

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 DOReMi - DIFFICULTY-ORIENTED REASONING EF- FORT MODELING OF SCIENCE PROBLEMS FOR LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

## ABSTRACT

We introduce DoReMi (Difficulty-Oriented Reasoning Effort Modeling), a structured framework leveraging an extended Bloom’s taxonomy to comprehensively characterize intrinsic problem difficulty for large language models on scientific reasoning tasks. DoReMi systematically annotates problems along six cognitive and methodological axes using judge large language models (LLM) distinct from those being evaluated, with human annotations confirming the validity of these assessments. We empirically quantify LLM reasoning effort through metrics including minimum reasoning tokens required for solution, expected number of attempted runs to first correct answer. Our validation demonstrates strong agreement across diverse judge LLMs spanning both open-source and proprietary LLMs. Evaluations on GPQA, ARC, and SuperGPQA reveal that our multidimensional difficulty fingerprints correlate strongly with and enable accurate predictive modeling of LLM reasoning effort. DoReMi enables principled difficulty-aware subset selection that substantially outperforms other baselines while providing interpretable diagnostics that uncover emergent reasoning capabilities across successive model generations. This framework offers actionable insights for benchmark design and targeted post-training improvements toward higher-order reasoning skills.

## 1 INTRODUCTION

Latest reasoning large language models (LLMs) have demonstrated significant progress in tackling complex reasoning tasks. However, clearly characterizing their capabilities remains challenging, as task difficulty often combines several partially overlapping factors including linguistic complexity, domain-specific knowledge, and the depth of reasoning involved. Consider a partial differential equation problem: an LLM might present a correct solution either by recalling a known theorem or by logically deriving the solution from fundamental principles. Both approaches are expressed through language and rely on familiarity with domain-specific notation, such as distinguishing between  $x$  and  $\vec{x}$ . When conventional benchmarks fold these heterogeneous challenges into a single accuracy score, they hide which capability—domain knowledge, deductive reasoning, or methodological complexity—were the real bottlenecks. Even seemingly finer-grained signals, like the accuracies of a question across a leaderboard packed with hundreds of LLMs, tell us little: most leaderboard entries are generic non-reasoning models, so their collective failure is like asking a roomful of laypeople to solve a PhD-level physics question—the near-universal miss reflects the respondents more than the task. Without a nuanced and principled way to measure difficulty for reasoning LLMs, tracking progress across model versions becomes inconsistent. New benchmarks risk being quickly saturated by state-of-the-art LLMs. Applications like curriculum learning or difficulty-aware subset sampling, which depend on understanding why a question is hard, remain mostly heuristic.

To address these limitations, we propose a *structured, multi-dimensional* evaluation framework grounded in educational theory, specifically Bloom’s taxonomy and its extensions Heer (2012). Our framework systematically annotates each scientific reasoning problem along six complementary axes: *Cognitive Level, Knowledge Dimension, Method Difficulty, Definition Completeness, Knowledge Breadth, and Number of Reasoning Steps*. Although not strictly orthogonal, these dimensions provide principled, theory-informed handles that expose facets of difficulty invisible to traditional static metrics such as SMOG, Gunning Fog, and Flesch-Kincaid scores McLaughlin (1969); Scott (2025);

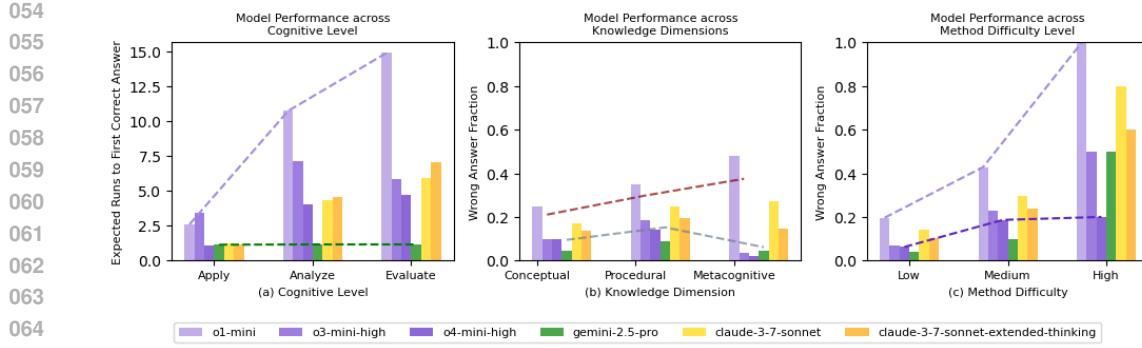


Figure 1: The performance of different LLMs long three Bloom axes: (a) cognitive level, (b) knowledge dimension, and (c) method difficulty.

Tanprasert & Kauchak (2021). The resulting interpretable “difficulty fingerprint” explicitly captures both the *what* (knowledge and cognition) and the *how* (methodological effort) necessary to reach a solution. This approach supports more nuanced performance comparisons across LLMs and helps evaluate and improve the difficulty of both current and future benchmarks.

One key application of this framework is to reveal unique insights into how intrinsic problem characteristics influence the reasoning effort required by LLMs for scientific problem-solving. Figure 1 highlights three important trends. **(i) Cognitive Level:** Early-generation reasoning models like `o1-mini` exhibit a steep increase in reasoning effort (measured by the expected number of runs to hit the first correct answer) as cognitive complexity increases, whereas advanced current-generation models such as `gemini-2.5-pro` demonstrate remarkable stability, maintaining consistently low reasoning effort across cognitive levels. **(ii) Knowledge Dimension:** The first-generation reasoning model `o1-mini` and the non-reasoning mode of `claude-3.7-sonnet` show a clear rise in error rates as problems require higher-order knowledge dimensions. In contrast, reasoning enabled (`claude-3.7-sonnet-extended-thinking`) and other later-generation reasoning LLMs exhibit robust, stable performance or even improved accuracy at the highest (metacognitive) dimension, suggesting emergent capabilities along the knowledge dimension axis. **(iii) Method Difficulty:** we observe improved robustness across successive generations **in the same model family** as `o1-mini` shows a sharp rise in errors from *Low* to *High* methodological difficulty, while `o3-mini-high` exhibits a more gradual increase and `o4-mini-high` maintains stable performance from *Medium* to *High*. These diagnostic insights reveal a gradual and dimension-specific emergence of higher-order reasoning skills. This pattern supports our hypothesis that scaled reinforcement learning (RL) during post-training may lead to effective transfer of reasoning strategies. It also enables stronger cognitive generalization across the Bloom-inspired axes. Our paper contributes three primary advancements:

- 1. Difficulty-Oriented Reasoning Effort Modeling (DoReMi).** We introduce a framework to model difficulty for reasoning-intensive science problems. By correlating structured difficulty metrics derived from an extended version of Bloom’s taxonomy with empirically measurable proxies of *reasoning effort*, such as the *minimum reasoning token length (MRT)* and *expected number of attempted runs to first correct answer (R2FCA)*, we establish interpretable, dimension-specific relationships between intrinsic problem characteristics and LLM reasoning demands. Based on these insights, we develop and validate predictive models capable of accurately identifying challenging (“high-effort”) problems across multiple established scientific reasoning benchmarks such as GPQA Rein et al. (2024), ARC Clark et al. (2018), and SuperGPQA Team et al. (2025).
- 2. Difficulty-Aware Downstream Applications.** Using the predictive capabilities of our DoReMi framework, we propose difficulty-aware subset selection, which adaptively prioritizes challenging problems. This helps preserve discriminative evaluation power even as aggregate benchmark performance saturates due to rapid advances in LLM capabilities. Experiments comparing our approach against static-difficulty baselines show significant improvements in correctly identifying truly challenging problems. Furthermore, these difficulty-aware selection strategies could enable more informed curriculum learning protocols and facilitate targeted benchmark refinement.
- 3. Interpretable Diagnostics of Reasoning Capabilities.** We leverage the fine-grained difficulty characterizations provided by DoReMi to construct a diagnostic framework for systematically analyzing reasoning strengths and weaknesses of LLMs. By examining performance stratified along multiple axes of Bloom’s taxonomy, we uncover distinct patterns in reasoning behaviors

108 across generations of models and post-training stages. Our analyses pinpoint precisely how and  
 109 where improvements manifest—highlighting emergent capabilities especially along cognitive  
 110 and knowledge dimension axes. These gained insights could provide potential guidance on  
 111 post-training methodologies to foster higher-order reasoning skills.  
 112

## 113 2 DESIGN PRINCIPLES

115 There is currently a lack of precise and reliable measures of intrinsic problem difficulty tailored  
 116 specifically to reasoning capabilities. Traditional static metrics, such as readability scores, overlook  
 117 cognitive and methodological complexity inherent to reasoning problems.  
 118

### 119 2.1 ANALOGY AND INTUITION: REASONING EFFORT AS A COGNITIVE BUDGET

121 We propose to operationalize problem difficulty through explicit, measurable proxies for *reasoning  
 122 effort*, directly mirroring cognitive processes observed in human scientific problem-solving:  
 123

**124 Expected Number of Attempted Runs to First Correct Answer (R2FCA)** Analogous to a mathe-  
 125 matician discarding one proof sketch after another until a promising idea emerges, we measure the  
 126 average number of independent attempts an LLM needs before it first produces the correct answer.  
 127 This metric reflects both the model’s exploratory persistence and its inherent stochasticity.  
 128

**129 Minimum Reasoning Token (MRT)** required to solve a problem from multiple sampled solutions,  
 130 retaining only the shortest successful one. This is the LLM analogue of a mathematician’s cleanest  
 131 proof, capturing the minimal cognitive and computational budget required to solve the problem.  
 132

133 Consider the cognitive process a mathematician undergoes when confronting a challenging theorem.  
 134 Typically, they begin by exploring various potential proof strategies—each attempt consuming  
 135 significant time, effort, and cognitive resources analogous to iterative “scratch paper” explorations.  
 136 As many explored pathways fail to yield immediate success, repeated exploratory iterations often  
 137 become necessary. Thus, if the probability of reaching a correct solution per attempt is low, the  
 138 expected number of exploratory attempts scales inversely with this probability. Additionally, even  
 139 after identifying a viable strategy, the final formal proof still incurs an inherent cognitive “cost”—the  
 140 minimal sequence of logical reasoning steps (or written tokens) required to rigorously articulate the  
 141 solution. In this analogy, our notion of *Expected Reasoning Cost (ERC)* aligns with:  
 142

$$143 \text{Expected Reasoning Cost (ERC)} \sim \underbrace{E[\text{R2FCA}]}_{\text{Exploratory Attempts}} \otimes \underbrace{\text{MRT}}_{\text{Minimal Solution Transcript}}$$

144 as a combined metric indicated by symbolic  $\otimes$  that jointly captures the iterative exploratory complex-  
 145 ity and the minimal cognitive burden to reach a successful solution. Applying this to LLM reasoning  
 146 under stochastic sampling, the repeated attempts are like the mathematician’s exploration, and the  
 147 generated reasoning tokens represent the cognitive resources used.  
 148

### 149 2.2 EXTENDED BLOOM’S TAXONOMY

150 To capture the multifaceted nature of science problem, we extend the original Bloom’s framework  
 151 into a *six-axis taxonomy*. Each axis is annotated directly from the problem statement and reference  
 152 solution, yielding a machine-parsable “difficulty fingerprint”.

- 153 **1. Cognitive Level:** Highest Bloom cognitive process required to solve the task: *Remember, Under-*  
 154 *stand, Apply, Analyze, Evaluate, or Create*. The level is chosen by locating the most demanding  
 155 mental operation that a correct solution must exhibit.
- 156 **2. Knowledge Dimension:** Type of knowledge invoked: *Factual, Conceptual, Procedural, or*  
 157 *Metacognitive*. This axis distinguishes mere recall from methodological know-how and self-  
 158 regulation of the reasoning process.
- 159 **3. Method Difficulty:** Degree of methodological novelty (*Low, Medium, High*). *Low* denotes routine,  
 160 textbook procedures; *Medium* requires minor adaptations or synthesis; *High* entails non-routine  
 161 combinations or inventive leaps.
- 162 **4. Definition Completeness:** Whether the statement fully specifies the solution space (*Complete*) or  
 163 leaves essential variables/criteria implicit (*Incomplete*), forcing the solver to supply assumptions.

162 5. **Knowledge Breadth** Disciplinary span of required knowledge: *Single-* versus *Multi-Discipline*.  
 163 The latter flags problems that integrate concepts from two or more distinct scientific fields.  
 164 6. **Number of Reasoning Steps:** Integer count of essential logical actions whose removal would  
 165 break the solution chain. Trivial paraphrases are excluded.

166 Together, these six axes disentangle the *what* (knowledge and cognition) from the *how* (method and  
 167 reasoning) of problem solving, enabling the reasoning effort modeling based on these metrics.  
 168

### 170 3 RELATED WORK

172 Recent advancements in reasoning LLMs have highlighted the need for precise measures of problem  
 173 difficulty, specifically within scientific tasks where reasoning complexity is prominent. Traditional  
 174 assessments have relied on static readability metrics, such as SMOG, Gunning Fog, or Flesch-  
 175 Kincaid McLaughlin (1969); Scott (2025); Tanprasert & Kauchak (2021), which fail to reflect deeper  
 176 cognitive demands required in complex reasoning tasks. Prompt-based approaches Rooein et al.  
 177 (2024) partially overcome these limitations by leveraging LLMs’ language understanding capabilities  
 178 to capture more abstract complexity; however, they primarily emphasize textual difficulty rather  
 179 than cognitive or methodological complexity. Compared to these static readability-focused metrics,  
 180 our DoReMi framework leverages an extended Bloom’s taxonomy Heer (2012), systematically  
 181 characterizing multiple dimensions of intrinsic cognitive and methodological difficulty.

182 Previous studies have employed Bloom’s taxonomy in the context of LLM evaluation Huber &  
 183 Niklaus (2025) and curriculum learning design Hase et al. (2024). Huber et al. Huber & Niklaus  
 184 (2025) classified benchmarks according to Bloom’s cognitive levels, revealing that LLM performance  
 185 predominantly excels at lower cognitive levels. Complementing this work, DoReMi enriches Bloom’s  
 186 taxonomy with additional methodological and metacognitive axes, directly linking these theory-  
 187 informed dimensions with empirical proxies of reasoning effort rather than accuracy alone.

188 Curriculum learning shows performance gains when aligning difficulty progression for LLM training.  
 189 To quantify sample-level hardness, previous work Hase et al. (2024) annotated the Bloom’s taxonomy  
 190 for each dataset based solely on human annotation. In contrast, we refined annotation with multiple  
 191 LLM judges and validated them with human annotations to ensure scalability and consistency.  
 192

193 Recent research investigates the relation between chain-of-thought (CoT) lengths and reasoning  
 194 success, suggesting a non-monotonic relationship and an optimal length dependent on problem  
 195 complexity and model capacity Wu et al. (2025). In comparison to this previous analysis, DoReMi ex-  
 196 plicitly correlates intrinsic problem properties to proxies of reasoning effort, modeling and predicting  
 197 effective reasoning difficulties across multiple science benchmarks.

### 199 4 SOLUTION

201 In this section, we describe our systematic methodology to predict LLM reasoning effort by combining  
 202 difficulty metrics from an extended 3D Bloom’s Taxonomy with learned reasoning patterns.  
 203

#### 205 4.1 PROBLEM DIFFICULTY QUANTIFICATION VIA 3D BLOOM’S TAXONOMY

207 We develop a systematic annotation process using our extended 3D Bloom’s taxonomy (Section 2.2)  
 208 to quantify problem difficulty across  $D = 6$  dimensions: *Cognitive Level, Knowledge Dimension,*  
 209 *Method Difficulty, Definition Completeness, Reasoning Steps, Knowledge Breath*.

210 For each problem, we employ  $K$  reasoning LLMs as automated judges. Each judge independently  
 211 evaluates the problem statement and reference solution, classifying difficulty along each Bloom  
 212 dimension with supporting rationale. Through iterative prompt refinement—testing over two dozen  
 213 variants with explicit decision criteria and illustrative examples—we achieved strong inter-judge  
 214 agreement. The optimized prompts are provided in Section 7.2. We also validated our approach by  
 215 comparing average LLM-as-a-judge scores against two human annotations on 100 GPQA questions,  
 with good alignment results observed in Section 5.1.

Final annotations are obtained by encoding categorical Bloom levels into ordinal scores  $b_{i,d} \in \{1, 2, \dots, L_d\}$  for dimension  $d$  and problem  $i$ , then averaging across judges. This yields a continuous multi-dimensional "difficulty fingerprint"  $B_i = \{b_{i,d}\}_{d=1}^D$  for each problem.

## 4.2 REASONING EFFORT METRICS

To capture the computational and algorithmic effort required for reasoning, we evaluate  $M$  reasoning models, each for  $R_m$  runs per question. We define four complementary effort metrics:

**(1) Wrong Answer Fraction (WAF):** The average failure rate across models:

$$\text{WAF}_i = \frac{1}{M} \sum_{m=1}^M \left( \frac{1}{R_m} \sum_{r=1}^{R_m} \mathbf{1}(s_{i,m,r} = 0) \right)$$

where  $s_{i,m,r} \in \{0, 1\}$  indicates correctness for question  $i$ , model  $m$ , run  $r$ .

**(2) Minimum Reasoning Tokens (MRT):** The minimum token count needed for success:

$$\text{MRT}_{i,m} = \begin{cases} \min\{t_{i,m,r} \mid s_{i,m,r} = 1\}, & \text{if any run succeeds} \\ \max\{t_{i,m,r} \mid r \in [R_m]\}, & \text{otherwise} \end{cases}$$

where  $t_{i,m,r}$  denotes reasoning tokens used.

**(3) Expected Runs to First Correct Answer (R2FCA):** The expected number of attempts needed:

$$\text{R2FCA}_i = \sum_{n=1}^{R_m} n \cdot P(n) + \epsilon$$

where  $P(n)$  is the empirical probability of first success on run  $n$ . Taking the expectation across many trials smooths out single-run volatility; the  $\epsilon$  floor prevents division by 0 when no run succeeds. R2FCA is intended to measure *problem solvability under repeated attempts*.

**(4) Reasoning Inconsistency (RI):** The diversity of reasoning trajectories:

$$\text{RI}_i = \frac{1}{R_m} \sum_{r=1}^{R_m} d(c, e_r), \quad c = \frac{1}{R_m} \sum_{r=1}^{R_m} e_r$$

where  $e_r$  are response embeddings,  $c$  is their centroid, and  $d(\cdot, \cdot)$  is cosine distance.

## 4.3 LEARNING TO PREDICT REASONING EFFORT

Our correlation analysis in Section 5.2 reveals that MRT exhibits the strongest relationship with Bloom metrics. Based on this finding, we develop a two-stage approach to predict reasoning effort:

**Stage 1: Model Aggregation.** We combine model-specific MRTs into a unified metric:

$$\text{MRT}_C = \sum_{m=1}^M w_m \cdot \text{MRT}_{i,m}$$

where weights  $w_m$  are learned via gradient descent to maximize F1-score on high-effort samples.  $\text{MRT}_C$  is distributed in four quantile-based bins — Minimum, Low, Medium, and High—ensuring clear separation of difficult problems from the abundant easy ones. These binned labels serve as targets/ground truth for training our reasoning-effort predictor.

$$\text{Category}(x) = \begin{cases} \text{Minimum}, & x < \mu - \sigma \\ \text{Low}, & \mu - \sigma \leq x < \mu \\ \text{Medium}, & \mu \leq x < \mu + \sigma \\ \text{High}, & x \geq \mu + \sigma \end{cases}$$

**Stage 2: Difficulty-to-Effort Mapping.** We train a neural classifier  $f_\theta : B_i \rightarrow \text{Category}$  that maps Bloom features to effort categories. The model is optimized using weighted cross-entropy loss with

270 emphasis on high-effort samples:  
 271

$$272 \quad \mathcal{L} = - \sum_{i=1}^N \sum_{c=1}^4 \alpha_c \cdot y_{i,c} \log(f_\theta(B_i)_c)$$

$$273$$

$$274$$

275 where  $\alpha_c$  are class weights and  $y_{i,c}$  are one-hot encoded labels.  
 276

#### 277 4.4 DOREMI ALGORITHM

278 Algorithm 1 summarizes our complete pipeline: (1) annotate bloom metrics using LLM judges,  
 279 (2) inference of reasoning LLMs to collect reasoning effort metrics, (3) learn optimal aggregation  
 280 weights, and (4) train a predictor from difficulty features to effort categories.  
 281

---

#### 282 **Algorithm 1** DoReMi: Difficulty-oriented Reasoning Effort Modeling

---

283  
 284 **Require:** Questions  $Q = \{q_i\}_{i=1}^N$ , Models  $M = \{m_j\}_{j=1}^M$ , Judges  $J = \{j_k\}_{k=1}^K$   
 285 **Ensure:** Effort predictor  $f_\theta : B \rightarrow \text{Category}$   
 286 1: // Phase 1: Difficulty Annotation  
 287 2: **for** each question  $q_i \in Q$  **do**  
 288 3:   Obtain Bloom features  $B_i$  via judge consensus (Section 3.1)  
 289 4: **end for**  
 290 5: // Phase 2: Effort Measurement  
 291 6: **for** each model  $m_j \in M$ , question  $q_i \in Q$  **do**  
 292 7:   Evaluate  $R_j$  runs, record correctness and token counts  
 293 8:   Compute  $\text{MRT}_{i,j}$  using Equation (2)  
 294 9: **end for**  
 295 10: // Phase 3: Learn Aggregation  
 296 11: Initialize weights  $w = [w_1, \dots, w_M]$  randomly  
 297 12: **while** not converged **do**  
 298 13:    $\text{MRT}_C \leftarrow \sum_j w_j \cdot \text{MRT}_{i,j}$   
 299 14:   Update  $w$  to maximize F1 on high-effort class  
 300 15: **end while**  
 301 16: // Phase 4: Train Predictor  
 302 17: Discretize  $\{\text{MRT}_{C,i}\}$  into categories  
 303 18: Train  $f_\theta$  on dataset  $\{(B_i, \text{Category}_i)\}_{i=1}^N$   
 304 19: **return** Trained predictor  $f_\theta$   


---

## 305 5 EXPERIMENTS AND RESULTS

306  
 307 Comprehensive experiments were conducted to validate our DoReMi framework. We evaluated 9 LLMs: three generations of OpenAI reasoning models (o1-mini, o3-mini-high,  
 308 o4-mini-high); three generations of Qwen-14B model family including two non-reasoning  
 309 variants (Qwen1.5-14B, Qwen2.5-14B) plus one reasoning variant (Qwen3-14B); Google’s  
 310 gemini-2.5-pro; and Anthropic’s claude-3.7-sonnet (non-reasoning mode) and  
 311 claude-3.7-sonnet-extended-thinking (reasoning mode). Each model answers every  
 312 problem 10 times to capture stochasticity, yielding 90 runs per problem.  
 313

314 Our evaluation spans three scientific reasoning benchmarks representing varying difficulty distributions:  
 315 GPQA (PhD-level STEM), ARC (K-12 science reasoning), and two specialized SuperGPQA  
 316 domains (particle physics and molecular biology) test deep domain expertise.  
 317

### 318 5.1 BLOOM SCORES: ALIGNMENT BETWEEN JUDGE GROUPS

319  
 320 To validate our automated Bloom annotations, we compared LLM-generated scores with hu-  
 321 man judgments on 100 GPQA problems. Two independent human annotators rated all six  
 322 Bloom dimensions; Human1 served as the reference. We used three open-source reason-  
 323 ing LLMs—DeepSeek-V3.1-Terminus DeepSeek-AI (2024), GPT-OSS-120B OpenAI  
 324 (2025), Qwen3-Next-80B-A3B-Thinking Yang et al. (2025)—to avoid overlap with our

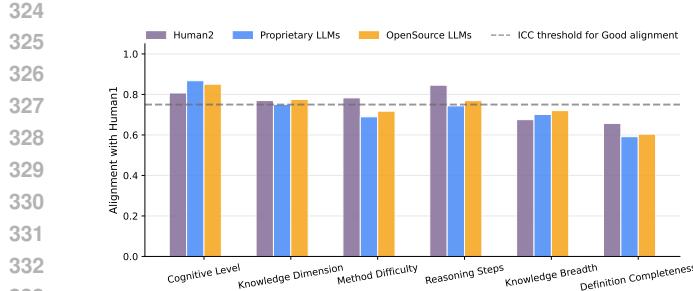


Figure 2: Human–model alignment across Bloom metrics. ICC(2,1) values measure inter-rater reliability between Human1 (reference) and three annotator groups.



Figure 3: Correlation Analysis of Reasoning Effort proxies across Bloom Axes

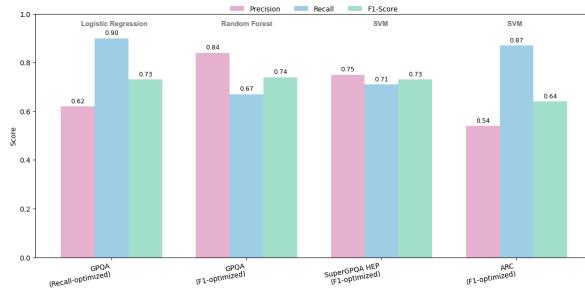


Figure 4: Test performance of optimal classification model for high-effort prediction on GPQA, SuperGPQA-HEP, and ARC using recall and f1-score optimization

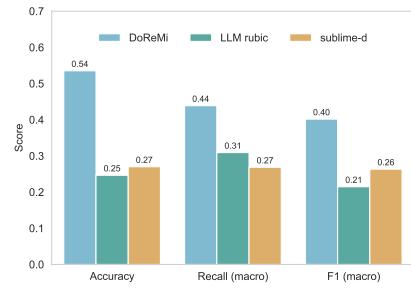


Figure 5: Benchmarking DoReMi with baselines in Predicting Reasoning Effort

evaluation models, and three proprietary LLMs—gemini-2.5-pro, gemini-2.5-flash, o3-mini—to test cross-group consistency.

Figure 2 reports Interclass Correlation Coefficient Leyland & Groenewegen (2014) - ICC(2,1) across judge groups. LLM judges show good-to-excellent agreement with humans on the axes most predictive of effort—*Cognitive Level*, *Knowledge Dimension*, and *Reasoning Steps*—with  $ICC > 0.75$ , and approach inter-human reliability on most remaining axes. Open-source and proprietary judges align closely with each other (Appendix Fig. 11).

Bloom-axis judging is rubric-driven, simpler than verifying full solutions, and stable across judge groups, mitigating LLM-as-judge risks. We therefore use LLM-derived Bloom features in DoReMi.

## 5.2 REASONING EFFORT MODELING RESULTS

We evaluated candidate effort proxies—WAF, R2FCA, UCA, and RI—and analyzed their correlations with Bloom axes (Fig. 3). MRT emerged as the strongest signal, so we adopt it as the primary proxy and aggregate it across  $M$  reasoning models into a combined metric  $MRT_c$ . Model-specific weights (Figs. 24, 25) are learned to optimize either recall- or  $F_1$ -oriented objectives. The resulting aggregate achieves  $\approx 80\%$  precision for identifying high-effort problems (details later). While a future composite could integrate R2FCA, WAF, and related proxies, the  $MRT_c$ -based approach is a strong, practical baseline.

We train classifiers to detect high-effort questions (high  $MRT_c$ ). Figure 4 summarizes precision–recall trade-offs under recall- and  $F_1$ -optimized settings. On **GPQA**, optimizing for recall selects logistic regression (recall 0.90, precision 0.62); optimizing for macro- $F_1$  favors a linear SVM (precision 0.81, recall 0.71). Five-fold CV yields  $F_1 = 63.4\% \pm 4.9\%$ , precision  $57.2\% \pm 2.4\%$ , and recall  $71.5\% \pm 9.3\%$  (Fig. 27). Our evaluation spans both science-focused benchmarks (GPQA, SuperGPQA-HEP) and general reasoning tasks (ARC) to demonstrate that Bloom taxonomy metrics can effectively capture reasoning difficulty across different domains. On **SuperGPQA-HEP**, SVM attains balanced performance (precision 75%, recall 71%). On **ARC**, SVM prioritizes recall

(87%) at lower precision (54%). Across benchmarks, the models remain interpretable and require no dataset-specific feature engineering, further supporting the domain-agnostic utility of Bloom taxonomy features for reasoning effort prediction.

### 382 5.3 DoReMi USE CASE 1: DIFFICULTY-AWARE SUBSET SELECTION

384 As state-of-the-art LLMs achieve near-saturation performance on many benchmarks (e.g., 85-88% on  
 385 GPQA), distinguishing between models becomes increasingly challenging. Difficulty-aware subset  
 386 selection addresses this by strategically sampling challenging problems to create more discriminative  
 387 evaluation sets. We compare three approaches for identifying high-effort problems:

388 **DoReMi (Ours):** Leverages learned reasoning effort models based on 3D Bloom taxonomy metrics  
 389 to predict  $\hat{MRT}_c$ , capturing nuanced aspects of problem difficulty beyond surface-level complexity.  
 390 The configurable effort metric  $MRT_c^*$  allows optimization for different objectives (recall vs. F1),  
 391 enabling flexible prioritization of either coverage or precision in identifying high-effort questions.

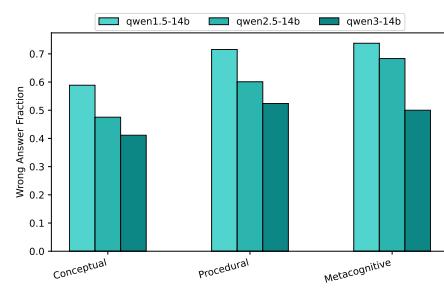
392 **Sublime-D Baseline:** A specialized variant within the SubLIME framework Xu et al. (2024) that  
 393 uses static readability metrics (Flesch, Gunning Fog) to estimate difficulty. While computationally  
 394 efficient, these surface-level metrics may miss deeper reasoning challenges that make problems truly  
 395 difficult for LLMs.

397 **LLM Rubric Baseline:** Employs a judge LLM (o4-mini-high) to directly classify reasoning effort by  
 398 analyzing problem statements and reference solutions holistically. This end-to-end approach serves  
 399 as a methodological contrast to DoReMi’s structured, multi-dimensional analysis (full prompt in  
 400 Appendix 7.6.1).

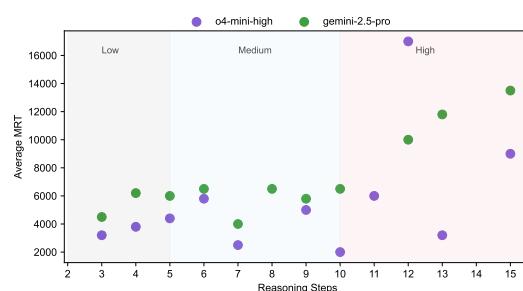
401 Figure 5 demonstrates DoReMi’s superior performance in identifying genuinely challenging prob-  
 402 lems. DoReMi achieves 54% accuracy in categorizing high reasoning effort problems, significantly  
 403 outperforming both the LLM rubric (25%) and Sublime-D (27%) methods. This 2 $\times$  improvement  
 404 suggests that structured analysis through Bloom taxonomy features provides more reliable difficulty  
 405 assessment than either static readability metrics or direct LLM classification.

406 The practical impact is substantial: when selecting a discriminative subset from benchmarks,  
 407 DoReMi’s higher accuracy ensures more “real” hard samples are included, preserving evaluation  
 408 power even as aggregate performance saturates. This capability enables more informed curriculum  
 409 learning protocols and facilitates targeted benchmark refinement as LLM capabilities continue to  
 410 advance.

### 411 5.4 DoReMi USE CASE 2 - INTERPRETABLE DIAGNOSTICS OF REASONING CAPABILITY



414  
 415 Figure 6: Qwen 14B model family analysis:  
 416 Wrong answer fractions along knowledge-  
 417 dimension axis



418 Figure 7: Average MRT vs Reasoning steps

424 **Cognitive Level.** Fig 1a shows the *expected number of R2FCA*, grouped by Bloom’s cognitive  
 425 levels. For o1-mini, the curve rises sharply—roughly quadrupling from APPLY to ANALYZE,  
 426 and again from ANALYZE to EVALUATE. Since this metric is inversely related to a model’s per-  
 427 attempt success rate, the steep slope shows that early-generation reasoning models experience an  
 428 exponential drop in hit rate as soon as multi-step evaluation or hypothesis testing is needed. In contrast,  
 429 gemini-2.5-pro maintains an almost flat profile. Its expected runs change little across the three

432 cognitive tiers, suggesting both a higher baseline competence and a much stronger ability to transfer  
 433 reasoning strategies as the cognitive level of the science problem increases.  $\circ 3\text{-mini}\text{-high}$  and  
 434  $\circ 4\text{-mini}\text{-high}$  fall between these extremes. This suggests a smooth, but not yet complete, scaling  
 435 path where each new model narrows the gap between **APPLY**-level heuristics and **EVALUATE**-level  
 436 analytical reasoning.

437 **Knowledge Dimension.** Fig 1b reports the *wrong-answer fraction* (WAF), grouped by Bloom’s  
 438 knowledge dimensions.  $\circ 1\text{-mini}$  again displays a monotonically rising error profile: its WAF  
 439 climbs from the **CONCEPTUAL** band through **PROCEDURAL** and peaks at the **METACOGNITIVE**  
 440 tier. Scaling post-training in  $\circ 3\text{-mini}\text{-high}$  and  $\circ 4\text{-mini}\text{-high}$  eliminate the gap between  
 441 **CONCEPTUAL** and **PROCEDURAL** questions and even invert the trend at **METACOGNITIVE**. A similar  
 442 transition is found in Claude: without reasoning tokens, performance of `claude-3.7-sonnet`  
 443 tracks  $\circ 1\text{-mini}$  almost exactly—error rates rise steadily with the knowledge dimension, suggest-  
 444 ing that the core model alone shares the same weakness. When reasoning tokens are enabled in  
 445 `claude-3.7-sonnet-extended-thinking`, the pattern breaks: errors grow only slightly  
 446 from **CONCEPTUAL** to **PROCEDURAL** and *decrease* at **METACOGNITIVE**. The open-source Qwen-  
 447 14B family shows the same story (Fig. 6): Qwen1.5-14B and Qwen2.5-14B—which lack RL-  
 448 based post-training—exhibit a monotonic rise from **CONCEPTUAL** to **METACOGNITIVE**, whereas the  
 449 reasoning-tuned Qwen3-14B reverses the trend with lower error at **METACOGNITIVE** than **PROCE-  
 450 DURAL**. Because RL appears only in the Qwen 2.5→3 transition, these results provide converging  
 451 evidence that RL might be the driver of the emergent metacognitive capability.

452 **Method Difficulty.** Fig 1c reveals a clear stratification in how successive model generations respond  
 453 to increasing *method difficulty*. For  $\circ 1\text{-mini}$ , the wrong-answer fraction rises sharply—from  
 454 roughly 20% at *Low* to 42% at *Medium*, and up to 100% at *High*.  $\circ 3\text{-mini}\text{-high}$  exhibits the  
 455 same monotonic trend but with a noticeably shallower slope. In contrast,  $\circ 4\text{-mini}\text{-high}$  shows  
 456 a qualitatively different pattern: its error rate increases from *Low* to *Medium* and then *plateaus*,  
 457 remaining statistically unchanged from *Medium* to *High*, which indicates a newfound robustness.  
 458 Compared to the knowledge-dimension results, these observations point to a *staggered emergence* of  
 459 higher-order abilities: sensitivity to *what* knowledge is required appears earlier (in  $\circ 3\text{-mini}\text{-high}$ ),  
 460 while resilience to *how* that knowledge must be operationalized emerges one generation later (in  
 461  $\circ 4\text{-mini}\text{-high}$ ). This suggests that scaled RL post-training yields non-uniform gains across  
 462 difficulty axes, with different dimensions reaching their inflection points at distinct stages.

462 **Solution-Step Complexity vs. Minimum Reasoning Tokens.** Figure 7 reveals a *piecewise* relation-  
 463 ship between the annotated number of solution steps and the *minimum* reasoning tokens required  
 464 by the most capable models in our study,  $\circ 4\text{-mini}\text{-high}$  (●) and `gemini-2.5-pro` (●). In the  
 465 **low-complexity regime** (steps < 5), both models quickly converge on concise answers—the average  
 466 reasoning token *falls* as the step count decreases. Once problems demand **more than 10 explicit  
 467 steps**, each additional step now incurs a significant rise in the minimum token count. This suggests  
 468 that beyond a critical complexity threshold the models must maintain longer context windows to  
 469 keep intermediate facts “alive,” and the cost scales super-linearly with step count. Intriguingly, in the  
 470 **transition band** ( $5 \leq \text{steps} \leq 10$ ) the correlation is almost flat for both systems. Together, these  
 471 three regimes highlight an emergent efficiency plateau followed by an exponential token explosion.

472

473

## 474 6 CONCLUSION

475

476

477 We proposed DoReMi, a structured Bloom-inspired framework for quantifying science problem  
 478 difficulty for LLMs. Evaluations on GPQA, ARC, and SuperGPQA demonstrated good correlations  
 479 (up to ( $r=0.52$ )) between our multidimensional difficulty fingerprints and empirical reasoning effort  
 480 metrics MRT and expected R2FCA. Our reasoning-effort prediction model significantly outperformed  
 481 static-difficulty baselines: DoReMi (54%) vs Sublime-D (27%). Moreover, through interpretable  
 482 diagnostics of reasoning capabilities, we identified emergent reasoning capabilities—such as improved  
 483 robustness to methodologically challenging problems and enhanced metacognitive monitoring—that  
 484 have appeared systematically across successive generations of reasoning LLMs. These insights could  
 485 potentially guide future targeted model improvements, curriculum design, and benchmark creation,  
 486 particularly highlighting the importance of careful post-training strategies for fostering higher-order  
 487 reasoning capabilities.

486 REFERENCES  
487

488 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and  
489 Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge.  
490 *arXiv:1803.05457v1*, 2018.

491 DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.

492

493 Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. The unreasonable effectiveness of easy  
494 training data for hard tasks, 2024. URL <https://arxiv.org/abs/2401.06751>.

495

496 Rex Heer. A model of learning objectives — based on \*a taxonomy for learning,  
497 teaching, and assessing: A revision of bloom’s taxonomy of educational ob-  
498 jectives\*. Center for Excellence in Learning and Teaching, Iowa State University,  
499 2012. URL <https://www.learningoutcomesassessment.org/wp-content/uploads/2019/10/RevisedBloomsHandout.pdf>. Three-dimensional representation  
500 of the revised Bloom’s taxonomy; updated January 2012.

501

502 Thomas Huber and Christina Niklaus. Llms meet bloom’s taxonomy: A cognitive view on large  
503 language model evaluations. In *Proceedings of the 31st International Conference on Computational*  
504 *Linguistics (COLING)*, pp. 5211–5246, Abu Dhabi, UAE, 2025. Association for Computational  
505 Linguistics. doi: 10.18653/v1/2025.coling-main.350. URL <https://aclanthology.org/2025.coling-main.350/>.

506

507 Alastair H. Leyland and Peter P. Groenewegen. *Intraclass Correlation Coefficient (ICC)*, pp.  
508 3367–3368. Springer Netherlands, Dordrecht, 2014. ISBN 978-94-007-0753-5. doi: 10.1007/  
509 978-94-007-0753-5\_1528. URL [https://doi.org/10.1007/978-94-007-0753-5\\_1528](https://doi.org/10.1007/978-94-007-0753-5_1528).

510

511 G. Harry McLaughlin. Smog grading-a new readability formula. *Journal of Reading*, 12(8):639–646,  
512 1969.

513

514 OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>.

515

516 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien  
517 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a  
518 benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.

519

520 Donya Rooein, Paul Rottger, Anastassia Shaitarova, and Dirk Hovy. Beyond flesch-kincaid: Prompt-  
521 based metrics improve difficulty classification of educational texts, 2024. URL <https://arxiv.org/abs/2405.09482>.

522

523 Brian Scott. The gunning fog index (or fog) readability  
524 formula. <https://readabilityformulas.com/the-gunnings-fog-index-or-fog-readability-formula/>, 2025. Accessed:  
525 2025-05-11.

526

527 Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation  
528 metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and*  
529 *Metrics (GEM 2021)*, pp. 1–14, Online, 2021. Association for Computational Linguistics. doi:  
530 10.18653/v1/2021.gem-1.1. URL <https://aclanthology.org/2021.gem-1.1/>.

531

532 M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao  
533 Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixing Deng, Shuyue Guo, Shian  
534 Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, Dehua Ma,  
535 Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, Ming Xu,  
536 Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang  
537 Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang,  
538 Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma,  
539 Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Yizhou Tan, Zili Wang, Chenqing

540 Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue,  
 541 Tianyang Zhan, Chun Zhang, Jingyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi  
 542 Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu  
 543 Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min  
 544 Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin,  
 545 Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines,  
 546 2025. URL <https://arxiv.org/abs/2502.14739>.

547 Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Un-  
 548 derstanding chain-of-thought length in llms, 2025. URL <https://arxiv.org/abs/2502.07266>.

549 Cong Xu, Gayathri Saranathan, Mohammad Parwez Alam, Arpit Shah, James Lim, Soon Yee Wong,  
 550 Foltin Martin, and Suparna Bhattacharya. Data efficient evaluation of large language models and  
 551 text-to-image models via adaptive sampling, 2024. URL <https://arxiv.org/abs/2406.15527>.

552 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
 553 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,  
 554 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
 555 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,  
 556 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui  
 557 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang  
 558 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger  
 559 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan  
 560 Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

561

## 562 7 TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

### 563 7.1 THE USE OF LLMs IN MANUSCRIPT PREPARATION

564 In accordance with ICLR 2026 submission guidelines, we disclose the use of Large Language Models  
 565 (LLMs) in the preparation of this manuscript. LLMs were employed as general-purpose assist tools  
 566 for the following specific purposes:

- 567 • **Writing assistance:** LLMs were used to improve clarity, grammar, and flow of certain  
 568 sections of the manuscript, particularly in refining technical explanations and ensuring  
 569 consistent terminology throughout the paper.
- 570 • **Literature review support:** LLMs assisted in identifying relevant related work and helped  
 571 structure the presentation of background material, though all cited works were independently  
 572 verified by the authors.
- 573 • **Code documentation:** LLMs were used to generate comments and documentation for  
 574 supplementary code materials to improve readability and reproducibility.

575 We emphasize that the core research ideas, experimental design, methodology, analysis, and conclusions  
 576 presented in this work are entirely the intellectual contribution of the human authors. LLMs did  
 577 not play a significant role in research ideation or the formulation of novel contributions. All factual  
 578 claims, experimental results, and scientific interpretations have been independently verified by the  
 579 authors.

580 The authors take full responsibility for all content in this manuscript, including any text that may have  
 581 been refined with LLM assistance. We confirm that no content generated by LLMs could be construed  
 582 as plagiarism or scientific misconduct, and all sources and prior work are properly attributed.

### 583 7.2 FULL BLOOM'S TAXONOMY PROMPT TEMPLATES

584 This appendix provides the full prompt templates used for automated, LLM-based annotation of science  
 585 problems along the extended 3D Bloom's taxonomy. Each template is optimized for consistency

594 and clarity, facilitating accurate evaluation by large language models. Prompts are presented verbatim  
 595 as used in our experiments.  
 596

597

598 **PROMPT 1: COGNITIVE LEVEL ASSESSMENT**

599

600 You are an expert **in** evaluating scientific problems using Bloom's  
 601 Taxonomy. Your task is to assess a given problem statement and its  
 602 reference solution to determine the highest cognitive process level  
 603 required for its resolution. Focus solely on the cognitive actions  
 604 necessary to solve the problem. Your analysis should be objective,  
 605 detailed, and applicable across all scientific subdomains.  
 606

607

608 The cognitive process levels (from lowest to highest) are defined as  
 609 follows:  
 610

611 1. Remember : Involves the recall or retrieval of factual information,  
 612 definitions, or previously learned material without modification or  
 613 interpretation.  
 614 \*Example tasks:\* Listing key facts, reciting definitions, or recalling  
 615 formulas.

616

617 2. Understand : Involves demonstrating comprehension by interpreting,  
 618 summarizing, or explaining concepts in your own words.  
 619 \*Example tasks:\* Paraphrasing theories, summarizing research findings,  
 620 or explaining the meaning of concepts.

621

622 3. Apply : Involves using known information, methods, or procedures in  
 623 specific or novel situations.  
 624 \*Example tasks:\* Solving standard problems using known formulas,  
 625 applying theories to new contexts, or executing established  
 626 procedures.

627

628 4. Analyze : Involves breaking complex information into parts to examine  
 629 relationships, identify patterns, and differentiate between  
 630 components.  
 631 \*Example tasks:\* Decomposing arguments, comparing and contrasting  
 632 elements, or mapping relationships within a system.

633

634 5. Evaluate : Involves making judgments based on criteria or standards  
 635 by critiquing, assessing, or validating theories, methods, or  
 636 evidence.  
 637 \*Example tasks:\* Critically assessing the validity of a hypothesis,  
 638 weighing evidence, or comparing alternative approaches.

639

640 6. Create : Involves synthesizing information to produce new or original  
 641 ideas, models, or solutions.  
 642 \*Example tasks:\* Formulating novel hypotheses, designing innovative  
 643 experiments, or constructing comprehensive models that integrate  
 644 diverse elements.

645

646 =====

647 Problem Statement:

648 <problem\_statement>  
 649 {problem\_statement}  
 650 </problem\_statement>

651

652 =====

653 Reference Solution:

654 <reference\_solution>  
 655 {reference\_solution if reference\_solution is not None else None }  
 656 </reference\_solution>

657

658 Instructions:

648 a. Read the problem statement and reference solution carefully.  
 649 b. Identify the specific cognitive actions required for a correct  
 650 solution.  
 651 c. Determine the highest cognitive process level (from the list above)  
 652 necessary to solve the problem.  
 653 d. Provide a detailed rationale for your rating by citing specific  
 654 elements from the problem statement and reference solution that  
 655 indicate the required cognitive processes.  
 656  
 657 Now output your answer as a JSON object with two keys: cognitive\_level  
 658 (which must be one of Remember, Understand, Apply, Analyze,  
 659 Evaluate, or Create) and rationale. Do not include any  
 660 additional commentary.

661  
 662 **PROMPT 2: METHOD DIFFICULTY ASSESSMENT**

663 You are an expert **in** evaluating scientific problems using Bloom's 3D  
 664 Taxonomy.  
 665  
 666 **\*\*Objective:\*\***  
 667 Evaluate the difficulty of the method required to solve a problem,  
 668 focusing exclusively on the methods characteristics. Base your  
 669 judgment solely on the method's nature-**not** on the overall complexity  
 670 of the scientific **or** biological context.  
 671  
 672 **\*\*Classification Criteria:\*\***  
 673 - **\*\*Low Method Difficulty:\*\***  
 674 - The problem **is** solved by directly applying a well-known, established  
 675 method exactly **as** described **in** standard references.  
 676 - No modifications, adaptations, **or** additional interpretative steps are  
 677 necessary.  
 678 - **\*\*Medium Method Difficulty:\*\***  
 679 - The solution requires adapting **or** modifying a known method to fit the  
 680 specific problem constraints **or** context.  
 681 - This may include reconciling contradictory observations, integrating  
 682 multiple pieces of evidence, **or** ruling out alternative hypotheses  
 683 before arriving at a conclusion.  
 684 - **Threshold:** If even a single additional step beyond the basic method  
 685 **is** needed (e.g., adjusting **for** indirect binding signals **or** combining  
 686 two standard analyses), classify **as** Medium.  
 687  
 688 - **\*\*High Method Difficulty:\*\***  
 689 - The problem demands a creative **or** non-routine approach that goes well  
 690 beyond standard adaptations.  
 691 - This includes devising new frameworks, integrating multiple disparate  
 692 methods **in** unconventional ways, **or** reasoning through unfamiliar **or**  
 693 abstract concepts.  
 694 - **Threshold:** If multiple adaptations, non-linear reasoning, **or**  
 695 innovative synthesis of several methods **is** required, classify **as** High  
 696 .  
 697  
 698 **\*\*Evaluation Steps:\*\***  
 699 1. **\*\*Examine the Method:\*\***  
 700 - Focus exclusively on the method employed **in** the solution. Ignore the  
 701 overall biological **or** technical complexity unless it directly  
 702 impacts the methods execution.  
 703  
 704 2. **\*\*Determine the Degree of Adaptation Required:\*\***  
 705 - **\*\*Direct Application (Low):\*\*** If the method **is** used **in** a textbook,  
 706 unmodified way.

```

702     - **Minor Adaptations (Medium):** If a standard method is modified or
703     augmented by one extra layer of interpretation or integration.
704     - **Significant Innovation (High):** If the solution requires
705     combining multiple methods, developing a new approach, or applying
706     the method in a highly non-standard or creative manner.
707
708 3. **Provide a Detailed Rationale:**
709     - Cite specific aspects of the method in the problem statement and
710     reference solution that indicate whether the method is used routinely
711     or has been significantly adapted.
712     - Explain why any additional steps or integrations push the
713     classification toward Medium or High.
714
715 4. **Maintain Objectivity:**
716     - Base your classification solely on the nature and execution of the
717     method not on the underlying scientific problem.
718     - Emphasize the number and significance of modifications required to
719     apply the method.
720
721 =====
722 Problem Statement:
723 <problem_statement>
724 {problem_statement}
725 </problem_statement>
726
727 =====
728 Reference Solution:
729 <reference_solution>
730 {reference_solution if reference_solution is not None else None }
731 </reference_solution>
732
733 =====
734
735 735 PROMPT 3: DEFINITION COMPLETENESS ASSESSMENT
736
737
738 You are an expert in evaluating scientific problems using Bloom's 3D
739 Taxonomy.
740 Please analyze the following problem statement and reference solution,
741 focusing exclusively on whether the problem is completely defined.
742 Consider the following:
743 - Is the problem completely defined with all necessary details provided?
744     If yes, answer yes; if it requires assumptions or strategic
745     decisions, answer no.
746 Provide a detailed rationale by citing specific aspects of the problem
747     statement and reference solution.
748
749 =====
750 Problem Statement:
751 <problem_statement>
752 {problem_statement}
753 </problem_statement>
754
755 =====
756 Reference Solution:
757 <reference_solution>
758 {reference_solution if reference_solution is not None else None }
759 </reference_solution>

```

756     Output your answer as a JSON object with two keys `completely_defined` (

757        yes or no ) and `rationale` . Do not include any additional

758        commentary.

759

760

761

762     **PROMPT 4: KNOWLEDGE DIMENSION ASSESSMENT**

763     You are an expert **in** evaluating scientific problems using Bloom's 3D

764        Taxonomy.

765     Please analyze the following problem statement and reference solution,

766        focusing exclusively on the type of knowledge required.

767     The knowledge dimensions (from lowest level to highest level) are defined

768        as:

- Factual : Involves recalling basic facts or definitions.
- Conceptual : Involves understanding theories, principles, and relationships.
- Procedural : Involves knowing how to perform methods or algorithms.
- Metacognitive : Involves awareness and control of one's own thinking processes.

774     Your tasks:

1. Determine the **\*\*highest level\*\*** of knowledge dimension(s) the problem targets.
2. Provide a detailed rationale for your classification by citing specific elements from the problem statement (and reference solution, if provided) .

780     Problem Statement:

```
<problem_statement>
{problem_statement}
</problem_statement>
```

784     Reference Solution:

```
<reference_solution>
{reference_solution if reference_solution is not None else None }
</reference_solution>
```

788     Please output your answer as a JSON object with two keys:

789        `knowledge_dimension` (highest level of Factual , Conceptual ,

790        Procedural , Metacognitive required) and `rationale` . Do not

791        include any additional commentary.

792

793

794

795     **PROMPT 5: KNOWLEDGE BREADTH ASSESSMENT**

796     You are an expert **in** evaluating scientific problems using an extended

797        version of Bloom's Taxonomy.

798     Focus exclusively on the 'Knowledge: Breadth' subdimension. This

799        dimension examines whether the problem

800        requires integrating knowledge from multiple disciplines.

801     By order of complexity, from low to high, this dimension's levels are

802        defined **as**:

- Single-Discipline : Does **not** require integrating knowledge **from** multiple disciplines.
- Multi-Discipline : Requires integrating knowledge **from** multiple disciplines.

806     Your tasks:

1. Analyze the problem statement **and** reference solution.
2. Determine **if** the problem involves multiple disciplines (i.e., Multi-Discipline ) **or is** confined to one.
3. Provide a detailed rationale citing specific elements **from** the text.

```

810
811 Output your answer as a JSON object with two keys: knowledge_breadth
812 and rationale . Do not include any additional commentary.
813
814 Problem Statement:
815 {problem_statement}
816
817 Reference Solution:
818 {reference_solution if reference_solution is not None else None }
819
820
821
822 PROMPT 6: REASONING STEPS ASSESSMENT
823
824 You are an expert in evaluating scientific problems using an extended
825 version of Bloom's Taxonomy.
826 Focus exclusively on the 'Reasoning: Number of Reasoning Steps Required
827 to Solve the Problem' subdimension. This dimension examines how many
828 distinct
829 reasoning or planning steps are required to solve the problem, and
830 whether the process is straightforward or complex.
831
832 Follow the guidelines below to count the number of distinct reasoning
833 steps:
834 1. **Identify Independent Logical Actions:**
835 - Count each separate act of analysis, deduction, computation, or
836 planning that contributes uniquely to arriving at the solution.
837 - A \step should represent an independent logical move rather than
838 a mere rephrasing or elaboration of an earlier step.
839
840 2. **Define the Boundaries of a Step:**
841 - A step begins when a new reasoning method or operation is introduced
842 (e.g., setting up an equation, deducing a chemical property, or
843 applying a theoretical principle).
844 - A step ends when that piece of reasoning has been fully applied or
845 resolved. Avoid splitting actions that are inherently part of one
846 unified idea.
847
848 3. **Ensure Each Step Is Essential:**
849 - Only count steps that are necessary for reaching the final
850 conclusion. Trivial clarifications, restatements of known facts, or
851 background information that does not directly contribute to the
852 reasoning process should not be counted as separate steps.
853 - Consider whether the removal of a step would leave a gap in the
854 logical progression toward the solution. If so, it must be counted.
855
856 4. **Differentiate Between Substeps and Major Steps:**
857 - When a step naturally divides into substeps, assess if those
858 substeps represent distinct reasoning actions. If they are tightly
859 interwoven and the separation does not change the logical flow, count
860 them as one step.
861 - Use clear criteria such as \new calculation, \new inference,
862 or \application of a different principle to decide on splitting or
863 merging substeps.
864
865 5. **Apply Consistent and Objective Criteria:**
866 - Use objective markers such as \setting up an equation, \balancing
867 mass or charge, \inferring molecular structure, \evaluating
868 experimental data, or \applying a theory to identify steps.
869 - Ensure that the criteria for what counts as a step are uniformly
870 applied, regardless of the scientific discipline (chemistry, physics,
871 biology, etc.).
872
873 6. **Document the Rationale for Each Counted Step:**

```

```

864     - For every reasoning step counted, provide a brief description that
865       clearly indicates why it is independent and essential to the overall
866       solution.
867     - This documentation should include reference to the specific part of
868       the problem or reasoning process it addresses.
869
870 7. **Sequence and Dependency Consideration:**
871     - Ensure that each counted step represents a sequential action that
872       builds on previous steps. Independent parallel reasoning that
873       converges into one conclusion may be considered as separate steps if
874       each contributes a distinct part to the final result.
875     - Avoid double-counting: if two pieces of reasoning essentially
876       support the same conclusion without introducing new independent
877       information, they should be merged into one step.
878 =====
879 Problem Statement:
880 <problem_statement>
881 {problem_statement}
882 </problem_statement>
883 =====
884 Reference Solution:
885 <reference_solution>
886 {reference_solution}
887 </reference_solution>
888 =====
889 Your tasks:
890 1. Analyze the problem statement and reference solution.
891 2. Output the number of distinct reasoning or planning steps required to
892   solve the problem.
893 3. Provide a detailed rationale citing specific elements from the text.
894
895 Output your answer as a JSON object with two keys: number_of_steps and
896   rationale. Do not include any additional commentary.
897
898

```

899 These prompt templates are designed to maximize annotation consistency and interpretability. For  
900 further implementation details, see Section 2.2.

901 Bloom Distribution Plots on GPQA, SuperGPQA and ARC To analyze how scientific problems  
902 are distributed across the Bloom taxonomy, we visualize the annotation results across three bench-  
903 marks: GPQA, SuperGPQA, and ARC. These distributions reveal the relative frequency of cognitive,  
904 knowledge-level, method difficulty demands placed by each benchmark.

905 **7.2.1 HUMAN ALIGNMENT FOR BLOOM METRICS**

906 **7.3 PLOTS FOR REASONING EFFORT METRICS**

907 We visualize the distribution of various reasoning metrics across models. These include:

- 908 • Wrong Answer Fraction (WAF) — Figure 12
- 909 • Minimum Reasoning Tokens for Correct Answer — Figure 13
- 910 • Expected Runs to Correctness — Figure 14
- 911 • Uncertainty in Correct Answers — Figure 15
- 912 • Inter-run Reasoning Inconsistency — Figure 16

913 These metrics serve as quantitative signals of reasoning effort and complexity in LLM behavior,  
914 complementing the structural annotations from the Bloom taxonomy.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

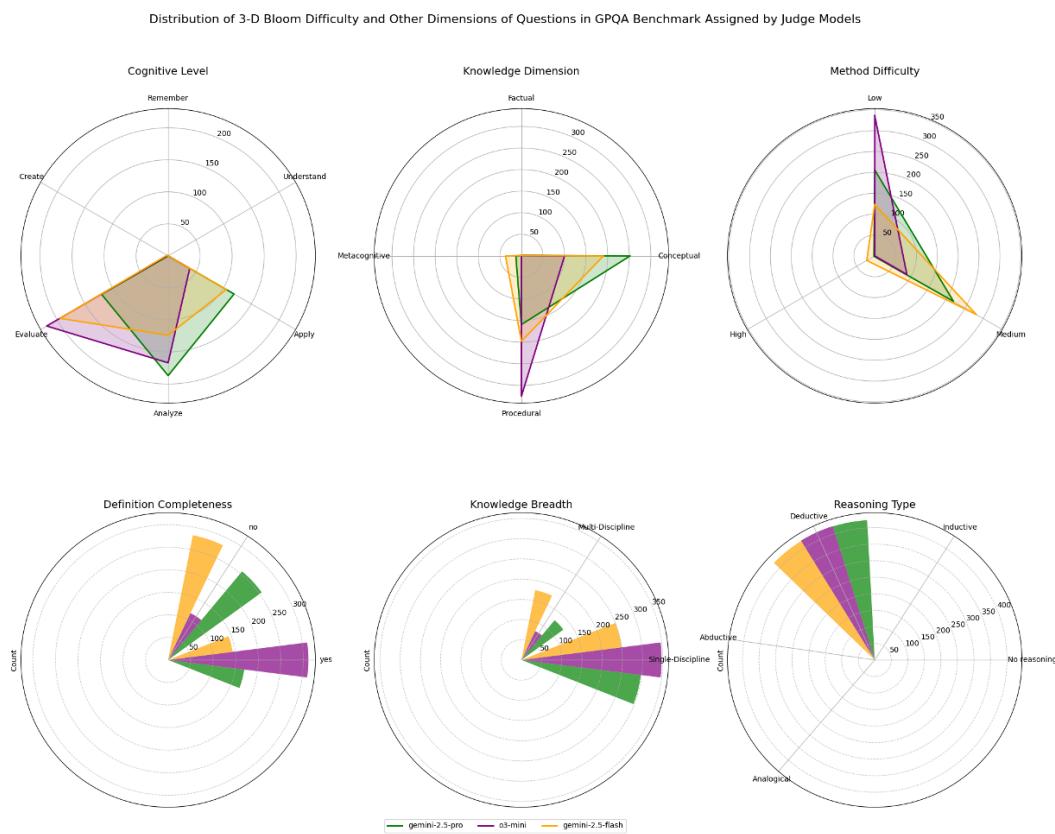


Figure 8: Bloom Metrics Distribution for GPOA



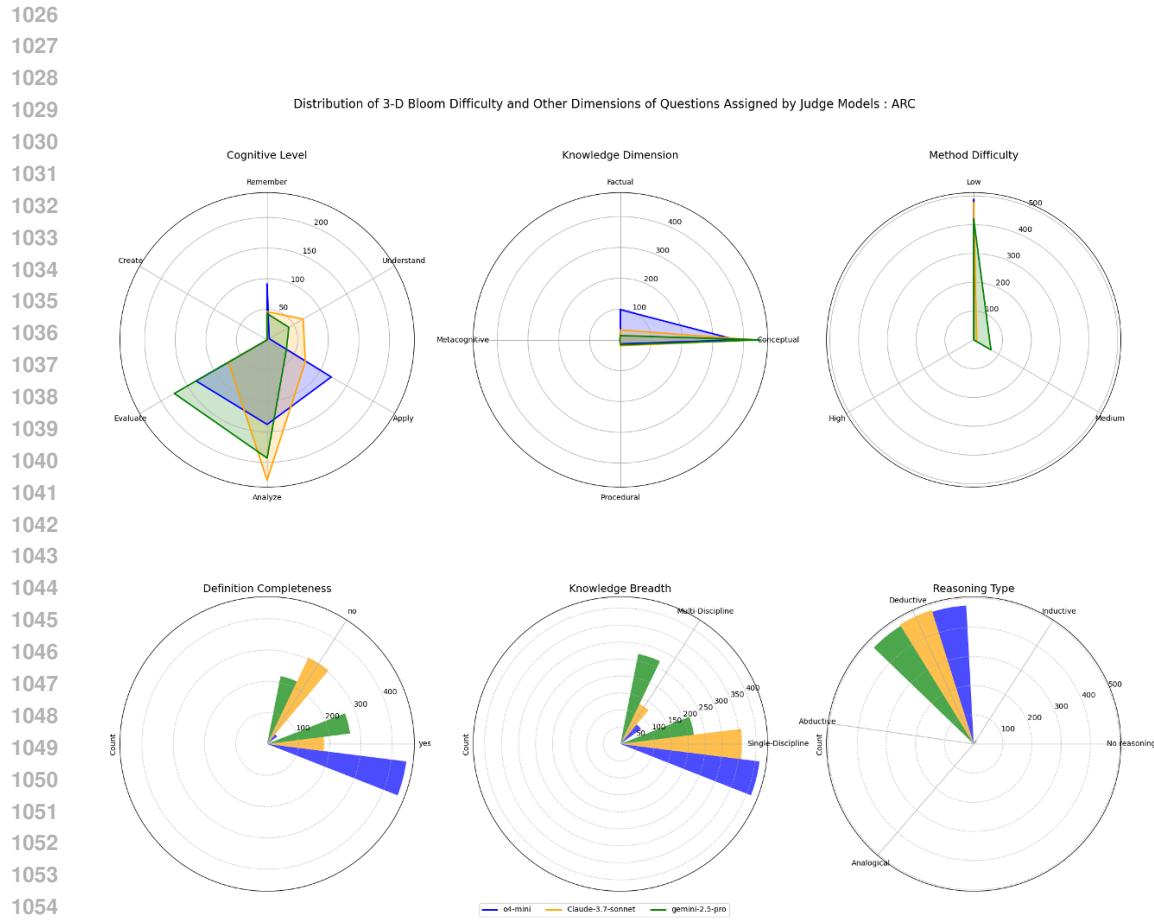


Figure 10: Bloom Metrics Distribution for ARC

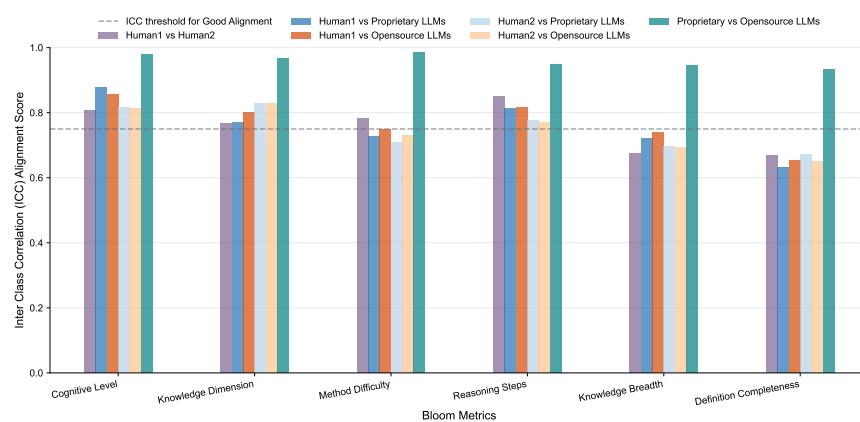


Figure 11: Human Alignments and LLM alignment

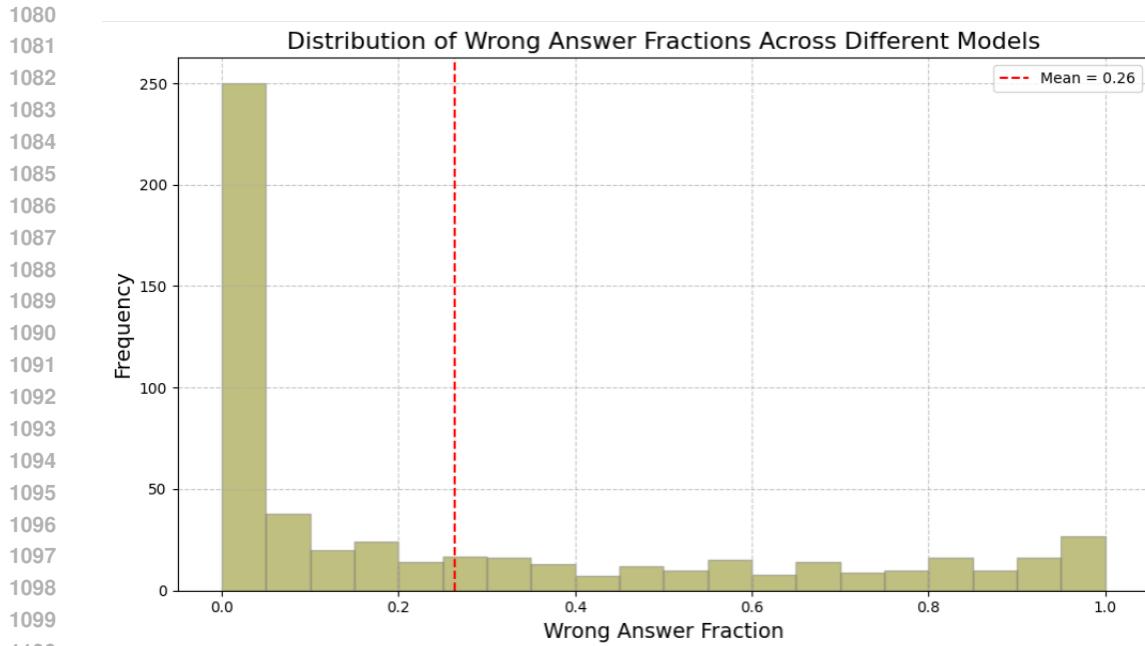


Figure 12: Distribution of Wrong Answer Fraction(WAF)

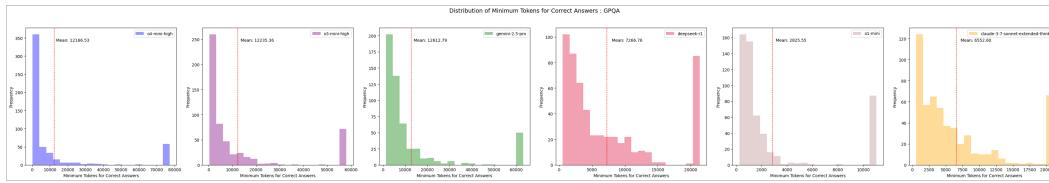
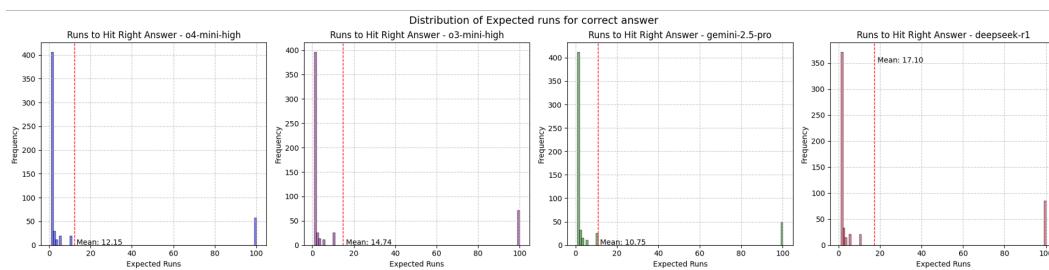
Figure 13: Distribution of Minimum Reasoning Token for Right Answer across Models  $M$ 

Figure 14: Distribution of Expected Runs for Correct Answer

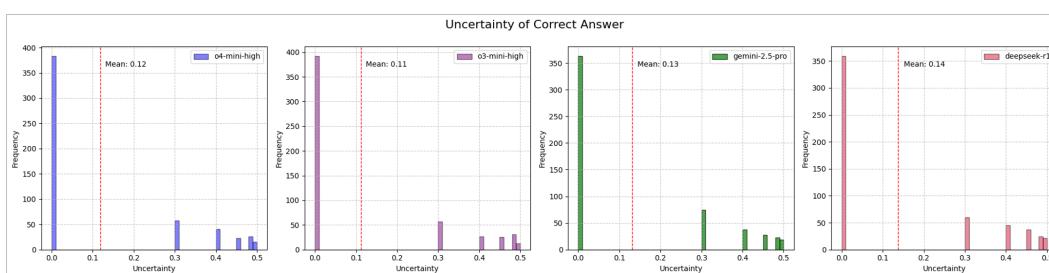


Figure 15: Distribution of Uncertainty of Correct Answers

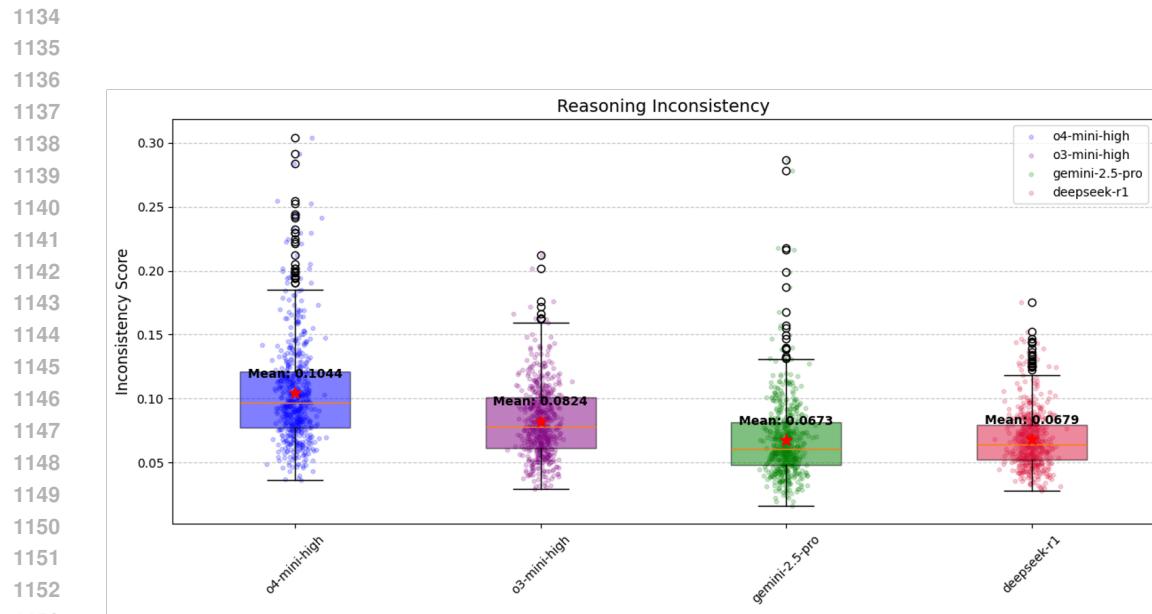


Figure 16: Distribution of Reasoning Inconsistency across 10 runs

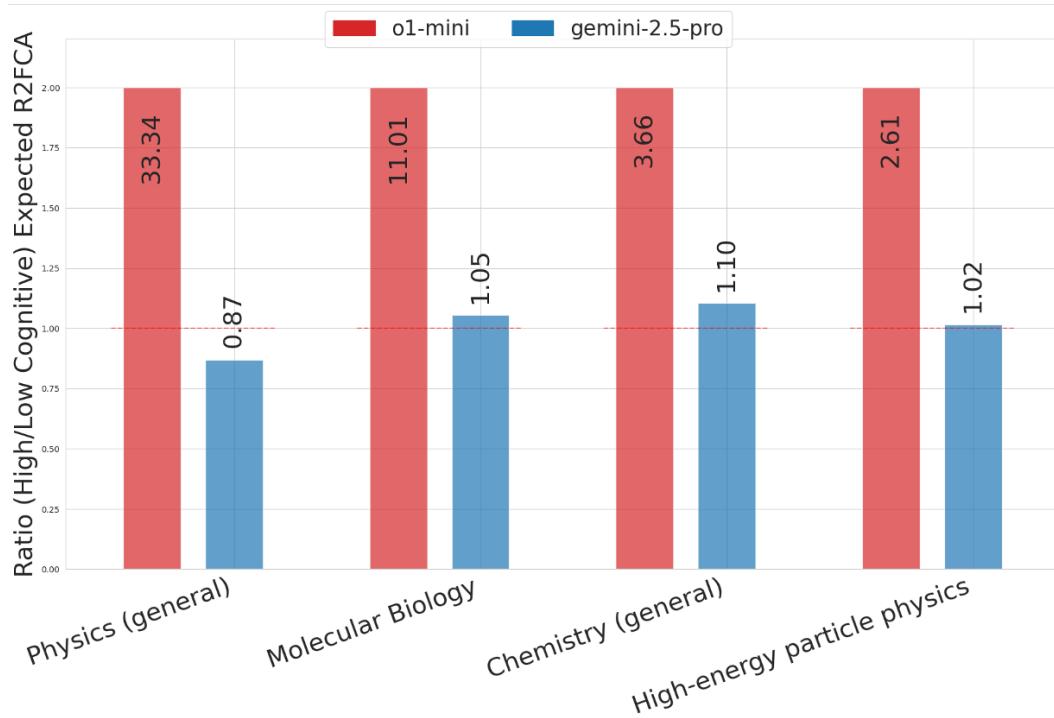


Figure 17: Ratio of High to Low Cognitive Expected R2FCA for o1-mini and Gemini-2.5-pro across 4 GPQA subfields

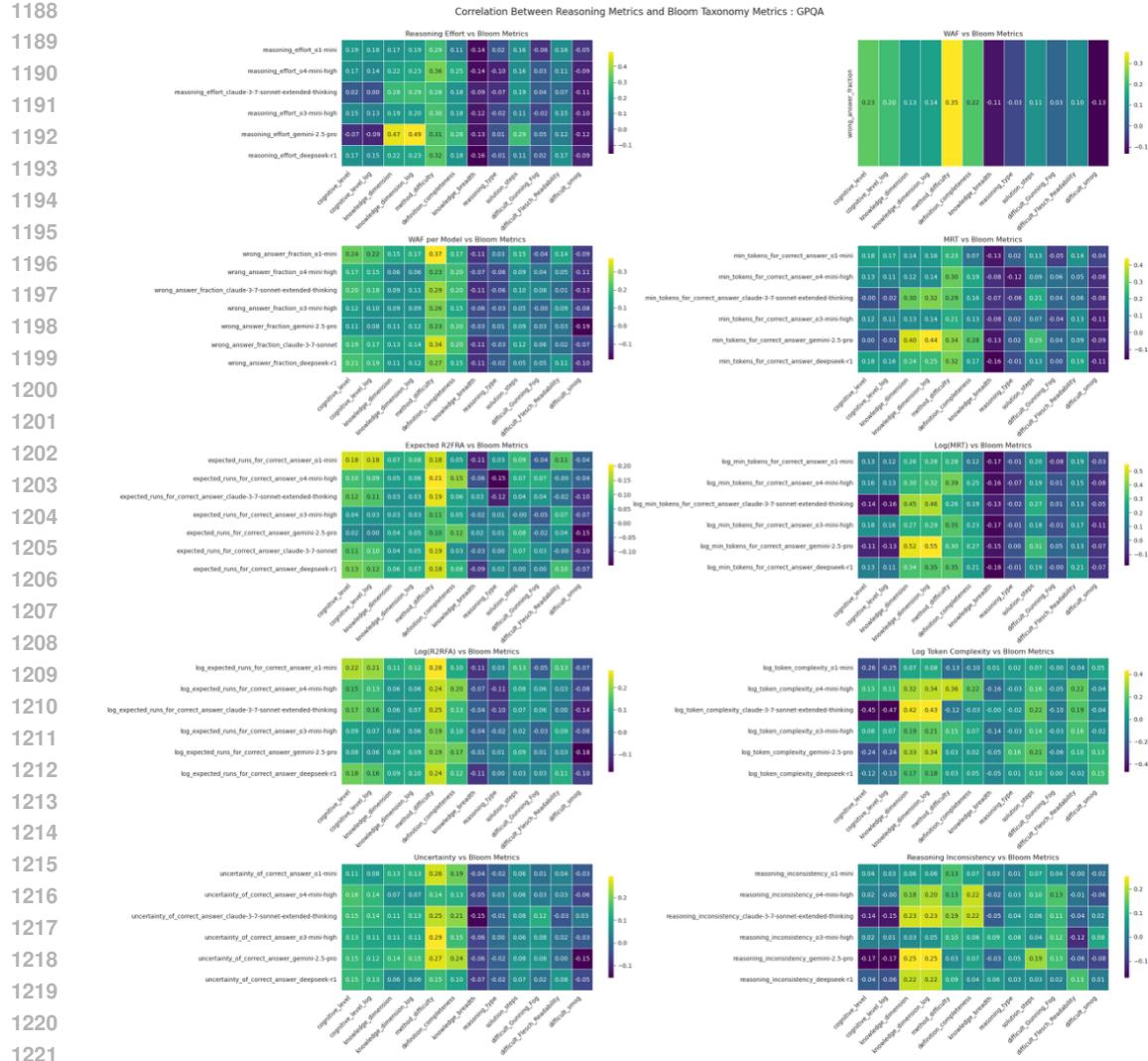


Figure 18: Correlation of Bloom Metrics across Reasoning Effort Metrics from individual reasoning models: GPQA

#### 7.4 EXTENDED CORRELATION ANALYSIS

For ARC we also leveraged the human labeled data provided by Clark et al. (2018) to perform correlation analysis to see if the human labeled metrics align well with the judge models Bloom prediction. The human labeled score was only for cognitive levels, it does not contain other dimensions of bloom taxonomy. We see that the human bloom scores correlate equally well as model assigned bloom scores across reasoning metrics.

#### 7.5 MODEL GENERATION COMPARISONS ON REASONING EFFORT

We compare different model generations—such as  $\text{o1-mini}$ ,  $\text{o3-mini}$ , and  $\text{o4-mini}$ —with respect to the reasoning effort required for problems of varying difficulty:

- Figure 22 compares minimum reasoning token lengths across knowledge difficulty classes.
- Figure 23 summarizes model behavior across effort metrics.

These comparisons illustrate emergent capabilities in later model generations, especially in handling high-difficulty tasks.

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255 Correlation Between Reasoning Metrics and Bloom Taxonomy Metrics : SuperGPQA HEP

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

Figure 19: Correlation of Bloom Metrics across Reasoning Effort Metrics from individual reasoning models : SuperGPQA

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

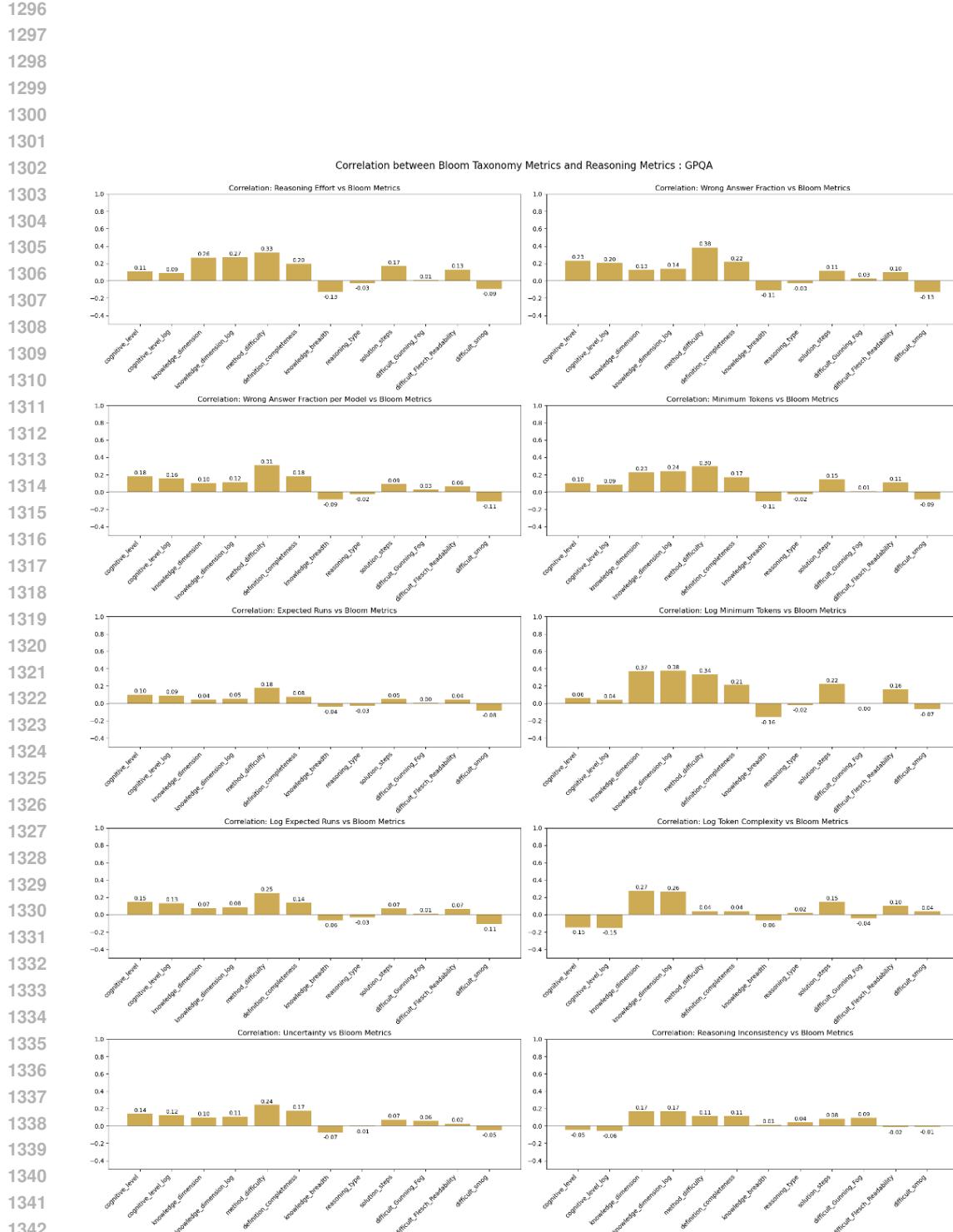


Figure 20: Overall correlation of Individual Reasoning Metrics w.r.t Bloom metrics

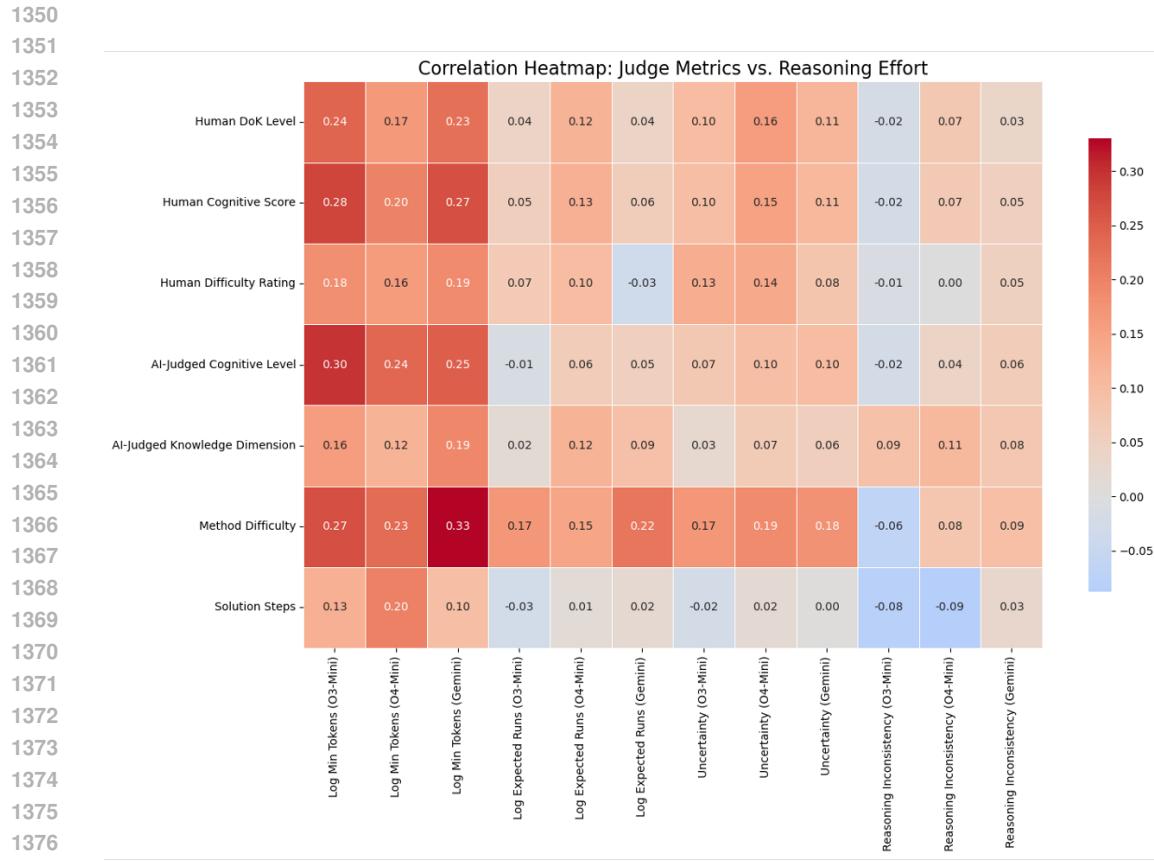


Figure 21: Correlation of Bloom Metrics across Reasoning Effort and Human Labeled Difficulty : ARC

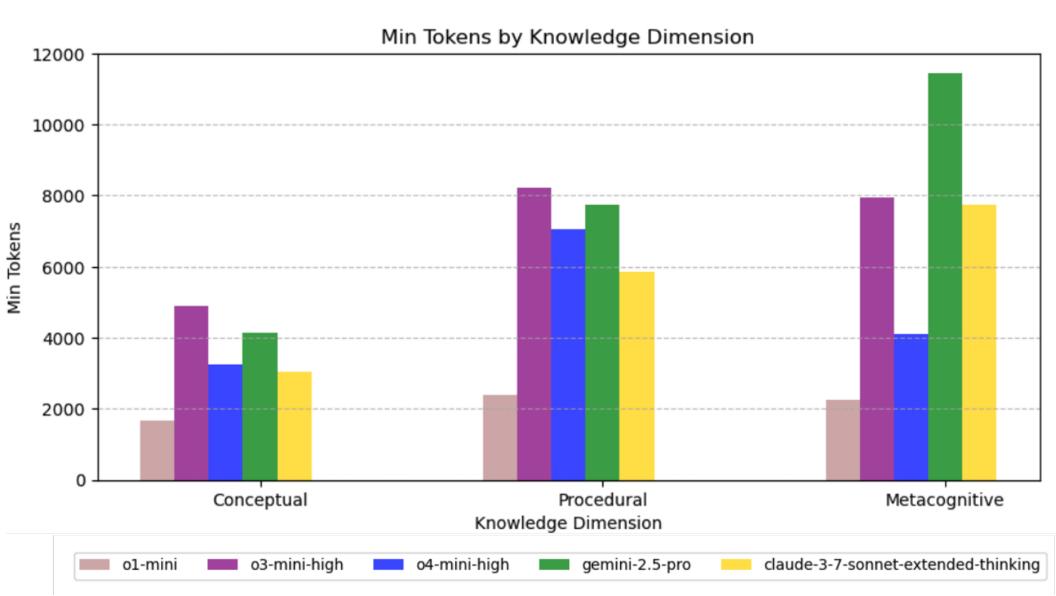


Figure 22: Minimum Reasoning Tokens to Right Answer across Knowledge Difficulty

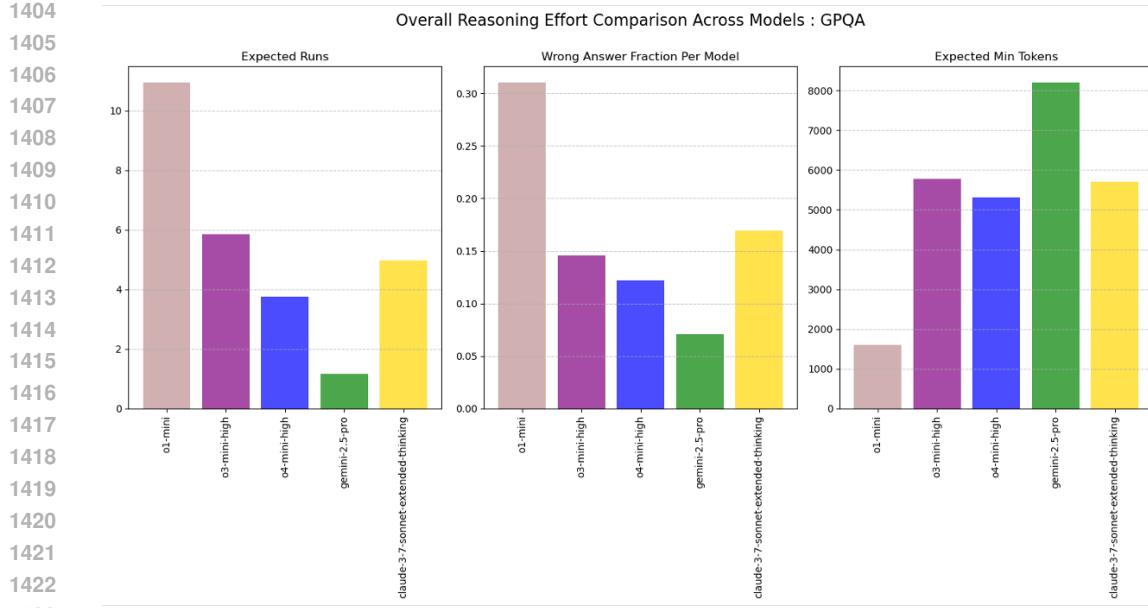


Figure 23: Overall analysis across Models

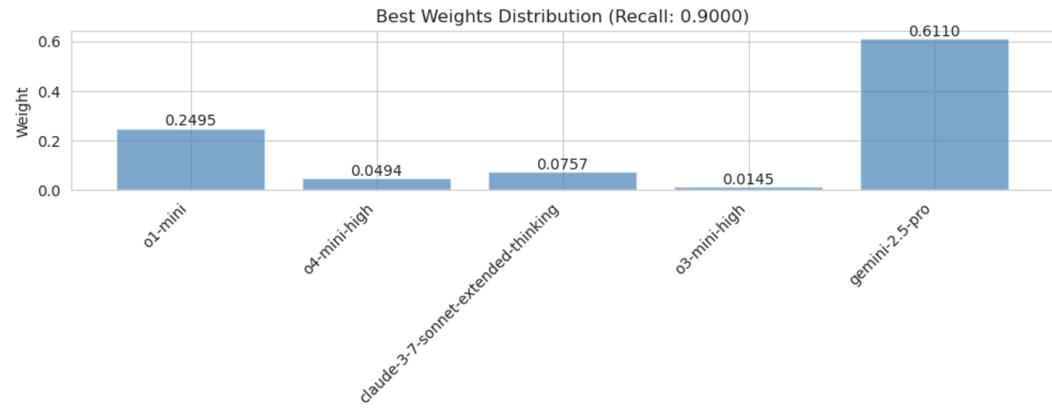


Figure 24: Weights Optimised for Recall : GPQA

## 7.6 REASONING EFFORT PREDICTION

Finally, we present results from Combined Reasoning effort model that predict reasoning effort from Bloom taxonomy features. Key findings include:

- Figure 24 and Figure 25 show weights learned for recall and F1-optimized classification, respectively.
- Figure 26 illustrates classifier performance for detecting high-effort reasoning cases optimized through f1.

### 7.6.1 PROMPT FOR LLM RUBIC CLASSIFICATION

Using a Judge LLM, in our experiment we have used o4-mini-high, to perform reasoning effort classification to serve as a baseline for DoReMi approach. The Judge LLM is given the problem statement, and reference solution as context to classify the reasoning effort required to solve the problem.

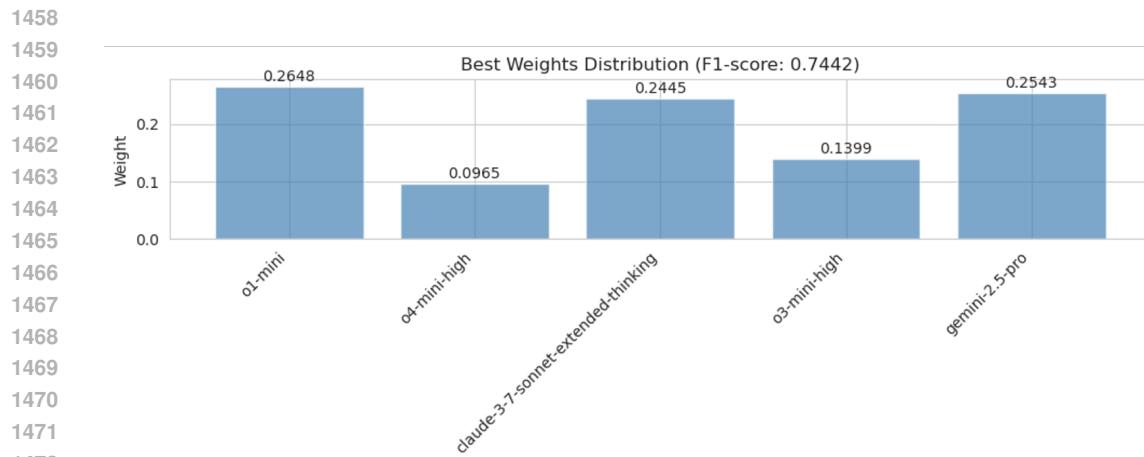


Figure 25: Weights Optimized for F1-score: GPQA

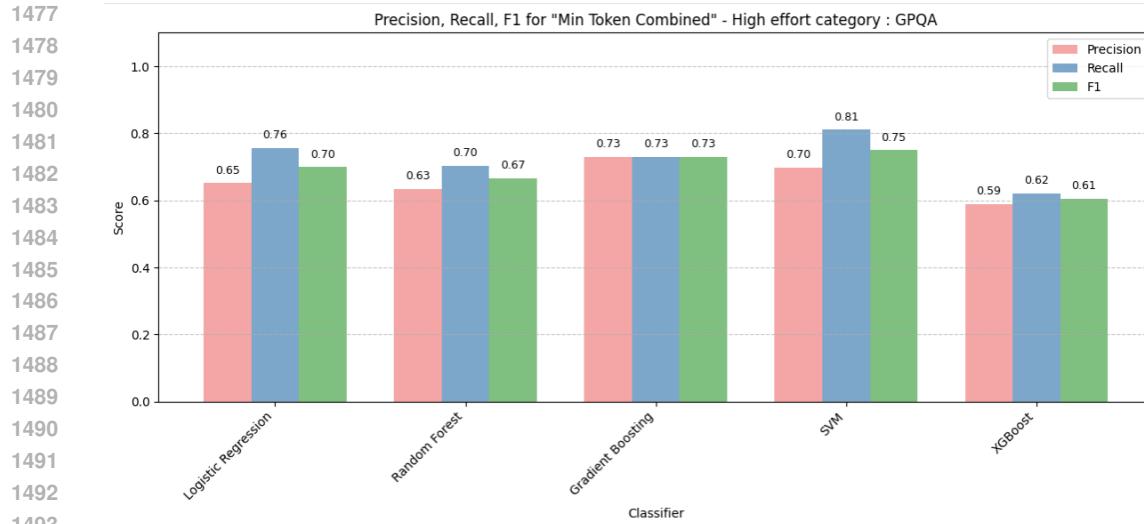


Figure 26: Model Prediction Metrics for Classifying High Reasoning Effort

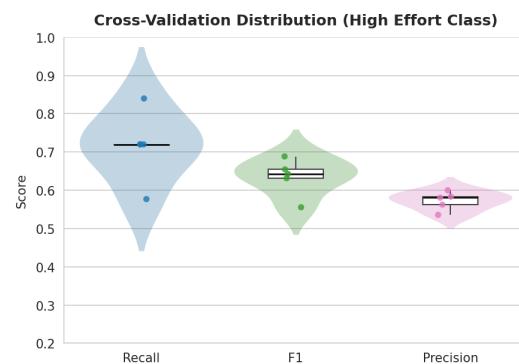


Figure 27: Confidence interval for predicting High Reasoning Effort class

```

1512 You are a top-notch scientist. Classify the reasoning effort required to
1513 solve the given scientific problem into exactly one category: Minimum
1514 , Low, Medium, or High.
1515
1516 ## Definitions
1517
1518 **Minimum:** Direct recall or single-step application. Problem is solved
1519 by retrieving and directly applying one known fact, formula, or
1520 procedure.
1521
1522 **Low:** Straightforward multi-step reasoning within a single concept.
1523 Require 2-4 logical steps using one domain of knowledge, with minimal
1524 abstraction or transformation.
1525
1526 **Medium:** Coordinated application of multiple concepts. Requires
1527 selecting appropriate methods, combining knowledge from 1-2 domains,
1528 or building intermediate representations to bridge problem and
1529 solution.
1530
1531 **High:** Complex integration across domains. Demands synthesizing
1532 concepts from 3+ domains, constructing elaborate models, navigating
1533 significant abstraction, or developing novel solution pathways.
1534 Sometimes require a large number of reasoning steps to solve the
1535 problem.
1536
1537 ## Guidelines
1538
1539 - Consider the conceptual complexity, not computational difficulty
1540 - If you consider the reasoning steps, use the cognitive steps an
1541 expert would perform, not the time required
1542 - Provide a rationale (2-3 sentences) that identifies the key
1543 reasoning operations and justifies your classification
1544
1545
1546 ## Problem Statement:
1547 {problem_statement}
1548
1549 ## Reference Solution:
1550 {reference_solution if reference_solution is not None else None }
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

```