CS46's inconsistent C/D votes at q=0.71 by 9.1 nats, yielding E with posterior 0.999.

B.2 WISE-DS ON SMART-840++: ONE EXAMPLE PER DIFFICULTY LEVEL

We next show one SMART-840++ puzzle from each of the five difficulty levels D1–D5 on which Wt. MV selects an incorrect answer and WISE-DS recovers the correct one. All debates use (G52+CS46)^2|G52: GPT-5.2 (G52) and Claude-Sonnet-4.6 (CS46) act as both solvers and reflectors, and GPT-5.2 serves as the Feedback Distiller, for $K{=}4$ rounds. WISE-DS is fitted separately for each difficulty level, over all puzzles of that level. Notation and the evidence decomposition follow Section B.1.

Across all five levels, one pattern dominates. CS46-as-reflector most often assigns the hedging weight $w{=}1$, and WISE-DS learns that this weight is *more* likely on incorrect responses than on correct ones: $\Psi_{\text{CS46}}(0,1) \in [0.48, 0.68]$ vs. $\Psi_{\text{CS46}}(1,1) \in [0.28, 0.44]$. Wt. MV counts every such weight as a positive vote, and so rewards the solver that repeats a wrong answer while its own reflector hedges. WISE-DS instead treats these weights as weak evidence *against* the endorsed answer. G52-as-reflector is strict, $\Psi_{\text{G52}}(0,0) \approx 0.86$–$0.92$, and its rare $w{=}2$ is highly diagnostic of a correct answer. Example D5 shows that WISE-DS does not simply trust the stronger model: there CS46 is correct and G52 is wrong.

EXAMPLE D1: SMART-840++, DIFFICULTY D1, MK-2020-11-12-17-D1-2



Question: Hannah drew three squares. The big square is in contact with the other two squares, as shown in the diagram. The numbers inside the smaller squares indicate their areas. What is the area of the big square? **Options:** (A) 0.0; (B) 25.39; (C) 10.39; (D) 8.18; (E) 47.6. **Ground truth:** **C**.

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	**D**	0	1	0.063	**D**	0	1	0.063
2	**C**	0	2	0.250	**D**	0	1	0.125
3	**E**	0	0	0.000	**D**	0	1	0.188
4	**C**	0	1	0.250	**D**	0	1	0.250

Wt. MV totals: **D**= 0.688, **C**= 0.500, **E**= 0 ⇒ Wt. MV selects **D**. WISE-DS posterior: **C**= 0.956, **A**= 0.020, **D**= 0.001 ⇒ **C**.

Rd	Solver G52	Solver CS46	Reflectors	Distiller → next round
1	Assumes side $= \sqrt{3.03} + \sqrt{1.44}$, area ≈ 8.65; picks the "closest" option **D** (8.18).	Same formula, area ≈ 8.65; **D**.	G52: 0 to both (contact points misused; $8.65 \neq 8.18$). CS46: 1 to both ("D4 is the closest option").	To G52: which exact corners does the red square touch? To CS46: which vertices coincide, and does that segment lie on one side of the red square?
2	The red side passes through the top-left corners of both squares, and the upper square is offset to the right; area = 10.39; **C**.	Repeats $(\sqrt{a} + \sqrt{b})^2 \approx 8.65$; **D**.	CS46: C "matches the correct geometric calculation" (2); own D (1). G52: 0 to both (thinks 47.6 is more plausible).	To G52: verify which two corners the red edge passes through, as one critic disputes. To CS46: list the touch points; are the small squares offset?
3	Adopts the reflector's alternative reading and a "visual sanity check"; **E** (47.6).	Same formula; **D**.	CS46: E wrong, "the correct calculation gives ≈10.39, C3" (0); own D (1). G52: 0 to both.	To G52: which corners are collinear with the red edge, and where is its bottom vertex? To CS46: which points are endpoints of one red side?
4	Places the bottom vertex at the corner of the 3.03 square and computes from the confirmed contacts; **C**.	New formula $2(A_1+A_2) = 8.94$; **D**.	CS46: C "approximately right" (1); own D "elegant" (1). G52: 0 to both.	–

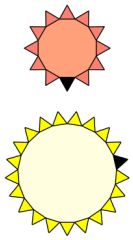
q_i^k (D1)	k=1	k=2	k=3	k=4
G52	0.720	0.720	0.720	0.772
CS46	0.432	0.478	0.499	0.499

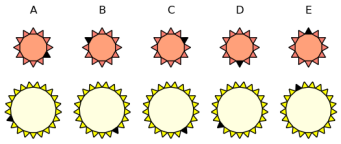
$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.284	0.030	0.686
G52, c=0	0.895	0.015	0.090
CS46, c=1	0.073	0.278	0.648
CS46, c=0	0.306	0.625	0.069

y	prior	solver	reflector	total
C	−1.51	−14.05	**−8.43**	**−23.99**
D	−1.79	**−11.47**	−17.35	−30.61

Analysis. CS46 repeats the same unjustified formula in all four rounds and picks the "closest" option, D. CS46-as-reflector gives each of these responses w=1, so D collects five small votes and wins Wt. MV (0.688 vs. 0.500). WISE-DS's solver term also favors D, by 2.6 nats, because D receives 5 of the 8 votes. The reflector term reverses this by 8.9 nats. At D1, a weight of 1 from CS46 is 2.2× more likely on an incorrect response (0.625 vs. 0.278), so the five hedged endorsements count against D. CS46's weight of 2 on G52's C in round 2 is 9.4× more likely on a correct one. G52-as-reflector's zeros, which fall on five D responses but only two C responses, further favor C (p=0.956).

EXAMPLE D2: SMART-840++, DIFFICULTY D2, MK-2021-1-2-20-D2-3





Question: The image shows a pair of gears, each marked by a single black tooth. After meshing them the upper gear rotates 4 teeth counter-clockwise, where will the black teeth be? **Options:** (A)–(E) as shown. **Ground truth:** C.

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	C	0	1	0.063	B	0	1	0.063
2	B	0	1	0.125	B	0	1	0.125
3	C	2	0	0.375	D	0	1	0.188
4	E	0	1	0.250	B	0	1	0.250

Wt. MV totals: B= 0.563, C= 0.438, E= 0.250, D= 0.188 ⇒ Wt. MV selects B.
WISE-DS posterior: C= 0.895, E= 0.058, B= 0.017 ⇒ C.

Rd	Solver G52	Solver CS46	Reflectors	Distiller → next round
1	Upper black tooth moves to the upper right, lower tooth to the lower area; C.	Upper tooth moves to the upper left (≈10 o'clock); B.	G52: 0 to both (upper-gear position derived incorrectly; B inconsistent with gear motion). CS46: 1 to both ("plausible").	To G52: recount the teeth (10 vs. ≈12); verify the CCW direction from 6 o'clock. To CS46: are the gears meshed? Give exact tooth counts.
2	"Corrects" the CCW tracing and matches upper right + lower right to B.	Same reasoning; B.	G52: 0 to both; notes that upper right + lower right is option C, not B. CS46: 1 to both ("B may accidentally be correct").	To G52: re-check the CCW direction on the page; count the teeth. To CS46: find the contact point; give exact tooth counts.
3	Recounts 10/20 teeth, tracks the marked tooth tooth-by-tooth, and finds that C, not B, matches; C.	D.	G52: "option C indeed shows upper right, lower right" (2); D does not match the rotation (0). CS46: C has "directional errors", B is better (0); own D (1).	To G52: 10/20 or 12/24 teeth? Define one tooth step. To CS46: distinguish D and E by clock position.
4	Recounts as 12/24 teeth; 120° vs. 60°; E.	Back to B.	G52: 0 to both (E rests on a faulty count). CS46: 1 to both.	–

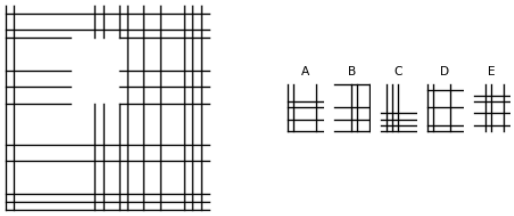
q_i^k (D2)	k=1	k=2	k=3	k=4
G52	0.718	0.720	0.729	0.781
CS46	0.487	0.487	0.541	0.572

$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.354	0.044	0.602
G52, c=0	0.912	0.017	0.071
CS46, c=1	0.080	0.309	0.611
CS46, c=0	0.295	0.640	0.065

y	prior	solver	reflector	total
C	−1.56	−14.72	**−8.47**	**−24.75**
B	−1.64	**−12.74**	−14.34	−28.72

Analysis. G52 starts with the correct answer C, is talked into B in round 2, and recovers C in round 3 after recounting the teeth, before drifting to E in round 4. CS46 mostly answers B. Every B response receives the hedging weight 1 from CS46-as-reflector, so B wins Wt. MV (0.563 vs. 0.438). The solver term favors B by 2.0 nats, since B receives 4 of the 8 votes. The reflector term favors C by 5.9 nats. G52's single weight of 2 on C in round 3 is 8.5× more likely on a correct response, while CS46's weights of 1 on B are 2.1× more likely on incorrect responses. WISE-DS thus selects C (0.895).

EXAMPLE D3: SMART-840++, DIFFICULTY D3, MK-2023-7-8-1-D3-3



Question: Horizontal and vertical lines are illustrated in the diagram, with one piece removed. Which of these could be the missing piece? **Options:** (A)–(E) as shown. **Ground truth: B**.

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	**B**	2	1	0.188	**A**	0	0	0.000
2	**B**	0	1	0.125	**A**	0	1	0.125
3	**A**	0	1	0.188	**A**	0	1	0.188
4	**B**	0	1	0.250	**A**	0	1	0.250

Wt. MV totals: **A**= 0.750, **B**= 0.563 ⇒ Wt. MV selects **A**. WISE-DS posterior: **B**= 0.941, **A**= 0.051 ⇒ **B**.

Rd	Solver G52	Solver CS46	Reflectors	Distiller → next round
1	B has three verticals on the right and three horizontals from the left, matching the gap; C and E are too dense and D is misaligned; **B**.	Places the hole at the "upper-left corner" with sparse verticals; **A**.	G52: own B consistent with the image (2); A incorrect (0). CS46: B "plausible" (1); on its own A, "B appears to be a better match" (0).	To G52: re-locate the hole; list every line touching each edge. To CS46: is the hole a corner or an interior cut-out?
2	Re-checks all four edges; **B**.	**A**.	G52: 0 to both (suggests C). CS46: 1 to both.	Both: enumerate the line stubs on each edge of the gap.
3	Claims it "overcounted" the boundary lines; switches to **A**.	**A**.	G52: 0 to both (A too sparse). CS46: 1 to both.	Both: do the vertical stubs on the top and bottom edges align?
4	Counts 3 vertical stubs on the top and bottom and 3 horizontal stubs on the left; rules out A; **B**.	Horizontal-only pattern; **A**.	G52: 0 to both. CS46: B "reasonable" (1); own A "questionable" (1).	–

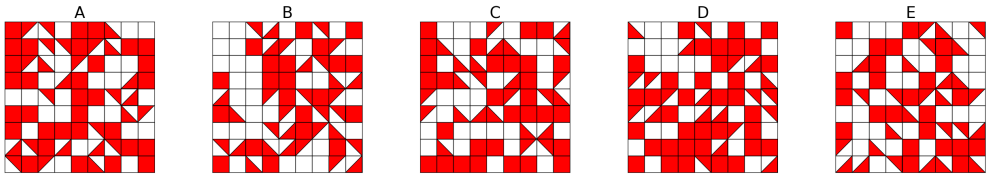
q_i^k (D3)	k=1	k=2	k=3	k=4
G52	0.587	0.665	0.665	0.753
CS46	0.563	0.593	0.593	0.593

$\Psi_j(c,w)$	w=0	w=1	w=2
G52, c=1	0.469	0.058	0.473
G52, c=0	0.858	0.017	0.125
CS46, c=1	0.061	0.386	0.553
CS46, c=0	0.440	0.484	0.076

y	prior	solver	reflector	total
B	−1.44	−12.78	**−9.60**	**−23.82**
A	−1.70	**−10.08**	−14.95	−26.73

Analysis. CS46 answers A in every round, and G52 joins it once, in round 3. A therefore receives 5 of the 8 votes, and CS46-as-reflector's steady weights of 1 make A the Wt. MV winner (0.750 vs. 0.563). The solver term also favors A, by 2.7 nats. The reflector term favors B by 5.4 nats, driven by the most informative weights. In round 1, CS46-as-reflector rejects its own answer A with w=0, which is 7.2× more likely on an incorrect response (0.440 vs. 0.061). G52-as-reflector endorses B with w=2 (3.8× more likely on a correct response) and scores all five A responses 0 (1.8× more likely on an incorrect one). Wt. MV discards zero weights entirely and so cannot use any of this evidence. WISE-DS selects B (0.941).

EXAMPLE D4: SMART-840++, DIFFICULTY D4, MK-2020-3-4-4-D4-2



Question: Paul sketched some shapes. Which one has the most shaded area? **Options:** (A)–(E) as shown. **Ground truth:** **D**.

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	**D**	2	0	0.125	**A**	0	1	0.063
2	**D**	1	0	0.125	**A**	0	1	0.125
3	**D**	0	1	0.188	**A**	0	1	0.188
4	**E**	0	1	0.250	**A**	0	1	0.250

Wt. MV totals: **A**= 0.625, **D**= 0.438, **E**= 0.250 ⇒ Wt. MV selects **A**. WISE-DS posterior: **D**= 0.988, **E**= 0.005, **A**= 0.002 ⇒ **D**.

Rd	Solver G52	Solver CS46	Reflectors	Distiller → next round
1	Weights full red cells as 1 and half-cells as 0.5 and compares the options; **D**.	Judges visual density ("A has the densest red coverage"); **A**.	G52: D "appears correct" (2); A incorrect, D looks more shaded (0). CS46: D incorrect, A looks denser (0); own A (1).	Both: give explicit counts of full and half cells for A vs. D; are all grids the same size?
2	Recount: A's dense region is mostly half-cells with more white gaps; **D**.	"A clearly shows the most red"; **A**.	G52: own D plausible but unverified (1); A (0). CS46: D (0); own A (1).	Both: tally every option (full, half) so a critic can reproduce it.
3	Tallies each option (e.g., E ≈36.5 units); D is largest; **D**.	Coverage estimates; **A**.	G52: 0 to both. CS46: D (1, "E could rival D"); own A (1).	To G52: are the grids 9×9? Recount E. To CS46: count instead of estimating density.
4	Reconsiders D vs. E and switches to **E** ("at least as dense as D").	**A**.	G52: 0 to both ("D appears to have the greatest shaded area"). CS46: 1 to both.	–

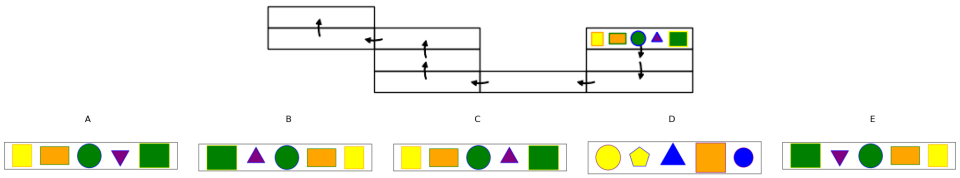
q_i^k (D4)	k=1	k=2	k=3	k=4
G52	0.666	0.696	0.761	0.764
CS46	0.473	0.473	0.473	0.473

$\Psi_j(c, w)$	w=0	w=1	w=2
G52, c=1	0.398	0.044	0.558
G52, c=0	0.916	0.033	0.051
CS46, c=1	0.085	0.372	0.543
CS46, c=0	0.273	0.683	0.045

y	prior	solver	reflector	total
D	−1.42	**−11.98**	**−12.88**	**−26.29**
A	−1.52	−13.70	−17.58	−32.80

Analysis. CS46 never counts cells and answers A in every round from visual impression, and CS46-as-reflector gives each A response w=1. Because v grows with k, these late hedged endorsements outweigh G52's early, well-supported D (Wt. MV: 0.625 vs. 0.438). In WISE-DS both terms favor D. The solver term favors D by 1.7 nats, because G52 is the more reliable solver at this level (q=0.67–0.76 vs. 0.47). The reflector term favors D by 4.7 nats: at D4, CS46's weight of 1 is 1.8× more likely on incorrect responses, and G52's round-1 weight of 2 on D is 10.9× more likely on correct ones. WISE-DS selects D (0.988), even though G52 itself abandons D in the final round.

EXAMPLE D5: SMART-840++, DIFFICULTY D5, MK-2020-1-2-15-D5-1



Question: Carl performs repeated flips on a set of cards. What's their final look? **Options:** (A)–(E) as shown (bottom row); the top diagram shows the flips. **Ground truth:** **E**.

Round	Solver G52 $\hat{a}$	w_{G52}	w_{CS46}	v	Solver CS46 $\hat{a}$	w_{G52}	w_{CS46}	v
1	**E**	0	2	0.125	**E**	0	2	0.125
2	**A**	0	1	0.125	**E**	0	2	0.250
3	**A**	0	1	0.188	**E**	0	1	0.188
4	**A**	2	1	0.750	**E**	0	1	0.250

Wt. MV totals: **A**= 1.063, **E**= 0.938 ⇒ Wt. MV selects **A**. WISE-DS posterior: **E**= 0.997, **A**= 0.003 ⇒ **E**.

Rd	Solver G52	Solver CS46	Reflectors	Distiller → next round
1	Overall reversal of the strip, and the triangle ends up pointing down; **E**.	Three flips (an odd number) reverse the order; **E**.	CS46: 2 to both ("coherent"; "matches a single reversal"). G52: 0 to both (triangle orientation guessed; flip sequence not analyzed).	To G52: label each flip; is it a full reversal or a fold of a segment? To CS46: triangle orientation? 4 or 5 cards? Count the flips.
2	Folds affect only segments, so the order is preserved and only the triangle flips; **A**.	Odd flip count ⇒ reversed; **E**.	CS46: own E (2); A "reasonable interpretation" (1). G52: 0 to both (speculative).	To G52: write the sequence after every flip step. To CS46: where are "3 flips" in the picture?
3	Folds at the marked lines; rules out B and E; **A**.	**E**.	CS46: 1 to both. G52: 0 to both.	To G52: which folds reach the symbol block? To CS46: 4 or 5 cards? Trace the triangle.
4	Tracks icon 1 relative to icon 5: order preserved; **A**.	**E**.	G52: own A "likely correct" (2); E unjustified (0). CS46: 1 to both.	–

q_i^k (D5)	k=1	k=2	k=3	k=4
G52	0.629	0.695	0.695	0.695
CS46	0.601	0.601	0.647	0.647

$\Psi_j(c,w)$	w=0	w=1	w=2
G52, c=1	0.493	0.039	0.467
G52, c=0	0.879	0.026	0.095
CS46, c=1	0.070	0.441	0.490
CS46, c=0	0.304	0.631	0.064

y	prior	solver	reflector	total
E	−1.73	**−10.08**	**−11.30**	**−23.11**
A	−1.50	−12.93	−14.43	−28.87

Analysis. Here the roles reverse: CS46 answers E correctly in every round, while G52 abandons E after round 1 and settles on A. In round 4, G52-as-reflector endorses its own A with w=2. With the round-4 factor k/K=1, this single response receives the largest vote weight of the debate (v=0.75) and flips Wt. MV to A (1.063 vs. 0.938). WISE-DS weighs this endorsement against the rest of the evidence. G52's weight of 2 on A is 4.9× more likely on a correct response. However, CS46's three weights of 2 on E in rounds 1–2 are each 7.7× more likely on a correct response, and CS46's weights of 1 on A count against A. Together, the reflector term favors E by 3.1 nats. The solver term adds 2.9 nats for E, since E receives 5 of the 8 votes from solvers of similar reliability. WISE-DS selects E (0.997). This example shows that WISE-DS does not simply defer to the solver that is stronger on average.

C Implementation Details

Debate Configuration: We ran up to a maximum of 4 debate rounds in all our experiments, with early stopping upon consensus (i.e., when all solvers produce the same response and all critics declare the answer is correct). For solution aggregation, we compare against methods implemented in the Crowd-Kit toolbox[2], specifically the original Dawid-Skene model.

Evaluation and Implementation Details: We evaluate with multiple-choice accuracy on SMART-840 and VisualPuzzles, and use the Phi-4 model to compare predicted and ground-truth answers on EvoChart-QA, after manually verifying Phi-4's judging quality. WISE is implemented in Python with asynchronous communication modules so that messages among solvers, critics, and the distiller are exchanged in parallel. We use VLMs from HuggingFace or via APIs, running experiments on L40 and A100 nodes with one dedicated GPU per VLM. Prompts are identical across all debate steps and models. Standard deviation is computed over at least three runs.

D Prompts used in our MAD formulations

Below, we list the different prompts (discussed in the main paper) that are employed at each stage of our framework.

PROMPT_1^S (Solver)

Please describe the problem, including the images if available. Your description of the image should be detailed and accurate so that an LLM without access to the image can solve the problem. After this, please solve this question with explanation of the intermediate steps. You must select an answer from one of the five options:
Puzzle: [PROBLEM]...[/PROBLEM]

Options: [OPTIONS] ... [/OPTIONS].

Write the final **answer option** as one of `A1, B2, C3, D4, E5`.

PROMPT^C (critic)

Following are the solutions of LLM in solving a math puzzle. The problem statement and the LLM solution are provided below. The problem may contain an image that needs to be used to understand the LLM's answer. You should not solve the problem yourself. Instead, please check if the solutions below are logically and mathematically consistent and the final answer is correct?

$[PROBLEM]$ ... $[/PROBLEM]$.

$[RESPONSE]$ ... $[/RESPONSE]$.

Your Task: Provide a detailed explanation of your assessment so that it can be used as feedback to LLM to improve the solution, including any mistakes in understanding the given image (if any).

Your response should end with a single line in the format:

$FINAL_SCORE : \langle value \rangle$

where $\langle value \rangle$ is 0 if final answer is the wrong, 1 if you are unable to confirm, and 2 if the final answer is correct.

[2]https://github.com/Toloka/crowd-kit

PROMPT^{D} (distiller)

The following are the feedback received from LLMs on a puzzle solution. There could be significant differences in the feedback provided by different LLMs, especially the LLMs may select different answer options. Your task is to pay attention to these differences in the feedback and the selected answer options. Please summarize all the feedback and provide a **set of actionable questions** for the solver to revise its solution in the next round. It is possible that the previous solution might have overlooked details in the given image (if any), and thus your summary should emphasize the importance of re-evaluating the image.

$[FEEDBACK]$ ... $[/FEEDBACK]$
...
$[FEEDBACK]$ ... $[/FEEDBACK]$

PROMPT$^{S}_{k>1}$ (Solver)

Following is the feedback received on your previous solution. If there was a mistake in your previous solution, use this feedback and the **list of actionable steps** provided to reconsider and re-evaluate your solution. You should re-evaluate your understanding the given image (if any) and the problem statement using the feedback. Specifically, based on the provided feedback, please provide a detailed explanation of the given image (if any).

Please provide the final answer option and explanation of the intermediate steps, as well as details of how the feedback were addressed in your revised final answer. Your final answer should be reported in your final line and must select one of the answer options: A1, B2, C3, D4, E5.

$[PROBLEM]$ ... $[/PROBLEM]$

$[OPTIONS]$ ... $[/OPTIONS]$

$[RESPONSE]$ ... $[/RESPONSE]$

$[FEEDBACK]$ ... $[/FEEDBACK]$

E Additional Results

In Tables 5, 6, and 7, we provide more detailed comparisons between WISE-DS and alternative aggregation methods per category in the dataset for Visual Puzzles, SMART-840++, and SMART-840 dataset. As is clear, WISE-DS generally leads to the best performance. **Note:** The weighted-majority method depends on the weights produced by the feedback critics in WISE and thus is technically not a baseline. However, we include it in the study to clearly show the benefit of the WISE-DS probabilistic consensus model against only using the weighted majority from the WISE weights.

On the SMART-840++ dataset, we see a clear drop in performance of the models as the difficulty levels increase. Recall (from Figure 3) that the new dataset contains programmatically augmented puzzles of increasing difficulties. The results in Table 6 shows that model performance is in high 60s for difficulty level 0, however drops by about 10% for levels 3-4, suggesting that the LLMs are still not really capable of grounded mathematical visual reasoning.

Our analysis indicates that failures on SMART-840++ are not primarily due to EM non-convergence, but arise from three consistent factors: (i) visual perception challenges (e.g., Puzzle_19: counting errors), (ii) correlated hallucinations across agents (e.g., Puzzle_61), and (iii) breakdown of long-horizon reasoning and state tracking (e.g., Puzzle_111, Puzzle_140). These errors persist even with multiple rounds, suggesting limits in current model reasoning rather than aggregation.

Method	Algorithmic	Analogical	Deductive	Inductive	Spatial	Average
Majority-vote	0.5712	0.5836	0.5755	0.5806	0.5703	0.5807
Weighted-majority	0.5913	0.5911	0.5878	0.6015	0.5879	0.5952
DS	0.5260	0.5390	0.5286	0.5351	0.5240	0.5316
Weighted-DS	0.4740	0.4907	0.4694	0.4801	0.4856	0.5438
WISE-DS (ours)	**0.6080**	**0.6115**	**0.5939**	**0.6186**	**0.6134**	**0.6126**

Table 5: Category wise Performance on Visual Puzzles using $(G52 + CS46)^2|G52$ configuration.

Method Difficulty	0	1	2	3	4	Average
Majority-vote	0.6078	0.6071	0.5113	0.5772	0.5547	0.5716
Weighted-majority	0.6471	0.6143	0.5338	0.561	0.5703	0.5853
DS	0.5686	0.5929	0.5113	0.4878	0.5703	0.5462
Weighted-DS	0.5817	0.6214	**0.5414**	**0.5528**	**0.5781**	0.5751
WISE-DS	**0.6732**	**0.65**	0.5113	0.5366	**0.5781**	**0.5898**

Table 6: Category wise Performance on SMART-840++ using $(G52 + CS46)^2|G52$ configuration.

Method Grades	1_2	3_4	5_6	7_8	9_10	11_12	Average
Majority-vote	0.7560	0.7560	0.7637	0.7637	0.7637	0.7637	0.7615
Weighted-majority	0.7664	0.7664	0.7765	0.7765	0.7765	0.7765	0.7750
DS	0.7589	0.7589	0.7676	0.7676	0.7676	0.7676	0.7659
Weighted-DS	0.7649	0.7649	0.7701	0.7701	0.7701	0.7701	0.7650
RECONCILE	0.643	0.672	0.589	**0.855**	**0.818**	**0.806**	0.731
WISE-DS	**0.7783**	**0.7783**	**0.7829**	0.7887	0.7929	0.7829	**0.784**

Table 7: Category wise Performance on SMART-840 using $(G52 + CS46)^2|G52$ configuration.

E.1 Comparison to Feedback-based Methods?

In Table 8, we compare methods under matched compute and LLM-call budgets, focusing on approaches that use LLM/VLM feedback for solution refinement, including debate-based methods Chen et al. (2023). We evaluate: (i) single-model majority voting; (ii) Self-Reflect Shinn et al. (2023) with 4 rounds of self-reflection per model plus majority voting; (iii) self-consistency, sampling each model 4 times without feedback; (iv) homogeneous feedback, where models iteratively refine their own solutions; and (v) RECONCILE Chen et al. (2023), a heterogeneous multi-agent debate framework. Apart Debate-based methods consistently outperform single-model baselines. RECONCILE improves weaker models by up to $\sim$7% (e.g., Gemma3) and yields smaller gains for stronger models (e.g., 1.5% for GPT-4.1), whereas WISE delivers substantially larger gains under comparable compute. Results with stronger, newer models such as (G52+CS46) show similar trends, with about a 5% improvement over RECONCILE.

F WISE Debate with Strong versus Weak Models

In this section, we present additional ablation studies that evaluate several key components of our framework. Before diving into the experiments, we first review the LLM/MLLMs we use, their abbreviations that we refer to them as and establish a baseline performance metric that we use to categorize the MLLM and LLM models in our subsequent analyses.

In Table 9, we provide publicly available information of the LLMs that we consider, their capabilities, and tasks they were trained for. Table 10 summarizes each model's accuracy, parameter size, modality capabilities, and baseline performance on the SMART-840 dataset. These scores are obtained using a single forward pass—without any debate rounds—thereby serving as the foundational performance reference. We report results for both image–text problems and text-only problems, as our later studies incorporate text-only LLMs as critics that help weigh and critique the solutions proposed by multimodal LLMs.

Method	LLMs / VLMs	Image+Text	Text	Overall
Single	GPT-4.1 (G41)	48.2	88.2	60.6
	GPT-5.2 (G52)	65.0	95.9	72.0
	Gemma3 (G3)	24.0	73.7	38.6
	Claude-Sonnet-3.7 (CS37)	35.3	71.0	44.3
	Claude-Sonnet-4.6 (CS46)	55.2	92.5	65.8
Self-Reflect Shinn et al. (2023)	GPT-4.1	50.9	88.9	61.2
	Gemma3	30.0	58.2	46.0
Self-Consistency Wang et al. (2022)	GPT-4.1	48.6	90.3	63.2
	CS37	35.3	68.4	44.2
	Gemma3	24.8	56.8	43.9
Homogeneous Liang et al. (2023)	GPT-4.1	49.5	88.4	59.7
	CS37	34.9	71.0	45.2
	Gemma3	30.9	74.2	44.4
Single	G41+CS37+Ge3	45.5	82.7	55.6
Self-Reflect	Ge3+CS37+G41	42.6	83.4	53.9
Self-Consistency	G41+CS37+Ge3	47.2	84.1	56.6
Homogeneous	G41+CS37+Ge3	43.5	81.6	53.9
RECONCILE	G41 + Ge3 + S35	50.9	90.1	62.0
RECONCILE	G52+CS46	62.7	97.3	73.1
WISE (ours)	$(G41+Ge3+S35)^2$|G41	**55.8**	**91.4**	**68.1**
WISE (ours)	$(G52+CS46)^2$|G52	**70.6**	**98.2**	**78.4**

Table 8: Comparison of feedback methods on the SMART-840 dataset. The first three rows show single models with feedback rounds; the fourth shows majority voting over multiple single-feedback models. WISE results highlight the importance of WISE-way for agent interactions.

Based on these baseline results, we group the models into strong and weak categories. The two best-performing models, GPT-4.1 and Claude-Sonnet, are designated as strong. The remaining multimodal models—GPT-4o, Gemma3, and Qwen-VL-2.5—are categorized as weak. Similarly, the two text-only models, Phi-4 and Llama3, are also considered weak for the purposes of this study. We emphasize that this classification is purely performance-driven and is intended to enable a systematic exploration of different combinations of weak and strong models as they take on varied roles in our subsequent evaluations.

LLM/LVLM	**Skills**	**Training Tasks**	Multimodal?
GPT-4o (G4o)	Reasoning; coding	Language modeling, code generation	Yes
GPT-4.1 (G41)	Coding and reasoning	Language, advanced coding tasks, reasoning	Yes
Claude-Sonnet (CS35 or S35)	Strong reasoning, coding, multilingual	General reasoning, coding, text understanding	Yes
Claude-Haiku (Claude 3)	Lightweight, efficient reasoning	General reasoning, summarization, text analysis	Yes
Qwen-VL 2.5 (QVL)	Reasoning	Multimodal reasoning, VQA, captioning, dialogue	Yes
Gemma3 (Ge3	Reasoning (presumed)	General multimodal tasks, language understanding	Yes
Phi-4 (ϕ4)	Math and coding reasoning	Language modeling, mathematics, coding	No
Llama3 (L3)	General language understanding	Language modeling, reasoning, dialogue	No

Table 9: LLMs and MLLMs that we use within our performance analysis.

LLMs / MLLMs	Size	Capability	Performance	Image+Text	Text	Overall
GPT-4.1	Unknown	VL	Strong	48.2	88.2	60.6
Claude-Sonnet	Unknown	VL	Strong	35.3	71.0	44.3
GPT-4o	Unknown	VL	Weak	33.6	67.7	42.4
Gemma3	12B	VL	Weak	24.0	73.7	38.6
Qwen-VL-2.5	7B	VL	Weak	25.0	52.5	37.2
Phi-4	3.8B	L	Weak	N/A	65.4	N/A
Llama3	8B	L	Weak	N/A	34.8	N/A

Table 10: Performance of single models on the SMART-840 dataset (no debate or aggregation).

Solvers	critics	LLMs / MLLMs	1_2	3_4	5_6	7_8	9_10	11_12	Overall
Strong	Strong	$(G41+S35)^2$|G4o	66.7	59.1	58.2	78.9	71.4	72.2	67.7
Strong	Weak	(G41+S35)×(G4o+CH3*+ϕ4)|G4o	59.2	58.3	49.3	73.3	60.7	69.3	61.7
Mix	Mix	$(G41+Ge3+S35)^2$|G4o	**65.0**	**61.8**	**56.6**	74.3	**75.5**	**73.0**	**68.1**
Weak	Weak	(G4o+QVL)×(Q+ϕ4+L3+G3)|G4o	51.4	39.4	41.0	53.7	53.9	58.7	49.7
Weak	L-only	G4o × (ϕ4+o1-mini)|G4o	49.2	45.4	47.1	54.6	56.9	55.8	51.5

Table 11: Performance of WISE on SMART-840 for varied weak and strong LLM configurations under approximately the **same number of LLM calls.** *CH3 stands for Claude-Haiku 3.

F.1 Performance on Varied LLM Configurations:

Table 11 reports results for strong, weak, and language-only model combinations in the WISE debate framework. We test solver–critic pairings in five categories: strong–strong, strong–mix, mix–mix, weak–weak, and weak–language-only (mix includes both strong and weak models). All runs use the same distiller (GPT-4o) and are evaluated on SMART-840 with WISE-DS. Strong–strong performs well, while mix–mix—our best setup—slightly improves the score (67.7 $\rightarrow$ 68.1), suggesting that controlled role diversity helps. The weak–weak combination performs below all other pairings, as anticipated, yet it still noticeably surpasses the best individual model in the pair. For example, GPT-4o alone achieves 42.4% accuracy (Table 10), but when used in a weak–weak WISE configuration, performance rises to 49.7%—comparable to RECONCILE when using similarly weak models. Unlike RECONCILE, which assumes all participating models have similar capabilities (e.g., all text-only or all vision–language), WISE explicitly assigns models to specialized roles, enabling more flexible and heterogeneous debate setups. This flexibility becomes especially advantageous in the weak–language-only scenario, where WISE reaches 51.5% accuracy. Under similar compute budgets, this is a 1.1% improvement over RECONCILE using GPT-4o—despite operating entirely in a weak setting.

Overall, our results show that strong–strong configurations predictably benefit WISE. However, heterogeneous pairings yield mixed trends: i) strong–mix combinations where the mix role is dominated by weak critics tend to underperform, likely due to lower-quality feedback; ii) weak–weak and weak–L-only setups can still deliver competitive gains, suggesting that WISE 's structured debate can absorb or correct suboptimal weighting from weaker critics. These findings highlight the robustness and adaptability of WISE in leveraging diverse model capabilities.

F.2 Ablation on the Feedback Weights?

In Table 12, we report a sensitivity analysis on the effect of weight space. We use a subset of 150 problems from SMART-840, grades 5-6. We use the latest multimodal LLMs: GPT-5.2 and Claude-Sonnet-4.6. As evident, the average variation in performance is minimal, but the variance increases.

Weight space	WISE Acc. (%)
$\{0, 1, 2\}$	65.3 ± 1.9
$\{0, 1, \ldots, 9\}$	65.3 ± 2.1
$\{0, 1, \ldots, 99\}$	66.6 ± 2.7

Table 12: Ablation of the weights range used in the critic feedback. In this study, we change the range of the weights. The set $\{0, 1, 2\}$ is the default range we use in our experiments.

F.3 Performance Under Token Limit:

In Figure 5, we examine the robustness of WISE when imposing constraints on the maximum number of tokens available to the underlying models. For each MLLM used within WISE, we apply a token cap that we vary from 512 to 8192 tokens. We observe that restricting the token budget noticeably alters the behavior of the individual models, often resulting in substantial fluctuations in their standalone performance. In contrast, WISE shows only minimal performance variance across the same range of token limits. Its multi-model structure effectively balances errors made by individual models, yielding a remarkably stable overall accuracy. For this experiment, we employ weighted majority voting within WISE to isolate the effect of token limitations and remove any influence from the WISE-Dawid–Skene component.

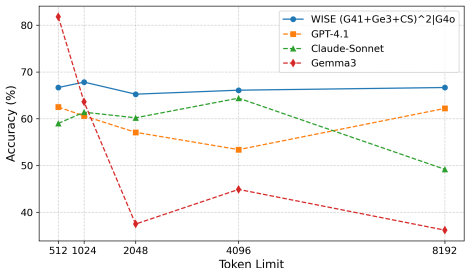


Figure 5: Token limit vs. performance on the SMART-840 1-2 split using WISE $(G41 + Ge3 + CS37)^2|G4o$. Single MLLMs (no debate) vary widely, but WISE 's debate-based combination is consistently robust and superior across token limits. Aggregation uses weighted majority.

G SMART-840++ Dataset, Statistics, and Generation Pipeline

G.1 Motivation:

As alluded to in the main paper, the key motivation for proposing the SMART-840++ dataset is to understand and analyze the capability of state-of-the-art multimodal large language models over three axes: i) generalization to never seen before problem content, ii) generalization to complex problem settings with controlled increase in the problem difficulty, and iii) robustness in performance to minor variations to the problem objectives. We note that our analysis uses publicly available problems from the SMART-840 dataset, presented in Cherian et al. (2024). The SMART-840 dataset contains 840 problems extracted from the Math Kangaroo (MK) Olympiads[3] from years 2020-2025. The problems span four areas: geometry, algebra, numerics, and logic, and is segregated into 6 categories, one for a pair of consecutive school grades from 1–12. We note that similar attempts at problem augmentation has been attempted to before, e.g., GSM-Symbolic Mirzadeh et al. (2024), Zebra-Logic Lin et al. (2025), and SMART-101 Cherian et al. (2023). While the first two prior works are for text-only domain, SMART-101 is for vision-language problems however is limited to only simple problems from 1–2 grades of the MK Olympiads. Through SMART-840++, we go beyond SMART-101 and propose new problems across grade curriculum.

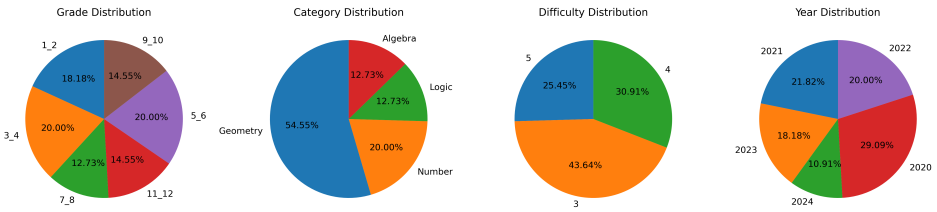


Figure 6: Distribution of problems in the SMART-840++ dataset. We also show the distribution of puzzles across grades, categories, difficulty levels, and years of MK competition. Note that the difficulties of the puzzles shown in the third plot above are with regards to the SMART-840 difficulty categories (i.e., 3: simple, 4: medium, 5: hard), which reflects the difficulty in algorithmic derivation. In SMART-840++, each puzzle is augmented into 5 new difficulty levels irrespective of them being in the original three difficulty categories, where the 'difficulty' in SMART-840++ predominantly reflects the scale of the problem irrespective of algorithmic difficulty.

G.2 Problem Selection:

To construct the SMART-840++ dataset, we selected 55 vision-and-language problem types, approximately balanced across grade levels. Because manually authoring large numbers of problems is labor-intensive and difficult to scale, we prioritized problems that could be programmatically regenerated using Python-based graphics toolboxes. Our approach is inspired by Cherian et al. (2023), which demonstrated programmatic reconstruction for MK problems, though limited to grades 1–2.

Programmatic synthesis offers several key advantages: i) it enables the generation of arbitrarily many problem instances; ii) by adjusting program parameters, we can control difficulty, such as using a 3×3 grid for easier tasks or a 10×10 grid for harder ones—both governed by the same underlying algorithm, but requiring different numbers of tokens and longer reasoning chains; and iii) it allows rapid creation of visual variety, including changes in colors, viewpoints, textures, and layouts.

Our overarching goal is to create novel problems that LLMs are unlikely to have encountered in training, thereby requiring genuine algorithmic reasoning that is invariant or agnostic to superficial variations in appearance or difficulty. Sample problem instances from SMART-840++ are shown in Figure 7. A full set containing one example instance from each of the 55 problem types appears at the end of this supplementary material (Figure 8). Dataset statistics are summarized in Figure 6. Each problem includes an image, a text question, and five answer options (image-based or text-based), with exactly one correct answer.

[3] mathkangaroo.org

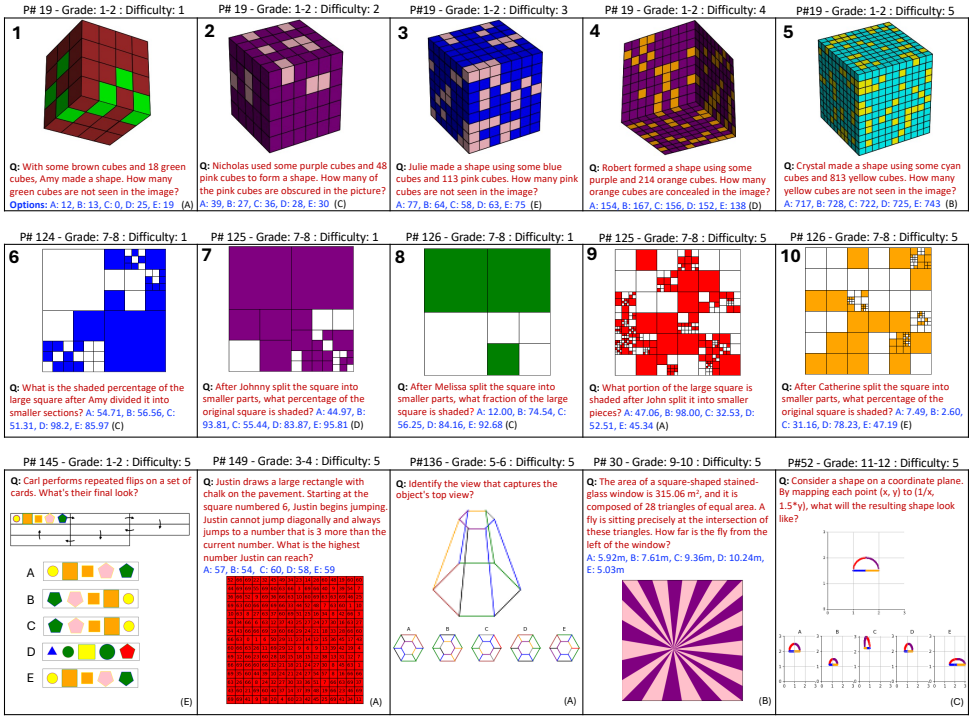


Figure 7: Example problems from the SMART-840++ dataset. We show variations in the newly generated puzzles across appearance and problem complexity, as well as diversity in problem text and options. For the below figure, we denote $D-$ to denote the difficulty levels for instances of the same problem. See Figure 8 for the full list of problems later in this supplementary materials.

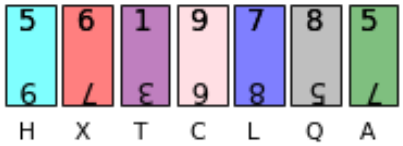
G.3 Algorithmic Augmentation Pipeline

As described above, we use Python to implement our augmentation pipeline for selected problems from the SMART-840 dataset. For each original problem, we first solved it manually, then wrote a Python program that (i) semantically reconstructs the problem and (ii) implements the corresponding solution logic. Each program is parameterized so that modifying its arguments generates new problem instances with arbitrary configurations, difficulty levels, and visual appearances. When visual attributes are not essential to the reasoning (e.g., colors, sizes, or minor layout choices), we randomly sample them from predefined sets.

For the text component, we begin by constructing a question template derived from the original problem, substituting puzzle-specific values where appropriate. We then use an LLM (e.g., Microsoft Copilot) to rephrase the text—using prompt constraints that preserve the underlying logical structure while allowing variations in style, tone, or character names. The resulting problem instances thus differ in appearance and wording from the originals, yet retain the same (or nearly the same) underlying algorithmic reasoning structure.

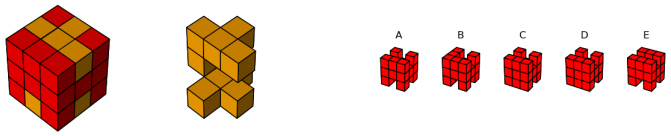
H SMART-840++ DATASET EXAMPLE INSTANCES

SMART-840++ Problem #1 – mk-2021-3-4-18: Pamela arranged 7 cards, each with two numbers, one written upside down. To balance the sums of the top and bottom rows, which card should they flip?



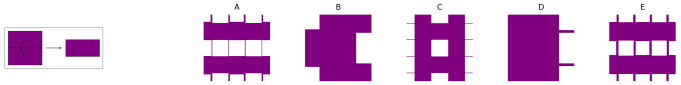
A: A, B: H, C: Q, D: X, E: T
Answer: E

SMART-840++ Problem #2 – mk-2021-7-8-8: The initial diagram displays a box constructed from orange, red cubes. The subsequent diagrams illustrate the orange sections of the cube. Which diagram shows the red part?



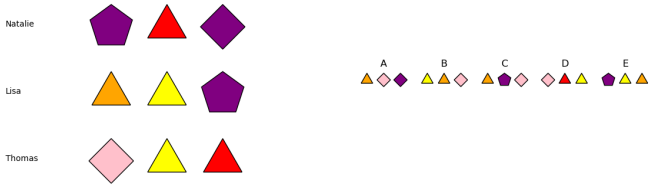
A: A, B: B, C: C, D: D, E: E
Answer: A

SMART-840++ Problem #3 – mk-2023-3-4-18: How does the paper look after Melissa folds it several times, cuts off some corners, and then unfolds it?



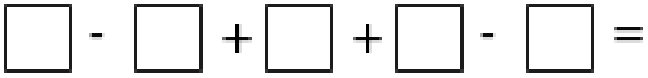
A: A, B: B, C: C, D: D, E: E
Answer: B

SMART-840++ Problem #4 – mk-2024-1-2-20: Thomas, Lisa, Natalie, Charles each possess 3 objects, each sharing exactly one object with every other child. Which object does Charles have?



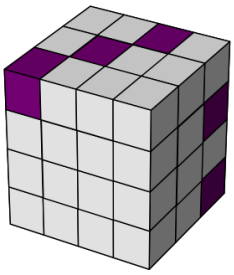
A: A, B: B, C: C, D: D, E: E
Answer: A

SMART-840++ Problem #5 – mk-2024-11-12-21: In a stack, Amber has several numbers written on each card. The numbers on each card are as follows: $[[17, 25], [10, 21], [48, 5], [36, 42], [8, 37]]$. The cards can be arranged in any order in the blank spaces of the figure, but only one number can be chosen per card. What is the smallest number that can be made?



A: -52, B: -54, C: -55, D: -53, E: -51
Answer: C

SMART-840++ Problem #6 – mk-2020-1-2-6: Katherine constructed a shape with some White cubes and 12 Purple cubes. How many of the Purple cubes cannot be seen in the image?



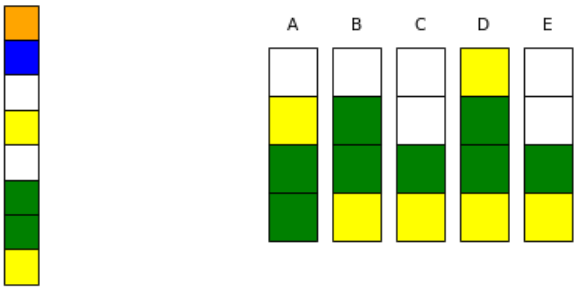
A: 7, B: 2, C: 18, D: 0, E: 19
Answer: A

SMART-840++ Problem #7 – mk-2021-5-6-17: Charles and Bridget, limited to their colored tokens, alternated turns to place tokens in two piles, starting with Charles. Which pair of piles could they not form?



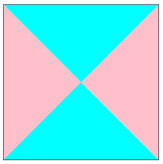
A: A, B: B, C: C, D: D, E: E
Answer: D

SMART-840++ Problem #8 – mk-2024-3-4-11: Tina created a block tower, as seen in the image. Then they remove the 1st block from the top. Then they remove the 3rd block from the top. Then they remove the 1st block from the top. Then they remove the 2nd block from the bottom. Which tower does Tina end up making?



A: A, B: B, C: C, D: D, E: E
Answer: C

SMART-840++ Problem #9 – mk-2020-9-10-20: Imagine a $square - shaped$ stained glass window with an area of 60.22 dm^2, divided into 4 equal triangles. A fly is sitting at the exact point where all triangles converge. How far is the fly from the bottom of the window?



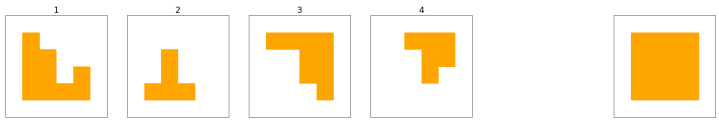
A: 6.26dm, B: 5.95dm, C: 3.88dm, D: 7.56dm, E: 7.86dm
Answer: C

SMART-840++ Problem #10 – mk-2020-3-4-4: Which figure, drawn by Thomas, contains the largest shaded section?



A: A, B: B, C: C, D: D, E: E
Answer: B

SMART-840++ Problem #11 – mk-2023-3-4-4: Which two pieces can Mrs. put together to match the shape on the right?



A: 3 and 4, B: 4 and 2, C: 4 and 1, D: 2 and 3, E: 2 and 1
Answer: C

Figure 8: SMART-840++ Problem #11 – mk-2023-3-4-4.