

# Automatically Generating Hard Math Problems from Hypothesis-Driven Error Analysis

**Anonymous authors**

Paper under double-blind review

## Abstract

Numerous math benchmarks exist to evaluate LLMs’ mathematical capabilities. However, most involve extensive manual effort and are difficult to scale. Consequently, they cannot keep pace with LLM development or easily provide new instances to mitigate overfitting. Some researchers have proposed automatic benchmark generation methods, but none focus on identifying the specific math concepts and skills on which LLMs are error-prone, and most can only generate category-specific benchmarks. To address these limitations, we propose a new math benchmark generation pipeline that uses AI-generated hypotheses to identify the specific math concepts and skills that LLMs struggle with, and then generates new benchmark problems targeting these weaknesses. Experiments show that hypothesis accuracy positively correlates with the difficulty of the generated problems: problems generated from the most accurate hypotheses reduce Llama-3.3-70B-Instruct’s accuracy to as low as 45%, compared to 77% on the original MATH benchmark. Furthermore, our pipeline is highly adaptable and can be applied beyond math to explore a wide range of LLM capabilities, making it a valuable tool for investigating how LLMs perform across different domains.

## 1 Introduction

Evaluating the mathematical reasoning of LLMs relies on benchmarks, yet most existing benchmarks are manually constructed and curated. This manual process does not scale, creating two persistent problems. First, because benchmark problems are static and publicly available, LLMs frequently encounter them during training—sometimes inadvertently, sometimes by design—leading to overfitted evaluations that overestimate true capability. Efforts to generate fresh instances that resemble the original benchmark (Zhang et al., 2024a) still rely on manual construction and therefore inherit the same bottleneck. Second, LLMs are evolving rapidly, but manual benchmark creation cannot keep pace: existing benchmarks quickly become obsolete, while new benchmarks targeting emerging capabilities are slow to appear.

To overcome these limitations, recent work has explored LLM-based benchmark generation. While faster than manual construction, current automatic methods share several shortcomings: (i) they are often restricted to a single problem format such as multiple-choice or simple QA; (ii) they do not identify the specific mathematical concepts and skills on which LLMs are weakest, and therefore cannot target generated problems toward those weaknesses; and (iii) they tend to be domain-specific, covering only a narrow slice of mathematics rather than adapting across topics or beyond math entirely.

We introduce a benchmark generation pipeline that addresses all three gaps. Our pipeline uses Hypogenic (Zhou et al., 2024), an LLM-based hypothesis generator, to analyze problems that a target LLM consistently fails and produce hypotheses about the mathematical concepts and skills underlying those failures. These hypotheses then guide the generation of new problems designed to probe the identified weaknesses. The pipeline is lightweight and adaptable: by modifying the hypothesis prompt, it can investigate non-mathematical factors that affect LLM performance (e.g., problem wording or solution length) or generate benchmarks in domains outside of mathematics.

We evaluate the pipeline on the MATH benchmark (Hendrycks et al., 2021) using Llama-3.3-70B-Instruct (Meta, 2024) as the target model, testing across five levels of mathematical concept granu-

054 larity and three hypothesis-generating LLMs. Our experiments show a positive correlation between  
055 hypothesis accuracy and the difficulty of the resulting problems: problems generated from the most  
056 accurate hypotheses reduce the target model’s accuracy to as low as 45%, compared to 77% on the  
057 original MATH benchmark. We also find that granularity matters—redundant or overlapping con-  
058 cept categories degrade both hypothesis quality and downstream problem quality.

059 Our contributions are as follows:

- 061 1. A generation pipeline that identifies mathematical concepts and skills on which LLMs are  
062 weak and produces targeted benchmark problems that are substantially more challenging  
063 than general-purpose benchmarks.
- 064 2. An empirical analysis showing that the granularity of concept categorization significantly  
065 affects both hypothesis accuracy and generated problem quality, with redundant categories  
066 degrading performance.

## 069 2 Related Works

071 Most existing math benchmarks, such as GSM8k (Cobbe et al., 2021) and MATH (Hendrycks et al.,  
072 2021), require extensive manual effort to construct. Even recent benchmarks rely on substantial  
073 human involvement: Glazer et al. (2024) enlist expert mathematicians to create and verify a dataset  
074 of challenging problems, and Chernyshev et al. (2024) manually curate a university-level benchmark.  
075 Because human-in-the-loop construction is slow and difficult to scale, it struggles to keep pace with  
076 LLM progress or to refresh problems quickly enough to prevent overfitting. This has motivated a  
077 growing body of work on partially or fully automatic benchmark generation.

079 **Extending existing benchmarks.** Several methods generate new problems by transforming or  
080 mimicking existing ones. Huang et al. (2024) train an LLM-based pipeline to produce new instances  
081 that match the length, semantic embedding, and diversity of a source benchmark. O’Brien et al.  
082 (2025) modify problems with counterfactual rules to test induction, deduction, and overfitting ten-  
083 dencies. Other work focuses on increasing difficulty: Wang et al. (2024) evolve the complexity of  
084 existing benchmarks through multi-agent interaction, and Zhang et al. (2024b) extract and perturb  
085 reasoning graphs to produce higher-complexity samples.

087 **Prompt-based generation.** Rather than extending existing data, some approaches generate bench-  
088 marks directly from prompts. Yuan et al. (2025) develop BenchMaker, which automatically elab-  
089 orates a user-provided prompt into detailed multiple-choice questions. Shashidhar et al. (2025) in-  
090 troduce YourBench, a framework that generates QA or MCQ problems using an input document as  
091 reference. A limitation of these prompt-based approaches is that they require either seed data or  
092 reference documents, making them difficult to apply in domains that lack such resources.

094 **Tool-integrated generation.** A third line of work combines multiple LLM capabilities in the gen-  
095 eration process. Shah et al. (2025) extract and categorize mathematical skills from the MATH bench-  
096 mark, then prompt an LLM to generate new problems by combining these skills. Peng et al. (2025)  
097 build an automatic proof-benchmark generator for algebraic geometry that produces problems re-  
098 siliant to guessing and superficial pattern matching.

099 **Hypothesis generation with LLMs.** Our pipeline builds on Hypogenic (Zhou et al., 2024), a  
100 framework for generating and evaluating natural-language hypotheses from labeled data using  
101 LLMs. Hypogenic iteratively proposes hypotheses, scores them against held-out examples, and  
102 returns the most accurate ones. We repurpose this machinery to hypothesize which mathematical  
103 concepts and skills underlie an LLM’s failures, then use those hypotheses to guide targeted problem  
104 generation.

106 Among existing automatic generation methods, few attempt to identify the specific concepts and  
107 skills on which LLMs are most error-prone, and none investigate how the granularity of concept  
categorization affects generation quality. Our work addresses both gaps.

### 3 Generation Pipeline

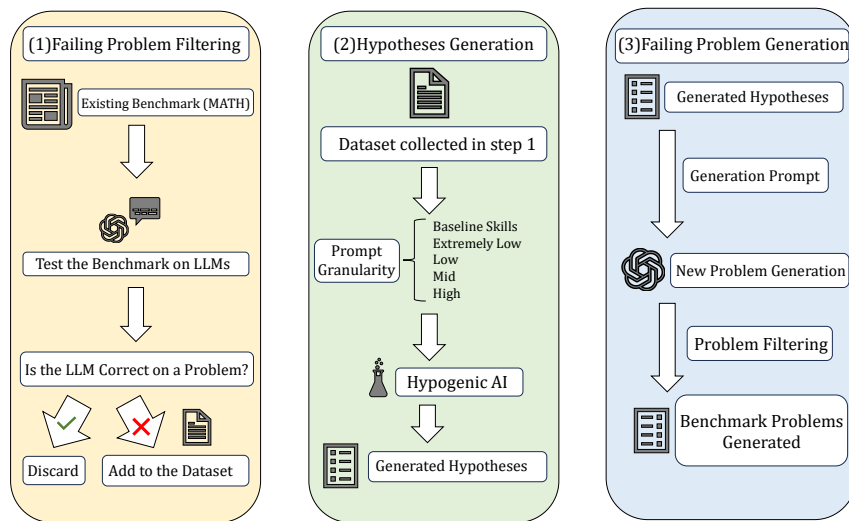


Figure 1: Overview of the three-stage generation pipeline: (1) filter problems that the target LLM consistently fails, (2) generate hypotheses about the concepts and skills underlying those failures, and (3) generate new problems guided by the hypotheses.

Our pipeline generates challenging math problems through three stages (Figure 1):

- Failing Problem Filtering:** Test the target LLM on an existing benchmark and extract problems it consistently answers incorrectly.
- Hypothesis Generation:** Use an LLM to generate hypotheses for which mathematical concepts and skills underlie the failures.
- Challenging Problem Generation:** Generate new problems that target the identified weaknesses, guided by the hypotheses from Stage 2.

#### 3.1 Failing Problem Filtering

We evaluate Llama-3.3-70B-Instruct (Meta, 2024) on the MATH benchmark using the model’s recommended decoding parameters (temperature=0.6, top\_p=0.9, top\_k=40, repetition\_penalty=1.2). Each problem is attempted five times; we retain only those that the model answers incorrectly on every attempt. Because these failures persist across all five trials, they are unlikely to result from sampling variance and more likely reflect systematic weaknesses in the model’s mathematical reasoning. The resulting set of consistently failed problems forms the input to the hypothesis generation stage.

#### 3.2 Hypothesis Generation

We run Hypogenic (Zhou et al., 2024) on the failed-problem dataset from Stage 1. Hypogenic proposes natural-language hypotheses about which mathematical concepts and skills are associated with the LLM’s failures, then scores each hypothesis by its *accuracy*: the fraction of samples in the dataset whose labels (correct/incorrect) are consistent with the hypothesis. We retain the top fifteen hypotheses per configuration for downstream evaluation (listed in Appendix C).

To investigate how the granularity of concept categorization affects hypothesis quality, we design five prompt variants (Appendix A), each providing Hypogenic with a different taxonomy of mathematical concepts and skills. Four taxonomies—*extremely low*, *low*, *mid*, and *high* granularity—were constructed by applying LLM-based extraction to the MATH benchmark at increasing levels of specificity. A fifth *baseline* taxonomy uses the full skill list extracted by Shah et al. (2025). We

also compare three LLMs as the backbone for Hypogenic: GPT-4o-mini, GPT-4.1-mini, and Qwen3-14B (with thinking disabled). Among these, GPT-4.1-mini produces the most accurate hypotheses, and the low-granularity prompt yields the best results overall. Figure 2 shows the hypothesis accuracy distributions across granularities for GPT-4.1-mini; results for the other two models appear in Appendix B.

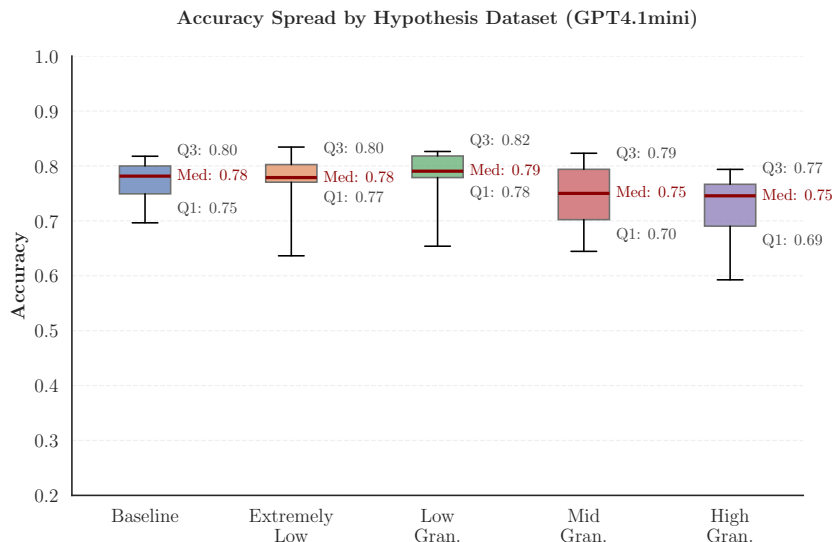


Figure 2: Hypothesis accuracy distributions across granularity levels (GPT-4.1-mini). The low-granularity prompt achieves the highest median and quartile accuracies. Accuracy increases from extremely low to low granularity, then decreases as granularity increases further.

Because the generated hypotheses capture patterns in the concepts and skills that the target LLM struggles with, they serve as natural guides for generating challenging problems. We use them as part of the generation prompt in Stage 3.

### 3.3 Challenging Problem Generation

We construct generation prompts by combining the hypotheses from Stage 2 with the mathematical concepts and skills present in the MATH benchmark (see Appendix D for full prompts). Since GPT-4.1-mini produces the most accurate hypotheses, we use its outputs for all problem generation. The generating LLM is Llama-3.3-70B-Instruct—the same model used in Stage 1—to control for model-specific variation throughout the pipeline (Hypogenic is the sole exception).

Because the generated problems target areas where LLMs are weak, the generating model itself is prone to producing flawed outputs. We therefore apply a multi-stage filtering process: we remove problems that are invalid, incorrect, incomplete, or ambiguous, and additionally exclude proof-style questions to standardize the output format and simplify evaluation. Answer keys are derived using DeepSeek-R1, GPT-o3, and GPT-5, then verified through cross-model agreement and manual validation. Sample generated problems for each granularity are shown in Appendix E.

## 4 Evaluating Problem Generation

We evaluate the quality of the generated problems by measuring how often the target LLM fails on them. For each granularity level, we select the two highest-accuracy hypotheses (from those listed in Appendix C) and generate problems following the procedure in Section 3.3.

### 4.1 Experimental Setup

Problem generation uses Llama-3.3-70B-Instruct with slightly elevated diversity settings (temperature = 1.0, top\_p = 0.9, top\_k = 50, repetition\_penalty = 1.05) to encourage variety. For each hypoth-

216 esis, we randomly sample 20 problems from the filtered pool, with answer keys derived as described  
 217 in Section 3.3.

218 To maintain consistency, we evaluate the generated problems using the same model and decoding  
 219 configuration as in Stage 1 (Llama-3.3-70B-Instruct; temperature=0.6, top\_p=0.9, top\_k=40, rep-  
 220 etition\_penalty=1.2). This avoids confounds from model-specific differences in problem-solving  
 221 ability, since we observe that larger or more capable models tend to produce substantially harder  
 222 problems under our prompting.

Granularity	Hypothesis	Llama-3.3 Solve Rate
Baseline Skills	1. The LLM is likely to fail in problems requiring calculation and conversion skills.	1. 90%
	2. The LLM shows difficulty on problems involving coordinate geometry and transformation skills together with graph understanding and interpretation.	2. 95%
Extremely Low Granularity	1. The LLM is likely to fail on problems involving the combination of Geometry and Algebra.	1. 70%
	2. The LLM is likely to fail on problems involving both Prealgebra and Algebra.	2. 90%
Low Granularity	1. The LLM is likely to fail on problems involving Modular arithmetic, divisibility, and integer properties.	1. 45%
	2. The LLM is likely to fail on problems involving spatial reasoning, geometric theorem application.	2. 60%
Mid Granularity	1. The LLM is more error-prone on problems involving function eval/composition.	1. 90%
	2. The LLM is more likely to fail on problems requiring the use of linear & systems concepts.	2. 95%
High Granularity	1. The LLM is likely to fail in problems involving function evaluation and basic transformations.	1. 55%
	2. The LLM is likely to fail on problems involving Integer arithmetic (+, -, ×, ÷).	2. 95%

252 Table 1: Selected hypotheses and Llama-3.3-70B-Instruct solve rates on problems generated from  
 253 each hypothesis, across granularity levels.

## 256 4.2 Experimental Results

257 Table 1 reports the two selected hypotheses per granularity and Llama-3.3-70B-Instruct’s solve rate  
 258 on each 20-problem set; Figure 4 visualizes these results.

259 **Hypothesis accuracy predicts problem difficulty.** Across all five granularity levels, the model  
 260 consistently performs worse on problems generated from the highest-accuracy hypothesis than on  
 261 those from the second-highest. This confirms that Hypogenic’s accuracy scores meaningfully reflect  
 262 the degree to which a hypothesis captures the model’s weaknesses: higher-accuracy hypotheses yield  
 263 harder problems.

264 **Granularity affects generation quality.** The trend in solve rates across granularities mirrors the  
 265 trend in the number of high-accuracy hypotheses (Figure 3). The low-granularity prompt produces  
 266 the most hypotheses with accuracy above 0.8, and problems generated from low-granularity hypothe-  
 267 ses are also the most challenging—reducing the model’s solve rate to as low as 45%. Under baseline  
 268 and mid granularities, even the best hypothesis produces problems on which the model scores at or  
 269

above its 77% MATH benchmark accuracy, suggesting that overly coarse categories lack specificity while overly fine ones constrain generation creativity.

**Redundant categories hurt performance.** Problems generated from baseline-skill hypotheses are the least challenging overall (Figure 4), indicating that redundant and overlapping concept categories degrade both hypothesis quality and the difficulty of the resulting problems.

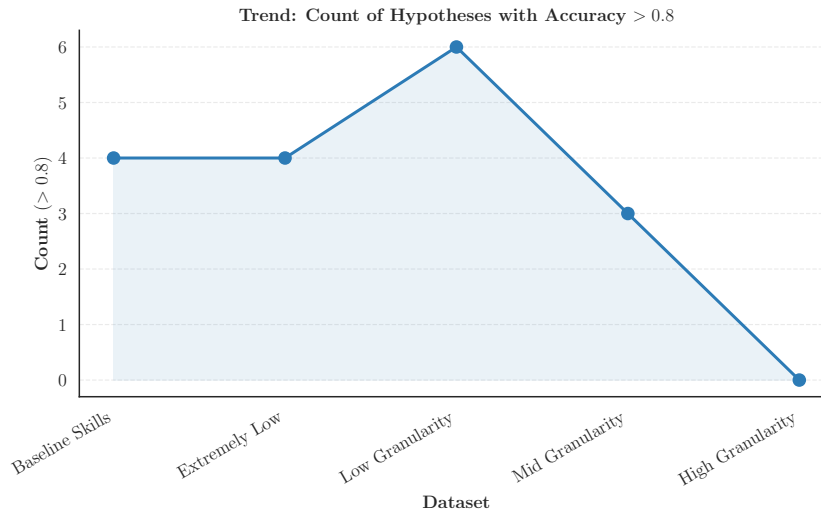


Figure 3: Trending of Number of Hypotheses with Accuracies Over 0.8 Using GPT4.1mini

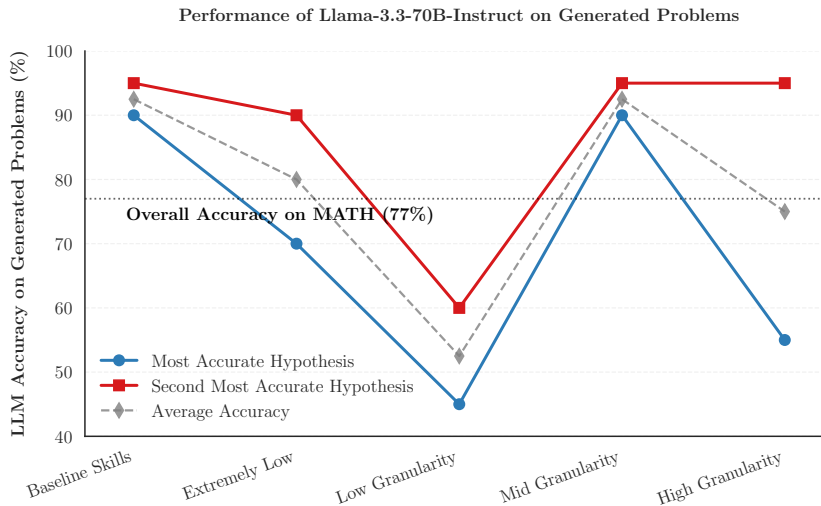


Figure 4: Llama-3.3-70B-Instruct solve rates on generated problems across granularity levels. The dashed line marks the model’s 77% solve rate on the original MATH benchmark (Meta, 2024). Solve rate decreases from extremely low to low granularity, then increases from low to high, mirroring the trend in hypothesis accuracy.

## 5 General Discussion

### 5.1 Advantages over Existing Methods

Our experiments confirm that hypothesis-guided generation produces problems that are meaningfully harder for the target LLM, and that hypothesis accuracy is a reliable predictor of problem difficulty. Compared to existing automatic generation approaches, our pipeline offers several advantages. First,

324 it explicitly identifies the concepts and skills on which an LLM is weakest, rather than generating  
325 problems uniformly or relying on surface-level difficulty heuristics. Second, it accounts for the  
326 granularity of concept categorization, a factor that, to our knowledge, no prior work has investigated,  
327 and shows that redundant or overlapping categories degrade generation quality. Third, the pipeline  
328 requires no manually curated seed set beyond an existing benchmark and involves minimal human  
329 intervention.

330 Beyond targeted benchmark construction, the generated problems can shed light on how LLMs inter-  
331 pret mathematical concepts differently from humans, potentially explaining why models sometimes  
332 solve advanced problems while failing on elementary ones. Because the pipeline’s behavior is con-  
333 trolled entirely through the Hypogenic prompt, it can be adapted to explore non-mathematical factors  
334 (such as problem wording, solution length, or the number of concepts per problem) or extended to  
335 domains outside mathematics.

## 337 5.2 Limitations

339 **Small evaluation sample.** Due to computational constraints, we evaluate only 20 problems per  
340 hypothesis. At this sample size, each individual error corresponds to a 5 percentage-point shift in  
341 solve rate, making the results sensitive to noise. Scaling to larger problem sets would yield more  
342 precise estimates and clearer trends across granularity levels.

344 **Generator capability.** Llama-3.3-70B-Instruct serves as both the target and the generator. We  
345 observe that the model produces more flawed problems when guided by high-accuracy hypotheses,  
346 precisely the areas where it is weakest. Although we filter these problems, generator limitations may  
347 still reduce the quality ceiling of the benchmark, particularly for the most targeted hypotheses.

349 **Correlation vs. causation.** The hypotheses identify statistical associations between concept labels  
350 and failures, but these associations are not necessarily causal. An LLM may fail on a problem labeled  
351 “modular arithmetic” for reasons unrelated to modular arithmetic itself.

- 353 1. *Sensitivity to wording:* the model may struggle with specific phrasings, and paraphrasing  
354 the same problem could yield a correct answer.
- 356 2. *Long-context degradation:* problems requiring extended reasoning chains may cause the  
357 model to lose coherence, independent of the mathematical content.
- 358 3. *Confounding skills:* problems nominally testing one concept often involve auxiliary skills  
359 (e.g., solving a geometry problem via systems of equations), and the true source of error  
360 may lie in the auxiliary skill rather than the labeled one.

362 Addressing this limitation is a natural extension of the pipeline: by modifying the Hypogenic prompt  
363 to generate hypotheses about non-content factors (e.g., wording complexity, solution length), the  
364 same framework can disentangle concept-level weaknesses from other sources of failure.

## 367 6 Conclusion

369 We presented an automatic math benchmark generation pipeline that uses LLM-powered hypothesis  
370 generation to identify mathematical concepts and skills on which a target LLM is weakest, then  
371 generates problems that specifically target those weaknesses. Our experiments show that hypothesis  
372 accuracy correlates with the difficulty of the resulting problems, with the low-granularity prompt  
373 producing the most accurate hypotheses and the hardest generated benchmarks. The pipeline requires  
374 only an existing benchmark as input and minimal human oversight, and can be adapted to investigate  
375 non-mathematical failure factors or extended to other domains by modifying the hypothesis prompt.  
376 In future work, we plan to scale evaluation to larger problem sets, test additional target models,  
377 and explore prompts that disentangle concept-level weaknesses from confounding factors such as  
problem wording and solution length.

## References

- 378  
379  
380 Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov,  
381 Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating math-  
382 ematical skills in llms. *arXiv preprint arXiv:2412.03205*, 2024.
- 383 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
384 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John  
385 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,  
386 2021.
- 387 Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caro-  
388 line Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Fron-  
389 tiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint*  
390 *arXiv:2411.04872*, 2024.
- 391 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,  
392 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv*  
393 *preprint arXiv:2103.03874*, 2021.
- 394 Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei  
395 Xiao, Jianfeng Gao, Lichao Sun, et al. Datagen: Unified synthetic dataset generation via large  
396 language models. In *The Thirteenth International Conference on Learning Representations*, 2024.
- 397  
398 Meta. Llama-3.3-70B-Instruct. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>, Dec 2024.
- 399  
400 Dayyán O’Brien, Barry Haddow, Emily Allaway, and Pinzhen Chen. Mathemagic: Generating dy-  
401 namic mathematics benchmarks robust to memorization. *arXiv preprint arXiv:2510.05962*, 2025.
- 402  
403 Yebo Peng, Zixiang Liu, Yaoming Li, Zhizhuo Yang, Xinye Xu, Bowen Ye, Weijun Yuan, Zihan  
404 Wang, and Tong Yang. Proof2hybrid: Automatic mathematical benchmark synthesis for proof-  
405 centric problems. *arXiv preprint arXiv:2508.02208*, 2025.
- 406  
407 Vedant Shah, Dingli Yu, Kaifeng Lyu, Simon Park, Jiatong Yu, Yinghui He, Nan Rosemary Ke,  
408 Michael Mozer, Yoshua Bengio, Sanjeev Arora, and Anirudh Goyal. Ai-assisted generation of  
409 difficult math questions, 2025. URL <https://arxiv.org/abs/2407.21009>.
- 410  
411 Sumuk Shashidhar, Clémentine Fourier, Alina Lozovskia, Thomas Wolf, Gokhan Tur, and  
412 Dilek Hakkani-Tür. Yourbench: Easy custom evaluation sets for everyone. *arXiv preprint*  
413 *arXiv:2504.01833*, 2025.
- 414  
415 Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. Benchmark self-  
416 evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*,  
417 2024.
- 418  
419 Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan  
420 Pan, Yao Hu, and Kan Li. Llm-powered benchmark factory: Reliable, generic, and efficient. *arXiv*  
421 *preprint arXiv:2502.01683*, 2025.
- 422  
423 Di Zhang, Jiatong Li, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing  
424 gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv*  
425 *preprint arXiv:2406.07394*, 2024a.
- 426  
427 Zhehao Zhang, Jiao Chen, and Diyi Yang. Darg: Dynamic evaluation of large language models  
428 via adaptive reasoning graph. *Advances in Neural Information Processing Systems*, 37:135904–  
429 135942, 2024b.
- 430  
431 Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis  
432 generation with large language models. In *Proceedings of EMNLP Workshop of NLP for Science*,  
433 2024. URL <https://aclanthology.org/2024.nlp4science-1.10/>.

## 432 A Hypogenic Prompts

### 433 A.1 Baseline Skill Prompt

```

436 prompt_templates:
437   observations: |
438     The math problem: ${problems}
439     The LLM's answer to the problem: ${answers}
440     The correctness of the answer: ${label}.
441
442   batched_generation:
443     system: |-
444       You are a professional math teacher and educational researcher.
445       Given a set of math problems and LLM-generated answers, we want to
446       ↪ generate hypotheses that predicts LLMs are more error-prone on
447       ↪ which particular math skills/concepts (or combinations of math
448       ↪ skills/concepts)
449       In other words, we want to understand what kinds of problem are
450       ↪ associated with correct or wrong labels, what kind of problem
451       ↪ makes LLM more likely to fail,
452       what kind of math concept or combination of math concepts make LLM
453       ↪ likely to fail the problem. Here is a form of math concepts
454       ↪ involved in the problems:
455       "absolute_value_skills",
456       "algebra_and_equations",
457       "algebraic_expression_skills",
458       "algebraic_manipulation_and_equations",
459       "algebraic_manipulation_skills",
460       "algebraic_skills",
461       "area_calculation_skills",
462       "arithmetic_operations",
463       "arithmetic_sequences",
464       "arithmetic_skills",
465       "average_calculations",
466       "base_conversion",
467       "basic_arithmetic",
468       "basic_arithmetic_operations",
469       "basic_trigonometry",
470       "calculating_and_understanding_combinations",
471       "calculation_and_conversion_skills",
472       "calculus",
473       "calculus_skills",
474       "circle_geometry_skills",
475       "circles",
476       "combinatorial_mathematics",
477       "combinatorial_operations_and_basic_arithmetic",
478       "combinatorics_and_probability_skills",
479       "combinatorics_knowledge",
480       "complex_number_manipulation_and_operations",
481       "complex_number_operations",
482       "complex_number_skills",
483       "complex_numbers",
484       "complex_numbers_related_skills",
485       "coordinate_geometry_and_transformation_skills",
486       "coordinate_systems",
487       "counting_and_number_theory",
488       "counting_principals",
489       "distance_and_midpoint_skills",
490       "division_and_remainders",

```

486 "exponent\_and\_root\_skills",  
487 "exponentials\_and\_logarithms",  
488 "exponentiation",  
489 "exponentiation\_rules",  
490 "factorials\_and\_prime\_factorization",  
491 "factoring\_skills",  
492 "factorization",  
493 "fractions\_and\_decimals",  
494 "function\_composition\_and\_transformation",  
495 "function\_composition\_skills",  
496 "function\_skills",  
497 "geometric\_relations",  
498 "geometric\_sequence\_skills",  
499 "geometric\_series\_comprehension",  
500 "geometry",  
501 "geometry\_and\_space\_calculation",  
502 "geometry\_triangle\_properties",  
503 "graph\_and\_geometry\_skills",  
504 "graph\_understanding\_and\_interpretation",  
505 "greatest\_common\_divisor\_calculations",  
506 "inequality\_skills",  
507 "inequality\_solving\_and\_understanding",  
508 "logarithmic\_and\_exponential\_skills",  
509 "matrix\_operations",  
510 "modular\_arithmetic",  
511 "multiplication\_and\_division",  
512 "number\_manipulation",  
513 "number\_theory",  
514 "number\_theory\_and\_arithmetic\_operations",  
515 "number\_theory\_skills",  
516 "other\_geometric\_skills",  
517 "parametric\_equations",  
518 "perimeter\_and\_area",  
519 "permutation\_and\_combinations",  
520 "polynomial\_skills",  
521 "prime\_number\_recognition\_and\_properties",  
522 "prime\_number\_theory",  
523 "probability\_and\_combinatorics",  
524 "probability\_concepts\_and\_calculations",  
525 "properties\_and\_application\_of\_exponents",  
526 "pythagorean\_skills",  
527 "quadratic\_equation\_skills",  
528 "quadratic\_equations\_and\_solutions",  
529 "ratio\_and\_proportion",  
530 "ratio\_and\_proportion\_skills",  
531 "recurrence",  
532 "recursive\_functions\_and\_sequences",  
533 "sequence\_analysis",  
534 "sequence\_and\_series\_analysis\_skills",  
535 "sequence\_and\_series\_skills",  
536 "sequences",  
537 "sequences\_series\_and\_summation",  
538 "simplification\_and\_basic\_operations",  
539 "solving\_equations",  
"solving\_inequalities",  
"solving\_linear\_equation",  
"solving\_system\_of\_equations",  
"summation\_and\_analysis\_of\_series",  
"three\_dimensional\_geometry",

540 "triangle\_geometry\_skills",  
541 "trigonometry\_skills",  
542 "understanding\_and\_application\_of\_functions",  
543 "understanding\_and\_applying\_combinatorics\_concepts",  
544 "understanding\_and\_applying\_floor\_and\_ceiling\_functions",  
545 "understanding\_and\_manipulation\_of\_rational\_functions",  
546 "understanding\_and\_utilizing\_infinite\_series",  
547 "understanding\_circle\_properties\_and\_algebraic\_manipulation",  
548 "understanding\_ellipse\_properties",  
549 "understanding\_logarithmic\_properties\_and\_solving\_equations",  
550 "understanding\_of\_fractions",  
551 "vector\_operations"

552 Using **!!!ONLY!!!** the given math concepts in the form, **!!!DO NOT!!!**  
553 ↪ include any guessings or conditions other than the concepts in  
554 ↪ the form!!!

555 **!!!Do NOT!!!** propose conditions or special particularities. For  
556 ↪ example, a hypothesis like "The LLM is likely to fail on a  
557 ↪ concept, especially/particularly when the problem is...(some  
558 ↪ special conditions)" or "The LLM is likely to fail on a concept  
559 ↪ because of (some condition)" is not allowed!

560 For example: "The LLM is likely to fail on problems that require  
561 ↪ function evaluation and transformations, especially when dealing  
562 ↪ with composite functions and inverses." is a BAD hypothesis that  
563 ↪ doesn't follow the previous requirement.

564 Another example: "The LLM is likely to fail on (some math concept)  
565 ↪ when the problem is complex" is also a BAD hypothesis that  
566 ↪ doesn't follow the instructions. 'complex' is very vague and is a  
567 ↪ condition that is NOT in the math concept. It will NOT be  
568 ↪ accepted.

568 **!!!Do NOT!!!** propose hypotheses that are based on a particular step  
569 ↪ in the problem! For example, a hypothesis like "The LLM is likely  
570 ↪ to fail on a concept at a (some step in problem solution)" is not  
571 ↪ allowed!

572 **!!!Do NOT!!!** propose any reasons behind the failures! For example, a  
573 ↪ hypothesis like "The LLM is likely to fail on a concept,  
574 ↪ because/due to ... (some reason)" is not allowed!

574 Again, Use **!!!Only!!!** the given math concepts in the form! Only means  
575 ↪ the hypotheses can only contain the concepts in the form! and no  
576 ↪ other things allowed!

577 Using anything other than the math concepts in the form will NOT be  
578 ↪ accepted and should NOT be proposed!

579 These hypotheses should identify specific patterns that occur across  
580 ↪ the provided problems and LLM-generated answers.  
581 please propose  $\{\text{num\_hypotheses}\}$  possible hypothesis pairs.

582 These hypotheses should identify specific patterns that occur across  
583 ↪ the provided problems and LLM-generated answers.

584 You should check carefully the specific solving steps by the LLM, and  
585 ↪ consider which particular  
586 ↪ step and which particular math concept/skill did the LLM make mistake  
587 ↪ on.

587 When proposing hypotheses, generate half hypotheses using a single  
588 ↪ math concept, and generate the other half by combining two or  
589 ↪ more math concepts.

590 Again, use **ONLY** and **VERBATIMLY** the provided math concepts from the  
591 ↪ list

592  
593 Each hypothesis should contain the following:  
a. A hypothesis about what particular math makes the LLM to fail

```

594
595     Generate them in the format of 1. [hypothesis], 2. [hypothesis], ...
596     ↪ ${num_hypotheses}. [hypothesis].
597     The hypotheses should analyze what particular math concept(s) are
598     ↪ associated with correctness or error.
599
600     user: |-
601     We have seen some math problems and LLM-generated answers:
602     ${observations}
603     Please generate hypotheses that are useful for predicting which
604     ↪ particular math concept and solution step does the LLM likely to
605     ↪ make mistakes on.
606     Propose ${num_hypotheses} possible hypotheses. Generate them in the
607     ↪ format of 1. [hypothesis], 2. [hypothesis], ...
608     ↪ ${num_hypotheses}. [hypothesis].
609     Proposed hypotheses:
610
611     inference:
612     system: |-
613     You are a professional math teacher and your job is to determine
614     ↪ whether a given answer to a math problem is correct or wrong.
615     From past experience, you have learned that LLMs are more likely to
616     ↪ fail on certain math concepts (or combination of math concepts).
617     You need to determine whether the learned pattern applies to the
618     ↪ current problem and answer, and then make your prediction.
619     Give your final answer in the format of "Final answer: answer", where
620     ↪ the answer is either "correct" or "wrong".
621
622     user: |-
623     Our learned pattern: ${hypothesis}
624     A math problem and its answer are the following:
625     Problem: "${problems}"
626     Answer: "${answers}"
627
628     Given the pattern you learned above, decide whether the answer is
629     ↪ correct or wrong.
630     Think step by step.
631     First step: Consider if the pattern can be applied to the answer.
632     Second step: Based on the pattern, is this answer correct or wrong?
633     Final step: give your final answer in the format of "Final answer:
634     ↪ answer"
635
636     multiple_hypotheses_inference:
637     system: |-
638     You are a professional math teacher and your job is to determine
639     ↪ whether a given answer to a math problem is correct or wrong.
640     From past experience, you have learned that LLMs are more likely to
641     ↪ fail on certain math concepts (or combination of math concepts).
642     You need to determine whether these patterns apply to the current
643     ↪ problem and answer, and then make your prediction.
644     Give your final answer in the format of "Final answer: answer", where
645     ↪ the answer is either "correct" or "wrong".
646
647     user: |-
648     Our learned patterns: ${hypotheses}
649     A math problem and its answer are the following:
650     Problem: "${problems}"
651     Answer: "${answers}"

```

648  
 649       Given the patterns you learned above, decide whether the answer is  
 650       ↪ correct or wrong.  
 651       Think step by step.  
 652       First step: Think about which pattern(s) can be applied to the  
 653       ↪ answer.  
 654       Second step: Based on the patterns, is this answer correct or wrong?  
 655       Final step: give your final answer in the format of "Final answer:  
 656       ↪ answer"

## 660 A.2 Extremely Low Granularity Prompt

661  
 662 **prompt\_templates:**  
 663   **observations:** |  
 664       The math problem:  $\{\text{problems}\}$   
 665       The LLM's answer to the problem:  $\{\text{answers}\}$   
 666       The correctness of the answer:  $\{\text{label}\}$ .  
 667  
 668   **batched\_generation:**  
 669   **system:** |-  
 670       You are a professional math teacher and educational researcher.  
 671       Given a set of math problems and LLM-generated answers, we want to  
 672       ↪ generate hypotheses that predicts LLMs are more error-prone on  
 673       ↪ which particular math skills/concepts (or combinations of math  
 674       ↪ skills/concepts)  
 675       In other words, we want to understand what kinds of problem are  
 676       ↪ associated with correct or wrong labels, what kind of problem  
 677       ↪ makes LLM more likely to fail,  
 678       what kind of math concept or combination of math concepts make LLM  
 679       ↪ likely to fail the problem. Here is a form of math concepts  
 680       ↪ involved in the problems:  
 681       1. Prealgebra  
 682       2. Algebra  
 683       3. Number Theory  
 684       4. Counting & Probability  
 685       5. Geometry  
 686       6. Precalculus  
 687       7. Advanced Auxiliary Topics  
 688  
 689       Using **!!!ONLY!!!** the given math concepts in the form, **!!!DO NOT!!!**  
 690       ↪ include any guessings or conditions other than the concepts in  
 691       ↪ the form!!!  
 692       **!!!Do NOT!!!** propose conditions or special particularities. For  
 693       ↪ example, a hypothesis like "The LLM is likely to fail on a  
 694       ↪ concept, especially/particularly when the problem is...(some  
 695       ↪ special conditions)" or "The LLM is likely to fail on a concept  
 696       ↪ because of (some condition)" is not allowed!  
 697       For example: "The LLM is likely to fail on problems that require  
 698       ↪ function evaluation and transformations, especially when dealing  
 699       ↪ with composite functions and inverses." is a BAD hypothesis that  
 700       ↪ doesn't follow the previous requirement.  
 701       Another example: "The LLM is likely to fail on (some math concept)  
 702       ↪ when the problem is complex" is also a BAD hypothesis that  
 703       ↪ doesn't follow the instructions. 'complex' is very vague and is a  
 704       ↪ condition that is NOT in the math concept. It will NOT be  
 705       ↪ accepted.

702       !!!Do NOT!!! propose hypotheses that are based on a particular step  
703       ↪ in the problem! For example, a hypothesis like "The LLM is likely  
704       ↪ to fail on a concept at a (some step in problem solution)" is not  
705       ↪ allowed!  
706       !!!Do NOT!!! propose any reasons behind the failures! For example, a  
707       ↪ hypothesis like "The LLM is likely to fail on a concept,  
708       ↪ because/due to ... (some reason)" is not allowed!  
709       Again, Use !!!Only!!! the given math concepts in the form! Only means  
710       ↪ the hypotheses can only contain the concepts in the form! and no  
711       ↪ other things allowed!  
712       Using anything other than the math concepts in the form will NOT be  
713       ↪ accepted and should NOT be proposed!  
714       These hypotheses should identify specific patterns that occur across  
715       ↪ the provided problems and LLM-generated answers.  
716       please propose  $\{\text{num\_hypotheses}\}$  possible hypothesis pairs.  
717       These hypotheses should identify specific patterns that occur across  
718       ↪ the provided problems and LLM-generated answers.  
719       You should check carefully the specific solving steps by the LLM, and  
720       ↪ consider which particular  
721       step and which particular math concept/skill did the LLM make mistake  
722       ↪ on.  
723       When proposing hypotheses, generate half hypotheses using a single  
724       ↪ math concept, and generate the other half by combining two or  
725       ↪ more math concepts.  
726       Again, use ONLY and VERBATIMLY the provided math concepts from the  
727       ↪ list

727       Each hypothesis should contain the following:  
728       a. A hypothesis about what particular math makes the LLM to fail,  
729       ↪ where this math is one of the seven math words from the list

730       Generate them in the format of 1. [hypothesis], 2. [hypothesis], ...  
731       ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
732       The hypotheses should analyze what particular math concept(s) are  
733       ↪ associated with correctness or error.

734       

735       user: |-  
736       We have seen some math problems and LLM-generated answers:  
737        $\{\text{observations}\}$   
738       Please generate hypotheses that are useful for predicting which  
739       ↪ particular math concept and solution step does the LLM likely to  
740       ↪ make mistakes on.  
741       Propose  $\{\text{num\_hypotheses}\}$  possible hypotheses. Generate them in the  
742       ↪ format of 1. [hypothesis], 2. [hypothesis], ...  
743       ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
744       Proposed hypotheses:

745       inference:  
746       system: |-  
747       You are a professional math teacher and your job is to determine  
748       ↪ whether a given answer to a math problem is correct or wrong.  
749       From past experience, you have learned that LLMs are more likely to  
750       ↪ fail on certain math concepts (or combination of math concepts).  
751       You need to determine whether the learned pattern applies to the  
752       ↪ current problem and answer, and then make your prediction.  
753       Give your final answer in the format of "Final answer: answer", where  
754       ↪ the answer is either "correct" or "wrong".

755       user: |-

```

756     Our learned pattern: ${hypothesis}
757     A math problem and its answer are the following:
758     Problem: "${problems}"
759     Answer: "${answers}"
760
761     Given the pattern you learned above, decide whether the answer is
762     ↪ correct or wrong.
763     Think step by step.
764     First step: Consider if the pattern can be applied to the answer.
765     Second step: Based on the pattern, is this answer correct or wrong?
766     Final step: give your final answer in the format of "Final answer:
767     ↪ answer"
768
769 multiple_hypotheses_inference:
770 system: |-
771     You are a professional math teacher and your job is to determine
772     ↪ whether a given answer to a math problem is correct or wrong.
773     From past experience, you have learned that LLMs are more likely to
774     ↪ fail on certain math concepts (or combination of math concepts).
775     You need to determine whether these patterns apply to the current
776     ↪ problem and answer, and then make your prediction.
777     Give your final answer in the format of "Final answer: answer", where
778     ↪ the answer is either "correct" or "wrong".
779
780 user: |-
781     Our learned patterns: ${hypotheses}
782     A math problem and its answer are the following:
783     Problem: "${problems}"
784     Answer: "${answers}"
785
786     Given the patterns you learned above, decide whether the answer is
787     ↪ correct or wrong.
788     Think step by step.
789     First step: Think about which pattern(s) can be applied to the
790     ↪ answer.
791     Second step: Based on the patterns, is this answer correct or wrong?
792     Final step: give your final answer in the format of "Final answer:
793     ↪ answer"
794
795 A.3 Low Granularity Prompt
796
797 prompt_templates:
798 observations: |
799     The math problem: ${problems}
800     The LLM's answer to the problem: ${answers}
801     The correctness of the answer: ${label}.
802
803 batched_generation:
804 system: |-
805     You are a professional math teacher and educational researcher.
806     Given a set of math problems and LLM-generated answers, we want to
807     ↪ generate hypotheses that predicts LLMs are more error-prone on
808     ↪ which particular math skills/concepts (or combinations of math
809     ↪ skills/concepts)
810     In other words, we want to understand what kinds of problem are
811     ↪ associated with correct or wrong labels, what kind of problem
812     ↪ makes LLM more likely to fail,

```

810 what kind of math concept or combination of math concepts make LLM  
811 ↪ likely to fail the problem. Here is a form of math concepts  
812 ↪ involved in the problems:

813   Math Concepts/Skills	
814   -----	
815   Expression manipulation, equation solving	
816   Modular arithmetic, divisibility, integer properties	
817   Combinatorics, probability modeling	
818   Spatial reasoning, theorem application	
819   Sequences, function analysis	
820   Multi-step reasoning, deduction, diagram use	
821   parabolas, ellipses, hyperbolas, GCD	

822 Using **!!!ONLY!!!** the given math concepts in the form, **!!!DO NOT!!!**  
823 ↪ include any guessings or conditions other than the concepts in  
824 ↪ the form!!!

825 **!!!Do NOT!!!** propose conditions or special particularities. For  
826 ↪ example, a hypothesis like "The LLM is likely to fail on a  
827 ↪ concept, especially/particularly when the problem is...(some  
828 ↪ special conditions)" or "The LLM is likely to fail on a concept  
829 ↪ because of (some condition)" is not allowed!

830 For example: "The LLM is likely to fail on problems that require  
831 ↪ function evaluation and transformations, especially when dealing  
832 ↪ with composite functions and inverses." is a BAD hypothesis that  
833 ↪ doesn't follow the previous requirement.

834 Another example: "The LLM is likely to fail on (some math concept)  
835 ↪ when the problem is complex" is also a BAD hypothesis that  
836 ↪ doesn't follow the instructions. 'complex' is very vague and is a  
837 ↪ condition that is NOT in the math concept. It will NOT be  
838 ↪ accepted.

838 **!!!Do NOT!!!** propose hypotheses that are based on a particular step  
839 ↪ in the problem! For example, a hypothesis like "The LLM is likely  
840 ↪ to fail on a concept at a (some step in problem solution)" is not  
841 ↪ allowed!

842 **!!!Do NOT!!!** propose any reasons behind the failures! For example, a  
843 ↪ hypothesis like "The LLM is likely to fail on a concept,  
844 ↪ because/due to ... (some reason)" is not allowed!

845 Again, Use **!!!Only!!!** the given math concepts in the form! Only means  
846 ↪ the hypotheses can only contain the concepts in the form! and no  
847 ↪ other things allowed!

848 Using anything other than the math concepts in the form will NOT be  
849 ↪ accepted and should NOT be proposed!

850 These hypotheses should identify specific patterns that occur across  
851 ↪ the provided problems and LLM-generated answers.  
852 please propose  $\{\text{num\_hypotheses}\}$  possible hypothesis pairs.  
853 These hypotheses should identify specific patterns that occur across  
854 ↪ the provided problems and LLM-generated answers.

855 You should check carefully the specific solving steps by the LLM, and  
856 ↪ consider which particular  
857 step and which particular math concept/skill did the LLM make mistake  
858 ↪ on.

859 When proposing hypotheses, generate half hypotheses using a single  
860 ↪ math concept, and generate the other half by combining two or  
861 ↪ more math concepts.

862 Again, use **ONLY** and **VERBATIMLY** the provided math concepts from the  
863 ↪ list

864 Each hypothesis should contain the following:

- 865 a. A hypothesis about what particular math makes the LLM to fail

864  
865       Generate them in the format of 1. [hypothesis], 2. [hypothesis], ...  
866       ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
867       The hypotheses should analyze what particular math concept(s) are  
868       ↪ associated with correctness or error.  
869

870 **user:** |-  
871       We have seen some math problems and LLM-generated answers:  
872        $\{\text{observations}\}$   
873       Please generate hypotheses that are useful for predicting which  
874       ↪ particular math concept and solution step does the LLM likely to  
875       ↪ make mistakes on.  
876       Propose  $\{\text{num\_hypotheses}\}$  possible hypotheses. Generate them in the  
877       ↪ format of 1. [hypothesis], 2. [hypothesis], ...  
878       ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
879       Proposed hypotheses:

880 **inference:**  
881 **system:** |-  
882       You are a professional math teacher and your job is to determine  
883       ↪ whether a given answer to a math problem is correct or wrong.  
884       From past experience, you have learned that LLMs are more likely to  
885       ↪ fail on certain math concepts (or combination of math concepts).  
886       You need to determine whether the learned pattern applies to the  
887       ↪ current problem and answer, and then make your prediction.  
888       Give your final answer in the format of "Final answer: answer", where  
889       ↪ the answer is either "correct" or "wrong".

890 **user:** |-  
891       Our learned pattern:  $\{\text{hypothesis}\}$   
892       A math problem and its answer are the following:  
893       Problem: " $\{\text{problems}\}$ "  
894       Answer: " $\{\text{answers}\}$ "  
895  
896       Given the pattern you learned above, decide whether the answer is  
897       ↪ correct or wrong.  
898       Think step by step.  
899       First step: Consider if the pattern can be applied to the answer.  
900       Second step: Based on the pattern, is this answer correct or wrong?  
901       Final step: give your final answer in the format of "Final answer:  
902       ↪ answer"

903  
904 **multiple\_hypotheses\_inference:**  
905 **system:** |-  
906       You are a professional math teacher and your job is to determine  
907       ↪ whether a given answer to a math problem is correct or wrong.  
908       From past experience, you have learned that LLMs are more likely to  
909       ↪ fail on certain math concepts (or combination of math concepts).  
910       You need to determine whether these patterns apply to the current  
911       ↪ problem and answer, and then make your prediction.  
912       Give your final answer in the format of "Final answer: answer", where  
913       ↪ the answer is either "correct" or "wrong".

914 **user:** |-  
915       Our learned patterns:  $\{\text{hypotheses}\}$   
916       A math problem and its answer are the following:  
917       Problem: " $\{\text{problems}\}$ "  
918       Answer: " $\{\text{answers}\}$ "

918  
 919 Given the patterns you learned above, decide whether the answer is  
 920 ↪ correct or wrong.  
 921 Think step by step.  
 922 First step: Think about which pattern(s) can be applied to the  
 923 ↪ answer.  
 924 Second step: Based on the patterns, is this answer correct or wrong?  
 925 Final step: give your final answer in the format of "Final answer:  
 926 ↪ answer"

#### 929 A.4 Mid Granularity Prompt

930  
 931 **prompt\_templates:**  
 932 **observations:** |  
 933 The math problem:  $\{\text{problems}\}$   
 934 The LLM's answer to the problem:  $\{\text{answers}\}$   
 935 The correctness of the answer:  $\{\text{label}\}$ .  
 936  
 937 **batched\_generation:**  
 938 **system:** |-  
 939 You are a professional math teacher and educational researcher.  
 940 Given a set of math problems and LLM-generated answers, we want to  
 941 ↪ generate hypotheses that predicts LLMs are more error-prone on  
 942 ↪ which particular math skills/concepts (or combinations of math  
 943 ↪ skills/concepts)  
 944 In other words, we want to understand what kinds of problem are  
 945 ↪ associated with correct or wrong labels, what kind of problem  
 946 ↪ makes LLM more likely to fail,  
 947 what kind of math concept or combination of math concepts make LLM  
 948 ↪ likely to fail the problem. Here is a form of math concepts  
 949 ↪ involved in the problems:  
 950  
 951 Math Concepts/Skills  
 952 -Addition/subtraction/multiplication/division (fractions, decimals),  
 953 ↪ PEMDAS, even/odd, factors, multiples, GCD, LCM, absolute value,  
 954 ↪ integer exponents, roots  
 955 -Linear & systems, inequalities, polynomial ops, factoring  
 956 ↪ quadratics, rational expressions, exponents & radicals, absolute  
 957 ↪ value eqs/ineqs, function eval/composition, matrix inverse  
 958 -Prime factorization, divisibility, GCD/LCM, modular arithmetic,  
 959 ↪ Euler's  $\phi$ , Chinese remainder, parity  
 960 -Permutations, combinations, binomial expansions,  
 961 ↪ inclusion-exclusion, basic probability types, enumeration  
 962 -Angles, triangle theorems (Pythagorean, similarity, congruence),  
 963 ↪ inradius, polygon angles, circle theorems, area/volume (2D/3D),  
 964 ↪ coordinate formulas, polyhedron metrics  
 965 -Polynomial division, factor theorem, rational functions/asymptotes,  
 966 ↪ nonlinear systems, inequalities  
 967 -Sequences & series, exponential/logarithmic equations, basic trig  
 968 ↪ identities/solutions, function inversion/transformation  
 969 -Conic sections, polynomial GCDs, De-Moivre, calculus (integration,  
 970 ↪ arclength, gradients, divergence, curl, Jacobian, Laplacian),  
 971 ↪ linear algebra (eigenvalues, RREF)  
 972  
 973 Using !!!ONLY!!! the given math concepts in the form, !!!DO NOT!!!  
 974 ↪ include any guessings or conditions other than the concepts in  
 975 ↪ the form!!!

972           !!!Do NOT!!! propose conditions or special particularities. For  
973           ↪ example, a hypothesis like "The LLM is likely to fail on a  
974           ↪ concept, especially/particularly when the problem is...(some  
975           ↪ special conditions)" or "The LLM is likely to fail on a concept  
976           ↪ because of (some condition)" is not allowed!  
977           For example: "The LLM is likely to fail on problems that require  
978           ↪ function evaluation and transformations, especially when dealing  
979           ↪ with composite functions and inverses." is a BAD hypothesis that  
980           ↪ doesn't follow the previous requirement.  
981           Another example: "The LLM is likely to fail on (some math concept)  
982           ↪ when the problem is complex" is also a BAD hypothesis that  
983           ↪ doesn't follow the instructions. 'complex' is very vague and is a  
984           ↪ condition that is NOT in the math concept. It will NOT be  
985           ↪ accepted.  
986           !!!Do NOT!!! propose hypotheses that are based on a particular step  
987           ↪ in the problem! For example, a hypothesis like "The LLM is likely  
988           ↪ to fail on a concept at a (some step in problem solution)" is not  
989           ↪ allowed!  
990           !!!Do NOT!!! propose any reasons behind the failures! For example, a  
991           ↪ hypothesis like "The LLM is likely to fail on a concept,  
992           ↪ because/due to ... (some reason)" is not allowed!  
993           Again, Use !!!Only!!! the given math concepts in the form! Only means  
994           ↪ the hypotheses can only contain the concepts in the form! and no  
995           ↪ other things allowed!  
996           Using anything other than the math concepts in the form will NOT be  
997           ↪ accepted and should NOT be proposed!  
998           These hypotheses should identify specific patterns that occur across  
999           ↪ the provided problems and LLM-generated answers.  
1000           please propose  $\{\text{num\_hypotheses}\}$  possible hypothesis pairs.  
1001           These hypotheses should identify specific patterns that occur across  
1002           ↪ the provided problems and LLM-generated answers.  
1003           You should check carefully the specific solving steps by the LLM, and  
1004           ↪ consider which particular  
1005           step and which particular math concept/skill did the LLM make mistake  
1006           ↪ on.  
1007           When proposing hypotheses, generate half hypotheses using a single  
1008           ↪ math concept, and generate the other half by combining two or  
1009           ↪ more math concepts.  
1010           Again, use ONLY and VERBATIMLY the provided math concepts from the  
1011           ↪ list

1012           Each hypothesis should contain the following:  
1013           a. A hypothesis about what particular math makes the LLM to fail

1014           Generate them in the format of 1. [hypothesis], 2. [hypothesis], ...  
1015           ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
1016           The hypotheses should analyze what particular math concept(s) are  
1017           ↪ associated with correctness or error.

1018           **user:** |-  
1019           We have seen some math problems and LLM-generated answers:  
1020            $\{\text{observations}\}$   
1021           Please generate hypotheses that are useful for predicting which  
1022           ↪ particular math concept and solution step does the LLM likely to  
1023           ↪ make mistakes on.  
1024           Propose  $\{\text{num\_hypotheses}\}$  possible hypotheses. Generate them in the  
1025           ↪ format of 1. [hypothesis], 2. [hypothesis], ...  
1026           ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
1027           Proposed hypotheses:

```

1026
1027 inference:
1028   system: |-
1029     You are a professional math teacher and your job is to determine
1030     ↪ whether a given answer to a math problem is correct or wrong.
1031     From past experience, you have learned that LLMs are more likely to
1032     ↪ fail on certain math concepts (or combination of math concepts).
1033     You need to determine whether the learned pattern applies to the
1034     ↪ current problem and answer, and then make your prediction.
1035     Give your final answer in the format of "Final answer: answer", where
1036     ↪ the answer is either "correct" or "wrong".
1037
1038   user: |-
1039     Our learned pattern: ${hypothesis}
1040     A math problem and its answer are the following:
1041     Problem: "${problems}"
1042     Answer: "${answers}"
1043
1044     Given the pattern you learned above, decide whether the answer is
1045     ↪ correct or wrong.
1046     Think step by step.
1047     First step: Consider if the pattern can be applied to the answer.
1048     Second step: Based on the pattern, is this answer correct or wrong?
1049     Final step: give your final answer in the format of "Final answer:
1050     ↪ answer"
1051
1052 multiple_hypotheses_inference:
1053   system: |-
1054     You are a professional math teacher and your job is to determine
1055     ↪ whether a given answer to a math problem is correct or wrong.
1056     From past experience, you have learned that LLMs are more likely to
1057     ↪ fail on certain math concepts (or combination of math concepts).
1058     You need to determine whether these patterns apply to the current
1059     ↪ problem and answer, and then make your prediction.
1060     Give your final answer in the format of "Final answer: answer", where
1061     ↪ the answer is either "correct" or "wrong".
1062
1063   user: |-
1064     Our learned patterns: ${hypotheses}
1065     A math problem and its answer are the following:
1066     Problem: "${problems}"
1067     Answer: "${answers}"
1068
1069     Given the patterns you learned above, decide whether the answer is
1070     ↪ correct or wrong.
1071     Think step by step.
1072     First step: Think about which pattern(s) can be applied to the
1073     ↪ answer.
1074     Second step: Based on the patterns, is this answer correct or wrong?
1075     Final step: give your final answer in the format of "Final answer:
1076     ↪ answer"
1077
1078 A.5 High Granularity Prompt
1079
1080 prompt_templates:
1081   observations: |
1082     The math problem: ${problems}
1083     The LLM's answer to the problem: ${answers}

```

1080           The correctness of the answer:  $\{label\}$ .

1081

1082 **batched\_generation:**

1083 **system:** |-

1084     You are a professional math teacher and educational researcher.

1085     Given a set of math problems and LLM-generated answers, we want to

1086     ↪ generate hypotheses that predicts LLMs are more error-prone on

1087     ↪ which particular math skills/concepts (or combinations of math

1088     ↪ skills/concepts)

1089     In other words, we want to understand what kinds of problem are

1090     ↪ associated with correct or wrong labels, what kind of problem

1091     ↪ makes LLM more likely to fail,

1092     what kind of math concept or combination of math concepts make LLM

1093     ↪ likely to fail the problem. Here is a form of math concepts

1094     ↪ involved in the problems:

1094     Math Concepts/Skills

1095     • Integer arithmetic (+,-,\*,÷) • Fraction operations (simplify,

1096     ↪ add/subtract, multiply/divide) • Decimal operations (convert,

1097     ↪ round, compare) • Order of operations (PEMDAS) • Even/odd

1098     ↪ determination • Divisibility rules (2,3,4,5,6,8,9,11) • Factors

1099     ↪ and multiples • GCD and LCM via prime factorization • Absolute

1100     ↪ value • Integer exponents (e.g., 2, 3) • Square roots & cube

1101     ↪ roots (perfect and approximations)

1102     • One-step, multi-step linear equations • Systems of linear equations

1103     ↪ (substitution/elimination) • Linear inequalities & graphing

1104     ↪ solution sets • Simplifying polynomials (combine like terms,

1105     ↪ distributive) • Factoring polynomials: GCF, quadratics,

1106     ↪ difference of squares • Quadratic solving: factoring, completing

1107     ↪ the square, quadratic formula, discriminant interpretation •

1108     ↪ Rational expressions: simplify, multiply/divide, domain

1109     ↪ restrictions • Radical expressions: simplify, rationalize

1110     ↪ denominators • Absolute value equations & inequalities • Function

1111     ↪ evaluation and basic transformations • Function composition •

1112     ↪ Graphing linear functions (slope-intercept, point-slope) • Slope,

1113     ↪ intercept, parallel & perpendicular lines • Matrix basics (2\*2

1114     ↪ inverses, determinants)

1114     • Prime identification & prime factorization • Fundamental theorem of

1115     ↪ arithmetic • Divisibility tests (2-11) • GCD/LCM via Euclidean

1116     ↪ algorithm • Modular arithmetic: congruences,

1117     ↪ addition/multiplication mod n • Solving linear congruences •

1118     ↪ Euler's phi function (n) • Chinese remainder theorem • Parity

1119     ↪ arguments (even/odd reasoning)

1119     • Factorials (n!) • Permutations: P(n,k) • Combinations: C(n,k),

1120     ↪ Pascal's triangle patterns • Distinguish permutations vs

1121     ↪ combinations • Binomial identities & expansion (e.g., (a+b)

1122     ↪ coefficients) • Inclusion-exclusion principle • Basic

1123     ↪ probability: P(A), conditional (P(A|B)), complement rule,

1124     ↪ independent vs dependent events • Probability trees and compound

1125     ↪ events • Counting by cases / systematic enumeration

1126

1127

1128

1129

1130

1131

1132

1133

- 1134 • Angle relations: vertical, alternate interior, supplementary,
- 1135 ↪ complementary • Triangle properties: Pythagorean theorem,
- 1136 ↪ area/perimeter, classification (isosceles, equilateral, right) •
- 1137 ↪ Triangle similarity & congruence (AA, SAS, SSS) •
- 1138 ↪ Inradius/exradius formulas via area • Polygon properties:
- 1139 ↪ interior/exterior angle sums • Circle properties: arc length,
- 1140 ↪ sector area, chords, inscribed angles, tangents, central angles •
- 1141 ↪ Geometric constructions & relationships (e.g., inscribed shapes)
- 1142 ↪ • Area: triangles, quadrilaterals, circles, sectors • Volume &
- 1143 ↪ surface area: prisms, cylinders, cones, spheres • Coordinate
- 1144 ↪ geometry: distance formula, midpoint, slope, line equations,
- 1145 ↪ intercepts • Analytical geometry: shifts, intersections, slopes
- 1146 ↪ in coordinate plane • Polyhedron basics (e.g., face/vertex counts
- 1147 ↪ via Euler's formula)
- 1147 • Polynomial long division & synthetic division • Factor theorem &
- 1148 ↪ Remainder theorem • Finding polynomial zeroes/roots • Rational
- 1149 ↪ functions: domain, graph behavior, asymptotes • Systems mixing
- 1150 ↪ linear and quadratic equations • Quadratic & higher-degree
- 1151 ↪ inequalities (sign analysis, test intervals) • Complex algebraic
- 1152 ↪ manipulations of expressions/inequalities
- 1153 • Sequences & series: arithmetic ( $a_n = a_1 + (n-1)d$ ) & geometric
- 1154 ↪ ( $a_n = a_1 \cdot r^{n-1}$ ); sum formulas • Binomial theorem and summation
- 1155 ↪ identities • Exponential functions and laws (b) • Logarithmic
- 1156 ↪ functions, properties/log rules, log equations • Exponential/log
- 1157 ↪ equations solving (change of base, convert) • Trigonometry: unit
- 1158 ↪ circle fundamentals, sin/cos/tan, right-triangle & standard-angle
- 1159 ↪ values • Basic trig identities (Pythagorean, double-angle,
- 1160 ↪ co-function) • Solving trig equations (within domain
- 1161 ↪ restrictions) • Inverse trigonometric functions (arcsin, arccos,
- 1162 ↪ arctan) • Function properties: domains, ranges, transformations
- 1163 ↪ (shifts/stretch), inverses, composite functions
- 1163 • Conic sections: properties of parabolas, ellipses, hyperbolas •
- 1164 ↪ Polynomial GCD computations • Complex numbers & De Moivre's
- 1165 ↪ theorem • Introductory calculus: symbolic integration, arclength
- 1166 ↪ calculations • Vector calculus: gradients, divergence, curl •
- 1167 ↪ Multivariable calculus: Jacobians, Laplacians • Linear algebra:
- 1168 ↪ eigenvalues/eigenvectors, RREF, characteristic polynomials

1169 Using **!!!ONLY!!!** the given math concepts in the form, **!!!DO NOT!!!**

1170 ↪ include any guessings or conditions other than the concepts in

1171 ↪ the form!!!

1172 **!!!Do NOT!!!** propose conditions or special particularities. For

1173 ↪ example, a hypothesis like "The LLM is likely to fail on a

1174 ↪ concept, especially/particularly when the problem is...(some

1175 ↪ special conditions)" or "The LLM is likely to fail on a concept

1176 ↪ because of (some condition)" is not allowed!

1177 For example: "The LLM is likely to fail on problems that require

1178 ↪ function evaluation and transformations, especially when dealing

1179 ↪ with composite functions and inverses." is a BAD hypothesis that

1180 ↪ doesn't follow the previous requirement.

1180 Another example: "The LLM is likely to fail on (some math concept)

1181 ↪ when the problem is complex" is also a BAD hypothesis that

1182 ↪ doesn't follow the instructions. 'complex' is very vague and is a

1183 ↪ condition that is NOT in the math concept. It will NOT be

1184 ↪ accepted.

1185 **!!!Do NOT!!!** propose hypotheses that are based on a particular step

1186 ↪ in the problem! For example, a hypothesis like "The LLM is likely

1187 ↪ to fail on a concept at a (some step in problem solution)" is not

1187 ↪ allowed!

1188           !!!Do NOT!!! propose any reasons behind the failures! For example, a  
1189           ↪ hypothesis like "The LLM is likely to fail on a concept,  
1190           ↪ because/due to ... (some reason)" is not allowed!  
1191           Again, Use !!!Only!!! the given math concepts in the form! Only means  
1192           ↪ the hypotheses can only contain the concepts in the form! and no  
1193           ↪ other things allowed!  
1194           Using anything other than the math concepts in the form will NOT be  
1195           ↪ accepted and should NOT be proposed!  
1196           These hypotheses should identify specific patterns that occur across  
1197           ↪ the provided problems and LLM-generated answers.  
1198           please propose  $\{\text{num\_hypotheses}\}$  possible hypothesis pairs.  
1199           These hypotheses should identify specific patterns that occur across  
1200           ↪ the provided problems and LLM-generated answers.  
1201           You should check carefully the specific solving steps by the LLM, and  
1202           ↪ consider which particular  
1203           step and which particular math concept/skill did the LLM make mistake  
1204           ↪ on.  
1205           When proposing hypotheses, generate half hypotheses using a single  
1206           ↪ math concept, and generate the other half by combining two or  
1207           ↪ more math concepts.  
1208           Again, use ONLY and VERBATIMLY the provided math concepts from the  
1209           ↪ list  
1210           Each hypothesis should contain the following:  
1211           a. A hypothesis about what particular math makes the LLM to fail  
1212           Generate them in the format of 1. [hypothesis], 2. [hypothesis], ...  
1213           ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
1214           The hypotheses should analyze what particular math concept(s) are  
1215           ↪ associated with correctness or error.  
1216  
1217           **user:** |-  
1218           We have seen some math problems and LLM-generated answers:  
1219            $\{\text{observations}\}$   
1220           Please generate hypotheses that are useful for predicting which  
1221           ↪ particular math concept and solution step does the LLM likely to  
1222           ↪ make mistakes on.  
1223           Propose  $\{\text{num\_hypotheses}\}$  possible hypotheses. Generate them in the  
1224           ↪ format of 1. [hypothesis], 2. [hypothesis], ...  
1225           ↪  $\{\text{num\_hypotheses}\}$ . [hypothesis].  
1226           Proposed hypotheses:  
1227           **inference:**  
1228           **system:** |-  
1229           You are a professional math teacher and your job is to determine  
1230           ↪ whether a given answer to a math problem is correct or wrong.  
1231           From past experience, you have learned that LLMs are more likely to  
1232           ↪ fail on certain math concepts (or combination of math concepts).  
1233           You need to determine whether the learned pattern applies to the  
1234           ↪ current problem and answer, and then make your prediction.  
1235           Give your final answer in the format of "Final answer: answer", where  
1236           ↪ the answer is either "correct" or "wrong".  
1237           **user:** |-  
1238           Our learned pattern:  $\{\text{hypothesis}\}$   
1239           A math problem and its answer are the following:  
1240           Problem: " $\{\text{problems}\}$ "  
1241           Answer: " $\{\text{answers}\}$ "

1242 Given the pattern you learned above, decide whether the answer is  
 1243 ↪ correct or wrong.  
 1244 Think step by step.  
 1245 First step: Consider if the pattern can be applied to the answer.  
 1246 Second step: Based on the pattern, is this answer correct or wrong?  
 1247 Final step: give your final answer in the format of "Final answer:  
 1248 ↪ answer"

1249

1250 **multiple\_hypotheses\_inference:**  
 1251 **system:** |-  
 1252 You are a professional math teacher and your job is to determine  
 1253 ↪ whether a given answer to a math problem is correct or wrong.  
 1254 From past experience, you have learned that LLMs are more likely to  
 1255 ↪ fail on certain math concepts (or combination of math concepts).  
 1256 You need to determine whether these patterns apply to the current  
 1257 ↪ problem and answer, and then make your prediction.  
 1258 Give your final answer in the format of "Final answer: answer", where  
 1259 ↪ the answer is either "correct" or "wrong".

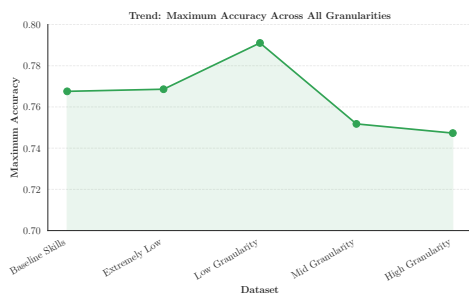
1260

1261 **user:** |-  
 1262 Our learned patterns:  $\{\text{hypotheses}\}$   
 1263 A math problem and its answer are the following:  
 1264 Problem: " $\{\text{problems}\}$ "  
 1265 Answer: " $\{\text{answers}\}$ "

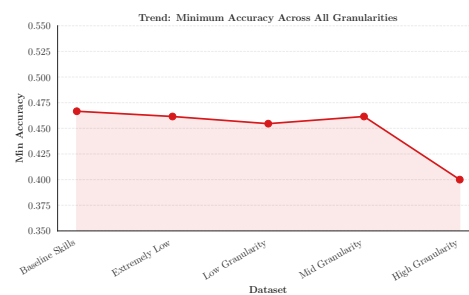
1266 Given the patterns you learned above, decide whether the answer is  
 1267 ↪ correct or wrong.  
 1268 Think step by step.  
 1269 First step: Think about which pattern(s) can be applied to the  
 1270 ↪ answer.  
 1271 Second step: Based on the patterns, is this answer correct or wrong?  
 1272 Final step: give your final answer in the format of "Final answer:  
 1273 ↪ answer"

## 1274 B Granularity and Performance

### 1275 B.1 GPT4Omini



1285 (a) Maximum Accuracy Trend Using GPT4Omini



1295 (b) Minimum Accuracy Trend Using GPT4Omini

Figure 5: Hypotheses Accuracy Trending for GPT4Omini Model

1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

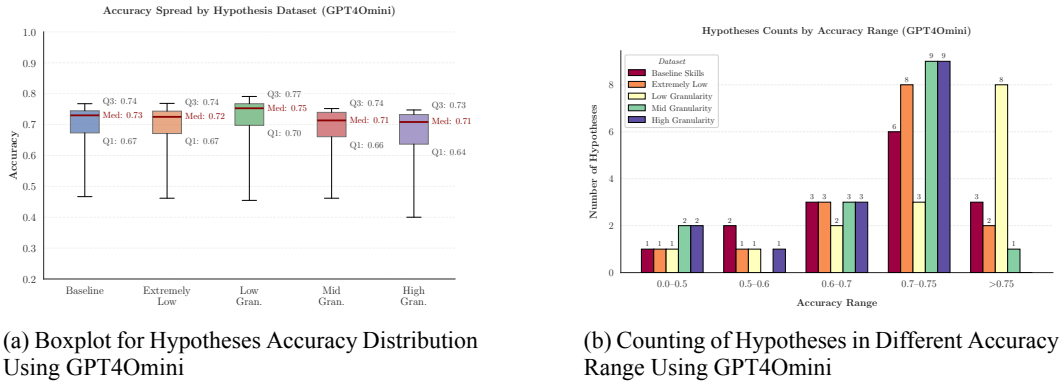


Figure 6: Distribution and Counting of Hypotheses Accuracies Using GPT4Omini

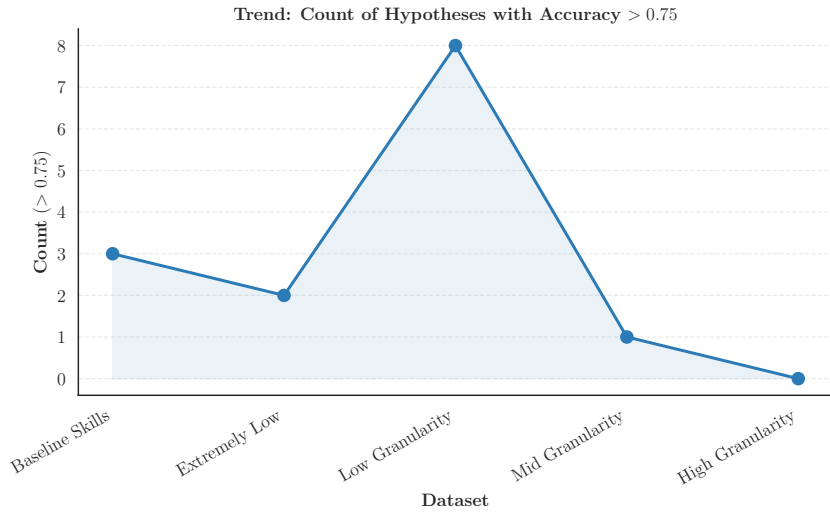


Figure 7: Trending of Number of Hypotheses with Accuracies Over 0.75 Using GPT4Omini

B.2 GPT4.1mini

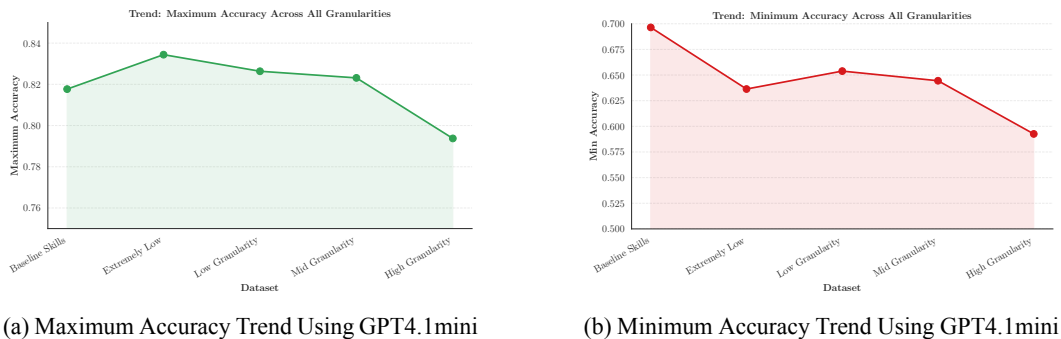


Figure 8: Hypotheses Accuracy Trending for GPT4.1mini Model

1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

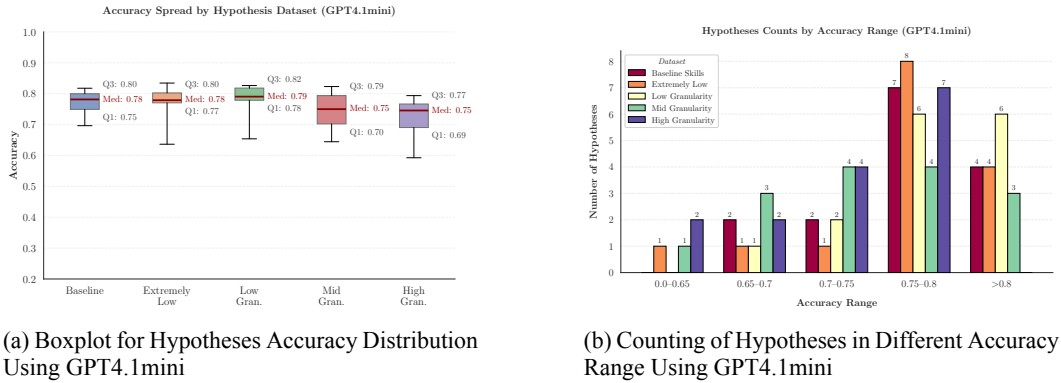


Figure 9: Distribution and Counting of Hypotheses Accuracies Using GPT4.1mini

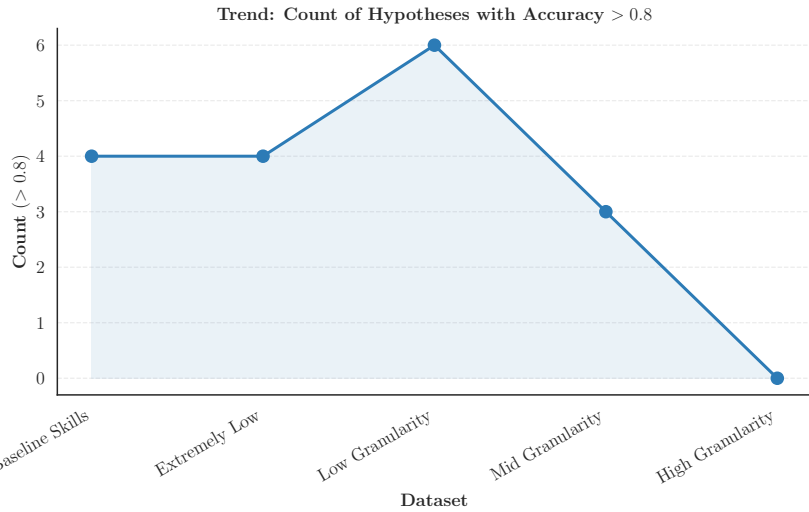


Figure 10: Trending of Number of Hypotheses with Accuracies Over 0.8 Using GPT4.1mini

B.3 Qwen3-14B(Nonthinking mode)

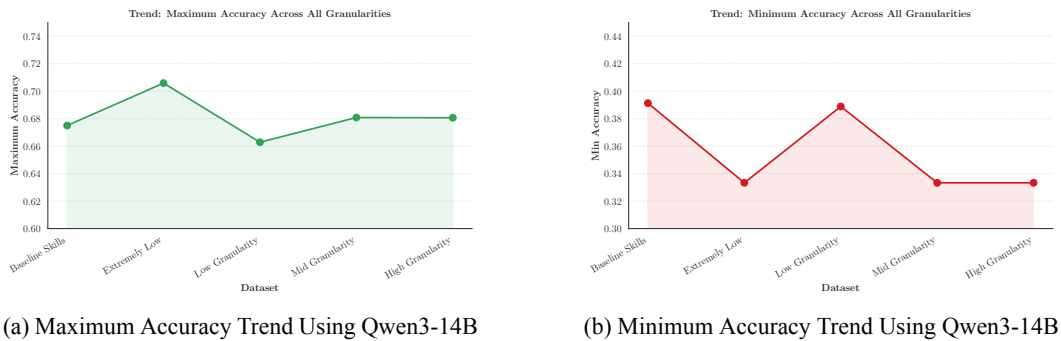


Figure 11: Hypotheses Accuracy Trending for Qwen3-14B Model

1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

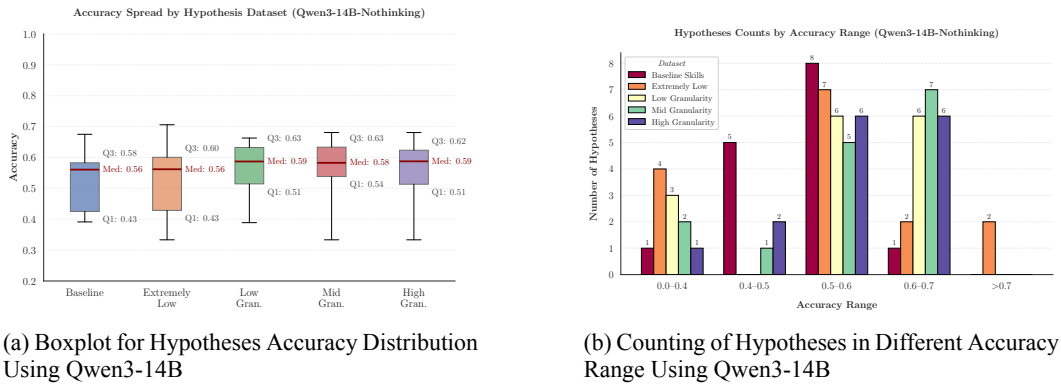


Figure 12: Distribution and Counting of Hypotheses Accuracies Using Qwen3-14B

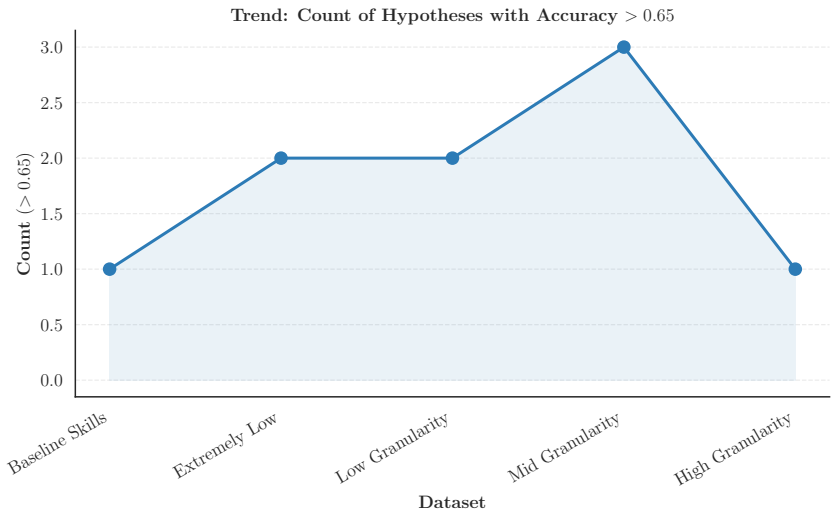


Figure 13: Trending of Number of Hypotheses with Accuracies Over 0.65 Using Qwen3-14B

## C Sample Generated Hypotheses

### C.1 Baseline Skills Hypotheses

- Hypothesis 1: "The LLM is likely to fail on problems involving combinatorics and probability skills combined with calculation and conversion skills." (accuracy: 0.6964285714285714)
- Hypothesis 2: "The LLM is likely to fail on problems requiring calculation and conversion skills." (accuracy: 0.8176795580110499)
- Hypothesis 3: "The LLM tends to fail on problems involving quadratic equations and solutions." (accuracy: 0.6981132075471698)
- Hypothesis 4: "The LLM is likely to fail on problems that involve function composition and transformation combined with understanding and application of functions." (accuracy: 0.7900552486187845)
- Hypothesis 5: "The LLM is likely to fail on problems involving solving system of equations." (accuracy: 0.7814569536423842)
- Hypothesis 6: "The LLM is likely to fail on problems involving geometry triangle properties combined with trigonometry skills." (accuracy: 0.7888198757763973)

1458 Hypothesis 7: "The LLM is likely to fail on problems involving number  
 1459 ↪ theory and arithmetic operations." (accuracy: 0.797546012269939)  
 1460 Hypothesis 8: "The LLM is likely to fail on problems involving geometric  
 1461 ↪ relations and triangle geometry skills." (accuracy: 0.7642857142857142)  
 1462 Hypothesis 9: "The LLM is likely to fail on problems combining algebraic  
 1463 ↪ manipulation skills and solving equations." (accuracy:  
 1464 ↪ 0.8022598870056498)  
 1465 Hypothesis 10: "The LLM is likely to fail on problems requiring solving  
 1466 ↪ system of equations." (accuracy: 0.8048780487804879)  
 1467 Hypothesis 11: "The LLM is likely to fail on problems involving arithmetic  
 1468 ↪ sequences combined with sequence and series skills." (accuracy:  
 1469 ↪ 0.7341772151898734)  
 1470 Hypothesis 12: "The LLM shows difficulty on problems involving coordinate  
 1471 ↪ geometry and transformation skills together with graph understanding  
 1472 ↪ and interpretation." (accuracy: 0.8157894736842105)  
 1473 Hypothesis 13: "The LLM is likely to fail on problems involving  
 1474 ↪ combinatorics and probability skills." (accuracy: 0.77)  
 1475 Hypothesis 14: "The LLM is likely to fail on problems involving  
 1476 ↪ probability concepts and calculations and permutation and  
 1477 ↪ combinations." (accuracy: 0.77)  
 1478 Hypothesis 15: "The LLM is likely to fail on problems involving solving  
 1479 ↪ system of equations and algebraic expression skills." (accuracy:  
 1480 ↪ 0.7090909090909091)

## 1481 C.2 Extremely Low Granularity Hypotheses

1482 Hypothesis 1: "The LLM is likely to fail on problems involving the  
 1483 ↪ combination of Geometry and Algebra." (accuracy: 0.8343949044585988)  
 1484 Hypothesis 2: "The LLM is likely to fail on problems involving the  
 1485 ↪ combination of Number Theory, Geometry, and Counting & Probability."  
 1486 ↪ (accuracy: 0.7737226277372263)  
 1487 Hypothesis 3: "The LLM is likely to fail on problems involving the  
 1488 ↪ combination of Algebra, Number Theory, and Counting & Probability."  
 1489 ↪ (accuracy: 0.7737226277372263)  
 1490 Hypothesis 4: "The LLM is likely to fail on problems involving the  
 1491 ↪ combination of Prealgebra and Counting & Probability." (accuracy:  
 1492 ↪ 0.7870370370370369)  
 1493 Hypothesis 5: "The LLM is likely to fail on problems involving a  
 1494 ↪ combination of Counting & Probability and Geometry." (accuracy:  
 1495 ↪ 0.7788461538461539)  
 1496 Hypothesis 6: "The LLM is likely to fail on problems involving a  
 1497 ↪ combination of Number Theory and Precalculus." (accuracy: 0.776)  
 1498 Hypothesis 7: "The LLM is likely to fail on problems involving both  
 1499 ↪ Geometry and Precalculus." (accuracy: 0.7671232876712328)  
 1500 Hypothesis 8: "The LLM is likely to fail on problems involving both  
 1501 ↪ Prealgebra and Algebra." (accuracy: 0.8295454545454546)  
 1502 Hypothesis 9: "The LLM is likely to fail on problems involving a  
 1503 ↪ combination of Geometry and Precalculus." (accuracy:  
 1504 ↪ 0.7985611510791367)  
 1505 Hypothesis 10: "The LLM is likely to fail on problems involving  
 1506 ↪ Prealgebra." (accuracy: 0.6904761904761905)  
 1507 Hypothesis 11: "The LLM is likely to fail on problems involving the  
 1508 ↪ combination of Geometry and Counting & Probability." (accuracy:  
 1509 ↪ 0.7368421052631579)  
 1510 Hypothesis 12: "The LLM is likely to fail on problems involving the  
 1511 ↪ combination of Prealgebra, Algebra, and Geometry." (accuracy:  
 ↪ 0.7972972972972971)  
 Hypothesis 13: "The LLM is likely to fail on problems involving Number  
 ↪ Theory, Geometry, and Precalculus." (accuracy: 0.8064516129032258)

1512 **Hypothesis 14:** "The LLM is likely to fail on problems involving a  
 1513 ↪ combination of Prealgebra and Algebra." (accuracy: 0.6363636363636364)  
 1514 **Hypothesis 15:** "The LLM is likely to fail on problems involving the  
 1515 ↪ combination of Prealgebra and Algebra." (accuracy: 0.8071428571428572)  
 1516  
 1517

### 1518 C.3 Low Granularity Hypotheses

1519  
 1520 **Hypothesis 1:** "The LLM is likely to fail on problems involving Spatial  
 1521 ↪ reasoning, geometric theorem application." (accuracy:  
 1522 ↪ 0.8092105263157894)  
 1523 **Hypothesis 2:** "The LLM is more likely to make mistakes on problems  
 1524 ↪ involving Combinatorics, probability modeling." (accuracy:  
 1525 ↪ 0.6538461538461539)  
 1526 **Hypothesis 3:** "The LLM is likely to fail on problems involving Multi-step  
 1527 ↪ reasoning, deduction, diagram use." (accuracy: 0.7851239669421488)  
 1528 **Hypothesis 4:** "The LLM is likely to fail on problems involving multi-step  
 1529 ↪ reasoning, deduction, diagram use." (accuracy: 0.8156424581005588)  
 1530 **Hypothesis 5:** "The LLM is likely to fail on problems involving Modular  
 1531 ↪ arithmetic, divisibility, integer properties." (accuracy:  
 1532 ↪ 0.8226950354609929)  
 1533 **Hypothesis 6:** "The LLM is likely to fail on problems that require  
 1534 ↪ Combinatorics, probability modeling and Spatial reasoning, theorem  
 1535 ↪ application." (accuracy: 0.7272727272727273)  
 1536 **Hypothesis 7:** "The LLM is likely to fail on problems involving spatial  
 1537 ↪ reasoning, geometric theorem application." (accuracy:  
 1538 ↪ 0.8225806451612904)  
 1539 **Hypothesis 8:** "The LLM is likely to fail on problems that require spatial  
 1540 ↪ reasoning, theorem application combined with sequences, function  
 1541 ↪ analysis." (accuracy: 0.8206521739130435)  
 1542 **Hypothesis 9:** "The LLM is likely to fail on problems that require  
 1543 ↪ Multi-step reasoning, deduction, diagram use and parabolas, ellipses,  
 1544 ↪ hyperbolas, GCD." (accuracy: 0.7904761904761904)  
 1545 **Hypothesis 10:** "The LLM is likely to fail on problems that require both  
 1546 ↪ Expression manipulation, equation solving and Sequences, function  
 1547 ↪ analysis." (accuracy: 0.782178217821782)  
 1548 **Hypothesis 11:** "The LLM is likely to fail on problems involving  
 1549 ↪ Combinatorics, probability modeling." (accuracy: 0.7798165137614679)  
 1550 **Hypothesis 12:** "The LLM is likely to fail on problems involving multi-step  
 1551 ↪ reasoning, deduction, diagram use combined with sequences, function  
 1552 ↪ analysis." (accuracy: 0.8263473053892217)  
 1553 **Hypothesis 13:** "The LLM is more likely to make mistakes on problems  
 1554 ↪ involving both Sequences, function analysis and Expression  
 1555 ↪ manipulation, equation solving." (accuracy: 0.7978723404255318)  
 1556 **Hypothesis 14:** "The LLM is likely to fail on problems involving Sequences,  
 1557 ↪ function analysis." (accuracy: 0.7777777777777778)  
 1558 **Hypothesis 15:** "The LLM is likely to fail on problems that require both  
 1559 ↪ Modular arithmetic, divisibility, integer properties and Spatial  
 1560 ↪ reasoning, theorem application." (accuracy: 0.723404255319149)

### 1559 C.4 Mid Granularity Hypotheses

1561 **Hypothesis 1:** "The LLM is more error-prone on problems involving function  
 1562 ↪ eval/composition." (accuracy: 0.8231292517006804)  
 1563 **Hypothesis 2:** "The LLM is more likely to fail on problems requiring the  
 1564 ↪ use of linear & systems concepts." (accuracy: 0.8200000000000002)  
 1565 **Hypothesis 3:** "The LLM is more error-prone on problems that involve angles  
 ↪ and triangle theorems." (accuracy: 0.8195876288659791)

1566 Hypothesis 4: "The LLM is more likely to make errors on problems involving  
 1567  $\rightarrow$  polynomial ops and factoring quadratics." (accuracy:  
 1568  $\rightarrow$  0.7999999999999998)  
 1569 Hypothesis 5: "The LLM is more likely to make mistakes on problems  
 1570  $\rightarrow$  involving Linear & systems." (accuracy: 0.7877094972067039)  
 1571 Hypothesis 6: "The LLM shows increased likelihood of error when problems  
 1572  $\rightarrow$  involve sequences & series." (accuracy: 0.7772020725388601)  
 1573 Hypothesis 7: "The LLM tends to make mistakes on problems involving linear  
 1574  $\rightarrow$  & systems." (accuracy: 0.75)  
 1575 Hypothesis 8: "The LLM is more likely to make mistakes on problems  
 1576  $\rightarrow$  combining exponential/logarithmic equations and function  
 1577  $\rightarrow$  inversion/transformation." (accuracy: 0.6941176470588237)  
 1578 Hypothesis 9: "The LLM is more likely to make mistakes on problems  
 1579  $\rightarrow$  involving Linear & systems together with inequalities." (accuracy:  
 1580  $\rightarrow$  0.7222222222222223)  
 1581 Hypothesis 10: "The LLM is more likely to fail on problems involving  
 1582  $\rightarrow$  polynomial division." (accuracy: 0.6444444444444443)  
 1583 Hypothesis 11: "The LLM is more likely to make mistakes on problems  
 1584  $\rightarrow$  involving polynomial division and factor theorem." (accuracy:  
 1585  $\rightarrow$  0.7102803738317754)  
 1586 Hypothesis 12: "The LLM is more likely to make mistakes on problems  
 1587  $\rightarrow$  involving function evaluation/composition combined with function  
 1588  $\rightarrow$  inversion/transformation." (accuracy: 0.7523809523809525)  
 1589 Hypothesis 13: "The LLM is more likely to produce incorrect answers on  
 1590  $\rightarrow$  problems involving GCD and LCM." (accuracy: 0.676923076923077)  
 1591 Hypothesis 14: "The LLM is more error-prone on problems involving circle  
 1592  $\rightarrow$  theorems." (accuracy: 0.6874999999999999)  
 1593 Hypothesis 15: "The LLM tends to fail on problems involving sequences &  
 1594  $\rightarrow$  series." (accuracy: 0.7272727272727274)

## 1595 C.5 High Granularity Hypotheses

1596  
 1597 Hypothesis 1: "The LLM is likely to fail on problems combining Integer  
 1598  $\rightarrow$  arithmetic (+,-,\*, $\div$ ) and Order of operations (PEMDAS)." (accuracy:  
 1599  $\rightarrow$  0.7837837837837838)  
 1600 Hypothesis 2: "The LLM is likely to fail on problems involving Logarithmic  
 1601  $\rightarrow$  functions, properties, and log rules." (accuracy: 0.6666666666666666)  
 1602 Hypothesis 3: "The LLM is likely to fail on problems involving Function  
 1603  $\rightarrow$  evaluation and basic transformations." (accuracy: 0.7937499999999996)  
 1604 Hypothesis 4: "The LLM is likely to fail on problems involving Integer  
 1605  $\rightarrow$  arithmetic (+,-,\*, $\div$ )." (accuracy: 0.7875000000000001)  
 1606 Hypothesis 5: "The LLM is likely to fail on problems involving Factorials  
 1607  $\rightarrow$  (n!) and Permutations P(n,k)." (accuracy: 0.7720588235294118)  
 1608 Hypothesis 6: "The LLM is likely to fail on problems involving Absolute  
 1609  $\rightarrow$  value equations and inequalities." (accuracy: 0.7456140350877193)  
 1610 Hypothesis 7: "The LLM is likely to fail on problems involving systems of  
 1611  $\rightarrow$  linear equations solved by substitution or elimination." (accuracy:  
 1612  $\rightarrow$  0.7611111111111111)  
 1613 Hypothesis 8: "The LLM is likely to fail on problems that combine Integer  
 1614  $\rightarrow$  exponents (e.g., 2, 3) and Square roots & cube roots (perfect and  
 1615  $\rightarrow$  approximations)." (accuracy: 0.7517241379310344)  
 1616 Hypothesis 9: "The LLM is likely to fail on problems involving systems of  
 1617  $\rightarrow$  linear equations (substitution/elimination)." (accuracy:  
 1618  $\rightarrow$  0.7560975609756095)  
 1619 Hypothesis 10: "The LLM is likely to fail on problems combining Complex  
 $\rightarrow$  algebraic manipulations of expressions/inequalities with Quadratic &  
 $\rightarrow$  higher-degree inequalities (sign analysis, test intervals)."  
 $\rightarrow$  (accuracy: 0.7372881355932204)

1620 Hypothesis 11: "The LLM is likely to fail on problems involving polynomial  
 1621  $\rightarrow$  long division and the factor theorem or remainder theorem." (accuracy:  
 1622  $\rightarrow$  0.7342657342657342)  
 1623 Hypothesis 12: "The LLM is likely to fail on problems involving both  
 1624  $\rightarrow$  Quadratic solving and Systems mixing linear and quadratic equations."  
 1625  $\rightarrow$  (accuracy: 0.7073170731707317)  
 1626 Hypothesis 13: "The LLM is likely to fail on problems requiring multi-step  
 1627  $\rightarrow$  linear equations." (accuracy: 0.5925925925925926)  
 1628 Hypothesis 14: "The LLM is likely to fail on problems combining Prime  
 1629  $\rightarrow$  identification & prime factorization with Divisibility tests (2-11)."  
 1630  $\rightarrow$  (accuracy: 0.6734693877551019)  
 1631 Hypothesis 15: "The LLM is likely to fail on problems combining Conic  
 1632  $\rightarrow$  sections: properties of parabolas, ellipses, hyperbolas and Analytical  
 1633  $\rightarrow$  geometry: shifts, intersections, slopes in coordinate plane."  
 1634  $\rightarrow$  (accuracy: 0.6363636363636364)

## 1636 D Problem Generation Prompts

1637  
 1638  
 1639 Your task is to make some math problems to evaluate students' performance  
 1640  $\rightarrow$  on certain math concepts and skills.  
 1641 You decide to make the problems more challenging, so that you can easily  
 1642  $\rightarrow$  test students' understanding  
 1643 of these math concepts/skills. You have also found certain rules that would  
 1644  $\rightarrow$  make the problems more difficult and challenging.  
 1645 You found that students are very likely to fail on:  
 1646 <one Hypothesis>  
 1647 Therefore, you decide to make geometric spatial reasoning problems that  
 1648  $\rightarrow$  involves the application of different geometric theorem so that the  
 1649  $\rightarrow$  problems would be challenging and fully reveal the students' abilities.  
 1650 You decide to make a total of {n} such problems that are similar to  
 1651  $\rightarrow$  problems in the MATH benchmark.  
 1652 Each problem should:  
 1653 1. Be clear and non-ambiguous.  
 1654 2. Be correct and have no factual mistakes. Check carefully before  
 1655  $\rightarrow$  giving the final questions.  
 1656 3. Express clearly so that the students can understand what you mean.  
 1657 4. The problems should have a specific answer, instead of being a proof  
 1658  $\rightarrow$  question. NO PROOF QUESTIONS!  
 1659 5. No true and false problems, yes or no problems, or explanation  
 1660  $\rightarrow$  problems. The problem should have a specific numeric or binary  
 1661  $\rightarrow$  answer.  
 1662 6. Make the problems challenging instead of just doing very simple  
 1663  $\rightarrow$  arithmetic. You can try to involve multiple steps or math  
 1664  $\rightarrow$  concepts/skills in a single problem.  
 1665 When giving the problem statements, you MUST state them in the format:  
 1666  $\rightarrow$  "Problem 1/2/3/...:". Each problem should be separated by a blank line.  
 1667  $\rightarrow$  You must make exactly {n} questions. No more, no less.""

## 1668 E Sample Generated Math Problems

### 1669 E.1 Baseline Skill Granularity Generated Problems

1670  
 1671 Hypothesis used: "The LLM is likely to fail on problems requiring calculation and conversion skills."  
 1672  
 1673

1674 "A car travels from City A to City B at an average speed of 60 km/h and  
1675 ↪ returns at an average speed of 40 km/h. What is the average speed of  
1676 ↪ the car for the entire trip if the distance between City A and City B  
1677 ↪ is 240 km?"

1678 "A bakery sells 250 loaves of bread per day. Each loaf of bread requires  
1679 ↪ 0.5 kg of flour. If the price of flour is \$1.50 per kg, how much does  
1680 ↪ the bakery spend on flour per day?"

1681 "A person invests \$5000 in a savings account that earns an annual interest  
1682 ↪ rate of 4%. How much will the person have in the account after 10  
1683 ↪ years, assuming the interest is compounded annually? (to 2 decimal  
1684 ↪ places)"

1685 "A rectangular prism has a length of 8 cm, a width of 5 cm, and a height  
1686 ↪ of 3 cm. What is the volume of the prism in cubic centimeters?"

1687 "A water tank can hold 1200 liters of water. If  $\frac{3}{4}$  of the tank is already  
1688 ↪ filled, how many more liters of water can be added to fill the tank?"

1689 "A group of friends want to share some candy equally. If they have 48  
1690 ↪ pieces of candy and there are 8 friends, how many pieces of candy will  
1691 ↪ each friend get?"

1692 "A person has \$1000 to invest in two different investments. Investment A  
1693 ↪ earns an annual interest rate of 5%, while Investment B earns an  
1694 ↪ annual interest rate of 7%. If the person invests  $x$  dollars in  
1695 ↪ Investment A and  $(1000 - x)$  dollars in Investment B, how much will the  
1696 ↪ person have after 5 years, assuming the interest is compounded  
1697 ↪ annually?"

1698 "A cylinder has a radius of 4 cm and a height of 10 cm. What is the volume  
1699 ↪ of the cylinder in cubic centimeters?"

1700 "A person buys 15 meters of fabric at \$8 per meter. If the person also  
1701 ↪ buys 20 buttons at \$0.50 each, what is the total cost of the fabric  
1702 ↪ and buttons?"

1703 "A car travels 250 miles in 5 hours. If the car uses 25 gallons of gas,  
1704 ↪ how many miles per gallon does the car get?"

1705 "A bakery sells 300 loaves of bread per day. Each loaf of bread requires  
1706 ↪ 0.75 kg of flour. If the price of flour is \$2.00 per kg, how much does  
1707 ↪ the bakery spend on flour per day?"

1708 "A person invests \$2000 in a savings account that earns an annual interest  
1709 ↪ rate of 6%. How much will the person have in the account after 5  
1710 ↪ years, assuming the interest is compounded annually? (to 2 decimal  
1711 ↪ places)"

1712 "A rectangular prism has a length of 10 cm, a width of 6 cm, and a height  
1713 ↪ of 4 cm. What is the volume of the prism in cubic centimeters?"

1714 "A water tank can hold 1500 liters of water. If  $\frac{2}{3}$  of the tank is already  
1715 ↪ filled, how many more liters of water can be added to fill the tank?"

1716 "A group of friends want to share some candy equally. If they have 72  
1717 ↪ pieces of candy and there are 12 friends, how many pieces of candy  
1718 ↪ will each friend get?"

1719 "A person has \$2000 to invest in two different investments. Investment A  
1720 ↪ earns an annual interest rate of 4%, while Investment B earns an  
1721 ↪ annual interest rate of 8%. If the person invests  $x$  dollars in  
1722 ↪ Investment A and  $(2000 - x)$  dollars in Investment B, how much will the  
1723 ↪ person have after 10 years, assuming the interest is compounded  
1724 ↪ annually?"

1725 "A sphere has a radius of 6 cm. What is the volume of the sphere in cubic  
1726 ↪ centimeters?"

1727 "A person buys 20 meters of fabric at \$10 per meter. If the person also  
↪ buys 30 buttons at \$0.75 each, what is the total cost of the fabric  
↪ and buttons? (to 2 decimal places)"

1728 "A car travels 300 miles in 6 hours. If the car uses 30 gallons of gas,  
 1729 ↪ how many miles per gallon does the car get?"

1730 "A bakery sells 400 loaves of bread per day. Each loaf of bread requires  
 1731 ↪ 0.5 kg of flour. If the price of flour is \$2.50 per kg, how much does  
 1732 ↪ the bakery spend on flour per day?"

1733  
 1734

1735 E.2 Extremely Low Granularity Generated Problems

1736

1737 Hypothesis used: "LLMs are likely to fail on problems involving the combination of Geometry and  
 1738 Algebra."

1739

1740 "In a right triangle with legs measuring 8cm and 15cm, find the length of  
 1741 ↪ the hypotenuse using algebraic expression."

1742 "Find the area of an equilateral triangle inscribed within a circle of  
 1743 ↪ radius 12cm, where each side of the triangle touches the circumference  
 1744 ↪ of the circle."

1745 "Determine the equation for a line passing through two points  $(x_1, y_1) =$   
 1746 ↪  $(-3, -2)$  and  $(x_2, y_2) = (4, 5)$ , and use it to calculate the x-coordinate  
 1747 ↪ at which this line intersects a horizontal line  $y=0$ ."

1748 "Calculate the volume of a pyramid whose base is a square with sides of  
 1749 ↪ length 10 cm and whose height is given as  $h = -b^2 + 30b - 100$ , where b  
 1750 ↪ represents the length of one side of the square base."

1751 "Use geometry to solve for the value of x (rounded to the nearest degree)  
 1752 ↪ in the equation  $\sin(x) = \sqrt{1 - \cos^2(x)}$ , given that  $\cos(x) = 3/5$ ."

1753 "A cube has side lengths of s cm, and its surface area decreases by 16% if  
 1754 ↪ it shrinks to  $3/5s$ . Using algebra, verify if there's any relationship  
 1755 ↪ between the change in dimensions and percentage decrease."

1756 "Derive the vertex of a parabola given its focus  $F=(h,k)=(2,3)$  and  
 1757 ↪ directrix line  $d:x=-5$ ."

1758 "Solve for z in the parametric equations representing the intersection of  
 1759 ↪ a plane and a line. Plane P is represented as  $ax+by+cz+d=0$ ; Line L  
 1760 ↪ passes through point  $Q=(q,r,s)=(1,2,-1)$ ; and direction vector  
 1761 ↪  $n=\langle n_x, n_y, n_z \rangle = \langle 2, 1, -3 \rangle$ . Assume coefficients:  $a=3, b=-2, c=1, d=9$ ."

1762 "Calculate the arc length (to four decimal places) along a curve defined  
 1763 ↪ by  $r(t)=t^2+t$  for t [a,b] where  $[a,b]=[0, ]$ , considering the Cartesian  
 1764 ↪ coordinates and the polar coordinates and give the sum of the two  
 1765 ↪ results."

1766 "Calculate the sum of areas of 25 congruent sectors cut out from a large  
 1767 ↪ circular disc of radius  $R=40\text{cm}$ . Given central angle  $=72^\circ$  corresponding  
 1768 ↪ to one sector."

1769 "Given parallel lines  $l_1:x-y=2$  and  $l_2:x+y=6$  intersecting a line  $l_3:y=x+1$ ,  
 1770 ↪ find the sum (to three decimal places) of their respective distances  
 1771 ↪ to the origin  $O(0,0)$ ."

1772 "If a cone is circumscribing a sphere, determine the sum of the radii  $r_1,$   
 1773 ↪  $r_2$  if it is known that they share the same height  $H=10$  and the ratio  
 1774 ↪ of volumes  $V_{\text{cone}}/V_{\text{sphere}}$  is equal to 27."

1775 "In the given diagram, two triangles  $\triangle ABC$  and  $\triangle ADE$  are shown. The  
 1776 ↪ coordinates of points A, B, C, D, and E are  $(0,0), (2,4), (4,2),$   
 1777 ↪  $(1,1),$  and  $(3,3)$  respectively. Find the area of triangle  $\triangle ADE$ ."

1778 "With reference to Problem 7, substitute the value of k back into the  
 1779 ↪ original equation for calculating p. Then consider finding another  
 1780 ↪ point lying upon parabola by taking a suitable average value (for  
 1781 ↪ instance  $k=p/q$  etc.) from its standard form."

"Evaluate the triple integral over region D defined as  $x^2+y^2+z^2 \leq R^2$  of  
 function  $f(x,y,z)=z/(x^2+y^2+z^2)^{(3/2)}$ , knowing  $R=5$  units."

"Two lines intersect at a point P. The equations of the two lines are  $2x +$   
 ↪  $3y = 7$  and  $x - 2y = -3$ . Find the coordinates of point P."

1782 "A circle centered at  $C(0,h)$  intersects line  $y=h+\sqrt{3}x$ . Assuming  $r$  is the  
 1783  $\hookrightarrow$  distance from center to  $(0,0)$ , find expression relating  $r$  &  $h$  under  
 1784  $\hookrightarrow$  specified constraints."  
 1785 "Calculate numerically the derivative  $dy/dx$  of implicit function:  
 1786  $\hookrightarrow 3xy+x^3-y^2=7$  for point  $(x_0,y_0)=(2,1)$ "  
 1787 "Find the area of a trapezoid with bases of length 8 and 12 and height of  
 1788  $\hookrightarrow$  6. Find its area."  
 1789 "In parallelogram ABCD, the coordinates of points A, B and C are  $(1,1)$ ,  
 $\hookrightarrow (4,4)$ , and  $(6,4)$  respectively. Find the coordinates of point D."

1791

### 1792 E.3 Low Granularity Generated Problems

1793

1794 Hypothesis used: "LLMs are likely to fail on problems involving Modular arithmetic, divisibility,  
 1795 integer properties."

1796

1797 "If  $a \equiv b \pmod{m}$ , then is it true that  $(ab)^n \equiv (ba)^n$   
 1798  $\hookrightarrow \pmod{m}$  for any positive integers  $n$ ?"  
 1799 "Determine the remainder when  $7^{100}$  is divided by 25."  
 1800 "Find all integers  $x$  satisfying  $x^2 + 6x - 8 \equiv 0 \pmod{11}$ ."  
 1801 "What is the smallest positive integer  $N$  such that  $N \equiv 1$   
 1802  $\pmod{4}$ ,  $N \equiv 1 \pmod{9}$ , and  $N \equiv 2 \pmod{13}$ ?"  
 1803 "For how many integers  $x$  where  $1 \leq x < 1000$ , does  $x^2 \equiv 12$   
 1804  $\pmod{19}$  not hold true?"  
 1805 "Let  $m = 13 * 7 * 3$ . How many solutions exist for  $x^2 \equiv 16$   
 1806  $\pmod{m}$ ?"  
 1807 "Evaluate the expression  $(10^{18} + 45) \pmod{23}$ ."  
 1808 "Given  $f(x) \equiv ax + b \pmod{29}$  and  $g(x) \equiv cx + d$   
 1809  $\pmod{29}$ , find the value of  $c$  given  $f(3) = g(3)$ ,  $a = 14$ ,  $b = 27$ ,  $d$   
 $\hookrightarrow = 22$ ."  
 1810 "For how many values of  $m$  does  $36m + 27$  yield a number divisible by 9  
 1811  $\hookrightarrow$  but not 81, where  $m$  is an integer between 1 and 100 inclusive?"  
 1812 "In the equation  $3x^2 + 2y^2 \equiv 15 \pmod{23}$ , determine the number  
 1813  $\hookrightarrow$  of possible remainders of  $x$  when divided by 23."  
 1814 "We define a sequence  $\{a_n\}$  as follows:  $a_n = 3a_{n-1} + 4$ . If  $a_0 = 1$ ,  
 1815  $\hookrightarrow$  find  $a_{2019} \pmod{7}$ ."  
 1816 "Find the number of distinct values for  $a$  such that the congruence  $x^2$   
 1817  $\hookrightarrow \equiv a \pmod{101}$  has at least two distinct solutions."  
 1818 "Consider the system  $x \equiv 2 \pmod{7}$ ,  $x \equiv 0 \pmod{3}$ ,  $x$   
 1819  $\hookrightarrow \equiv 6 \pmod{17}$ . Find the unique solution  $x \pmod{357}$ ."  
 1820 "Solve for  $x$ :  $9x \equiv 30 \pmod{99}$ ."  
 1821 "Consider the polynomial  $f(x) = 4x^4 + ax^3 + bx^2 + cx + 21$ , with  
 1822  $\hookrightarrow$  integer coefficients  $a, b, c$ . If  $r$  is an integer root of this  
 1823  $\hookrightarrow$  polynomial, find the number of possible remainders of  $r$  modulo 7."  
 1824 "Evaluate  $(64^{201} + 37) \pmod{97}$ ."  
 1825 "A positive integer  $n$  is called 'good' if the sum of its digits is a  
 1826  $\hookrightarrow$  multiple of 7. Determine how many good numbers exist between 100000  
 1827  $\hookrightarrow$  and 999999, inclusive."  
 1828 "Determine the remainder when  $7^{100} + (-8)^{200}$  is divided by 15."  
 1829 "How many four-digit multiples of 16 are there which do NOT contain the  
 1830  $\hookrightarrow$  digit 5?"  
 1831 "Given that  $p$  and  $q$  are prime numbers such that  $pq = 35$ , find the value of  
 $\hookrightarrow p^q + q^p \pmod{12}$ ."

1832

### 1833 E.4 Mid Granularity Generated Problems

1834

1835 Hypothesis used: "LLMs are more error-prone on problems involving function evaluation/composition."

1836 "A bakery sells a total of 250 loaves of bread per day. Among the breads,  
 1837  $\rightarrow$  wheat bread has price 5 and white bread has price 7. If they sell a  
 1838  $\rightarrow$  total of 1510, how many white breads did they sell?" $\square$   
 1839 "Solve for y when  $2x + 5y = 11$  and  $x - 2y = -3$ ." $\square$   
 1840 "Find the value of z in the system of equations:  $\begin{cases} nz + 2w = 7 \\ nz - w = 1 \end{cases}$ " $\square$   
 1841 "In a factory, there are two types of machines, X and Y. Machine X  
 1842  $\rightarrow$  produces 200 units per hour and machine Y produces 300 units per hour.  
 1843  $\rightarrow$  The cost of running machine X is \$100 per hour while the cost of  
 1844  $\rightarrow$  running machine Y is \$150 per hour. If the factory operates for 8  
 1845  $\rightarrow$  hours and wants to produce at least 2800 units with a maximum budget  
 1846  $\rightarrow$  of \$1200, how many hours should each type of machine run?" $\square$   
 1847 "A car rental company has two types of cars, compact and SUV. The rental  
 1848  $\rightarrow$  fee for a compact car is \$40 per day and an SUV costs \$60 per day. The  
 1849  $\rightarrow$  company has 50 compact cars and 30 SUVs available. If the total  
 1850  $\rightarrow$  revenue from renting out all the compact cars and 20 SUVs is \$3400,  
 1851  $\rightarrow$  how much money will the company make if it rents out 25 compact cars  
 1852  $\rightarrow$  and all the SUVs?" $\square$   
 1853 "Given two equations:  $\begin{cases} nx + 2y - 3z = 5 \\ 2x - y + z = 3 \end{cases}$  Find the value of  
 1854  $\rightarrow$  x when  $y = 2$  and  $z = 1$ ." $\square$   
 1855 "Tom has been saving money for a new bike and has \$120 in his piggy bank.  
 1856  $\rightarrow$  He wants to buy a bike that costs \$180. His parents have agreed to  
 1857  $\rightarrow$  give him an allowance of \$5 per week for doing chores. However, he  
 1858  $\rightarrow$  spends \$2 per week on snacks. How many weeks will it take Tom to save  
 1859  $\rightarrow$  enough money to buy the bike?" $\square$   
 1860 "A store has two types of discounts: 10% off all items and a  
 1861  $\rightarrow$  buy-one-get-one-free deal on a specific product. If a shirt normally  
 1862  $\rightarrow$  costs \$80 and the store offers both discounts, how much will the  
 1863  $\rightarrow$  customer pay for two shirts?" $\square$   
 1864 "A group of friends want to split a bill evenly. If the total bill comes  
 1865  $\rightarrow$  out to be \$240 and there are 8 friends, but 2 of them didn't eat  
 1866  $\rightarrow$  anything, how much will each friend who ate pay?" $\square$   
 1867 "A farmer has 15 acres of land and plants two types of crops, corn and  
 1868  $\rightarrow$  soybeans. Corn requires 20 gallons of water per acre and soybeans  
 1869  $\rightarrow$  require 15 gallons per acre. If the farmer has 450 gallons of water  
 1870  $\rightarrow$  available, how many acres of corn can he plant?" $\square$   
 1871 "Solve for x in the equation:  $2x + 3(5 - x) = 14$ ." $\square$   
 1872 "If it takes 5 machines 5 minutes to make 5 widgets, how long will it take  
 1873  $\rightarrow$  100 machines to make 100 widgets?" $\square$   
 1874 "A bakery is having a sale on bread. A loaf of bread normally costs \$2,  
 1875  $\rightarrow$  but during the sale, it's 15% off. If a customer buys 2 loaves of  
 1876  $\rightarrow$  bread, how much will they pay?" $\square$   
 1877 "Solve the system of equations:  $\begin{cases} nx + y + z = 6 \\ 2x - y - z = 0 \\ -x + 2y - \\ 3z = -2 \end{cases}$ " $\square$   
 1878 "A student scored 75, 85, and 90 on three quizzes. What score must the  
 1879  $\rightarrow$  student get on the next quiz to have an average of 85?" $\square$   
 1880 "In a right triangle, the length of the hypotenuse is 10 inches and one  
 1881  $\rightarrow$  leg is 6 inches. What is the length of the other leg?" $\square$   
 1882 "A company has two departments: sales and marketing. The sales department  
 1883  $\rightarrow$  has 20 employees and the marketing department has 15 employees. The  
 1884  $\rightarrow$  average salary in the sales department is \$40000 and the average  
 1885  $\rightarrow$  salary in the marketing department is \$50000. What is the total  
 1886  $\rightarrow$  payroll for the company?" $\square$   
 1887 "Solve for y in the equation:  $ny - 2(y - 3) = 7$ " $\square$   
 1888 "A cyclist travels 20 miles in 2 hours. If she increases her speed by 25%,  
 1889  $\rightarrow$  how far will she travel in 3 hours?" $\square$   
 "Solve the equation:  $(x + 2)(x - 3) = 0$  for x."

1890 E.5 High Granularity Generated Problems  
1891

1892 Hypothesis used: "The LLM is more likely to make mistakes on "problems involving function eval-  
1893 uation and basic transformations."

1894 "Find  $f(-x)$  if  $f(x) = (2x^2 + x)/(x+7)$ "  
 1895 "Given  $g(t) = t^2 - 4t + 9$  and  $h(x) = 3x - 12$ , find  $(g \text{ composite } h)(0)$ "  
 1896 "If  $m(a) = 8a - 11$  and  $n(b) = b + 15$ , then  $m(n(2))$  equals"  
 1897 "Let  $p(u) = u^2 - 10u + 16$  and  $q(v) = v - 14$ , find  $p(q(21))$ "  
 1898 "Evaluate  $f(6)$  for  $f(x) = |x - 4| / (x - 3)$ "  
 1899 "For  $r(w) = w^2 - 13w + 40$  and  $s(y) = y^2 + 22y - 24$ , calculate the  
 1900  $\hookrightarrow$  product  $(r*s)(-2)$ "  
 1901 "Determine  $k(z)$  when  $z=5$  if  $k(z) = z/(z-9) + 17/z$ "  
 1902 "Solve  $j(c)$  where  $c=-3$  if  $j(c) = \sqrt{(c^2)-25}/c+3$ , where  $j(c)$  can be  
 1903  $\hookrightarrow$  complex."  
 1904 "If  $d(k) = k^3 - 6k^2 + 9k - 10$  and  $e(j) = 5j + 18$ , find  $d(e(-2))$ "  
 1905 "Calculate  $i(1)$  if  $l=2$  and  $i(l)=(l+5)^2/l-5$ "  
 1906 "Evaluate  $b(f)$  where  $f=10$  for  $b(f) = \text{abs}(f-19)/f+7$ "  
 1907 "Given  $o(p) = p^2 - 2p + 5$  and  $t(d) = d^2 - 7d - 1$ , find the product  
 1908  $\hookrightarrow (o*t)(5)$ "  
 1909 "Let  $a(r) = 7r^2 + 15r + 29$ , evaluate  $a(2r)$  given  $r=4$ "  
 1910 "Find  $y(5)$  for  $y(x) = \max\{x, 30-x\}$ "  
 1911 "Find the value of  $t(h)$  where  $h=-15$  if  $t(h)=\sqrt{81-h^2}+h/h-5$ . The value  
 1912  $\hookrightarrow$  can be complex."  
 1913 "Evaluate  $f(-7)$  if  $f(m) = m*(m-6)*(m-12)+(m-7)^2+(m+2)^2$ "  
 1914 "Let  $c(g) = 23g + 31$ , and let  $d(f) = f + 41$ , find  $c(d(12))$ "  
 1915 "For  $x(g) = g/g-9+11/g$ , determine the value of  $x(6)$ "  
 1916 "compute the average value of  $f(\theta)$  over  $0$  to  $2\pi$  for  $f(\theta) = \sin^2$   
 1917  $\hookrightarrow \theta + \cos^2 2\theta$ "  
 1918 "If  $t(s) = s/(s-6)$ , find the value of  $t(t(t(2)))$ "  
 1919  
 1920  
 1921  
 1922  
 1923  
 1924  
 1925  
 1926  
 1927  
 1928  
 1929  
 1930  
 1931  
 1932  
 1933  
 1934  
 1935  
 1936  
 1937  
 1938  
 1939  
 1940  
 1941  
 1942  
 1943