Too long; didn't solve

Anonymous ACL submission

Abstract

Mathematical benchmarks consisting of a range of mathematics problems are widely used to evaluate the reasoning abilities of large language models, yet little is known about how their structural properties influence model behaviour. In this work, we investigate two structural length variables, prompt length and solution length, and analyse how they relate to model performance on a newly constructed adversarial dataset of expert-authored mathematics problems. We find that both prompt and solution lengths correlate positively with increased model failure across models. We also include a secondary, exploratory analysis of cross-model disagreement. Under a difficulty-adjusted normalised analysis, both variables retain weak negative associations with realised model separation, slightly stronger for prompt length. Overall, our main robust finding is that structural length is linked to empirical difficulty in this dataset.

1 Introduction

The modern landscape of large language model (LLM) evaluation is increasingly shaped by advances in reasoning-oriented models. In the context of mathematical reasoning, benchmarks such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), MathArena (Balunović et al., 2025), OlympiadBench (He et al., 2024), AGIEval (Zhong et al., 2023) and MathVista (Lu et al., 2023) have become standard tools for evaluating the capabilities of LLMs. These typically comprise problems designed to stress multi-step reasoning chains and often require a single numerical or symbolic answer. Sustained improvements in model performance have also recently motivated the development of more sophisticated benchmarks, such as FrontierMath (Epoch AI, 2023), BIG-Bench Extra Hard (Kazemi et al., 2025) and GSM-Symbolic (Mirzadeh et al., 2024), aimed at probing the limits of current systems across a range of metrics.

Evaluation results on these benchmarks are typically reported by aggregating performance across categorical variables such as topic or difficulty level. While informative, these are discrete and can be both coarse and subjective, which may obscure structural patterns in model performance (Zhou et al., 2025). Related work has also highlighted broader challenges in evaluating reasoning systems, including issues of verification and reliability in mathematical outputs (Petrov et al., 2025). Similar concerns about the limitations of static benchmark evaluation have been raised in the broader literature on benchmark design (Kiela et al., 2021). In an effort to circumvent these limitations, we propose to examine continuous structural features of problems, focusing in particular on the word count of the problem statement and its associated solution, both authored by the same person.

The idea of studying prompt-level features appears extensively in the literature (Liu et al., 2023; Zhuo et al., 2024; Mizrahi et al., 2024; Hsieh et al., 2024; Zhang et al., 2024) but has rarely been exploited in the specific arena of LLM mathematical reasoning. Unlike categorical labels, the structural length of prompts and provided solutions is a measurable, model-agnostic quantity worthy of exploration. In this work, we analyse how these structural variables relate to model failure on an adversarially constructed dataset of original expert-authored math problems, and, secondarily, how they relate to cross-model disagreement.

2 Our dataset

Our dataset comprises a collection of 607 complex mathematics problems crafted by a team of Master's degree holders, PhDs, professors, domain experts and IMO medalists, specifically designed to induce failures in SOTA large language models. The process of prompt design leveraged an innovative, proprietary method to achieve a high-quality

standard and ensure absolute originality. Each problem went through a rigorous quality control layer involving human and LLM-verifiers. None of the problems tested in this work were drawn from publicly available sources, nor are they accessible online, safeguarding the analysis from data contamination. Because the benchmark consists of original olympiad-style and IMO-flavoured problems written specifically for this evaluation, the structural effects we observe are unlikely to be artifacts of repeated exposure to familiar public-domain items. A sample of the style of problem-solution pair used in this evaluation can be found in Appendix A.

Each problem or prompt in this collection is categorised by a topic: Geometry, Combinatorics and Discrete Mathematics, Counting and Probability, Algebra, Linear Algebra, Number Theory and Calculus. Prompts are in turn also labelled as either high school, undergraduate or graduate level. Regardless of their assigned level, all problems require complex multi-step reasoning.

Prompts are tested against five different models: ChatGPT 5, ChatGPT 4.1, GPT OSS 120b, Gemini 2.5 Flash, and Claude Sonnet 4.5. Each model performs five independent attempts per task, and we store this information as a fail count $k_{i,m}$ in $\{0, 1, 2, 3, 4, 5\}$ for problem i and model m.

It is worth noting that problems in this dataset were designed to have a single, ground-truth final answer in the form of an integer. Thus, each run returns a binary fail/success result depending on whether the final answer was reached. The dataset also includes full step-by-step solutions to each problem written by its original author.

We define the failure fraction per problem, per model, as the average fail count,

$$x_{i,m} = \frac{k_{i,m}}{5}\,, \tag{1}$$

so that $x_{i,m} \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. This quantity captures empirical instability and error rate of model m on problem i. To summarise the empirical difficulty of problem i, we define its mean failure fraction across models,

$$\mu_i = \frac{1}{M}\sum_{m=1}^{M} x_{i,m}\,, \tag{2}$$

where M is the total number of models. For our dataset, $M = 5$.

We present a visual of the distribution of the total number of problems in our dataset across their given level labels and mean failure fraction in Figure 1.

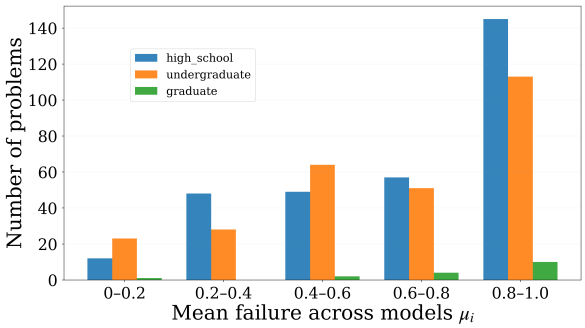


Figure 1: Number of problems, color-coded by level, producing mean failure rate μ_i across all five models.

As is evident from Figure 1, the dataset contains a large number of high-school- and undergraduate-labelled problems, and the histogram is skewed toward the high mean-failure region. This reflects the adversarial nature of the collection: tasks were deliberately authored by human experts to be difficult for current models to solve. In other words, the benchmark is not merely a passive aggregation of existing problems, but a purpose-built evaluation set designed to expose weaknesses in contemporary mathematical reasoning systems.

Statistics on datasets of this kind are often performed based on somewhat arbitrarily given tags, such as the aforementioned topic and level. The labelling of a given problem under a certain level is largely subjective and dependent on several factors, such as differences in education systems. Similarly, topic labels are very coarse variables. Problems that mix different topics are reduced to a single, again subjective choice among all possibilities, and information is thus lost. All in all, the discrete nature of these problem-level variables makes for but a limited analysis.

Throughout this work, instead, we choose to focus on two objectively measurable quantities: the word count in the problem statement and in its given solution. In the following sections, we study the impact of these structural variables on model performance and cross-model disagreement.

3 Structural Length as a driver of difficulty

A preliminary analysis of our data of interest immediately reveals a clear visual correlation between a model's failure fraction $x_{i,m}$ and both the word count in the statement of problem i, and the word

count of its step-by-step solution. As evident from the scatter plots in Figures 2a and 2b, a clear monotonic trend of degrading performance exists as these two structural variables increase.

This trend is visible across all models analysed, despite baseline differences in average ability. We note that, although prompt lengths vary substantially across problems, all tasks remain comfortably within the context window limits of the evaluated models. The naturally arising question is thus whether prompt and solution length are in fact structural drivers of empirical problem difficulty, or whether they simply act as proxies for latent mathematical complexity.

A secondary question is whether these structural variables are also related to cross-model disagreement, although such analyses require care, because disagreement measures are mechanically constrained by mean failure.

In what follows, we analyse prompt and solution lengths as structural drivers of empirical difficulty, and provide a more tentative, exploratory analysis of their relationship to cross-model disagreement.

Prompt Length

To validate our observations, we analyse the relationship between prompt length, measured as the raw word count in problem i, and mean failure per problem, as defined in (2) across all models. Spearman's rank correlation yields $\rho(\text{prompt length}, \mu) = 0.28$, with $p \ll 0.001$, which indicates a small but statistically significant positive association. Longer prompts are therefore more likely to produce model errors on average, which confirms the visual trend in figure 2a.

This positive association is observed across all model families, suggesting that sensitivity to verbosity is not exclusive to a single architecture.

Solution Length

A similar relationship emerges when considering the length of the provided solution. Spearman's rank correlation between solution length and mean failure is $\rho(\text{solution length}, \mu) = 0.32$, with $p \ll 0.001$, indicating a small-to-moderate positive association between the two. As with prompt length, longer solutions are positively correlated with higher model failure on average, consistent with the visual trend in Figure 2b.

However, solution length may more directly reflect underlying mathematical complexity than prompt length does, a distinction we examine further in the following section.

3.1 Exploratory analysis of cross-model disagreement

Having established that both prompt length and solution length are associated with model failure, we now briefly examine whether these structural features are also related to cross-model disagreement.

To quantify disagreement on a given problem, we define the variance of failure fractions across models,

$$\mathrm{Var}_i = \frac{1}{M} \sum_{m=1}^{M} (x_{i,m} - \mu_i)^2 \,. \tag{3}$$

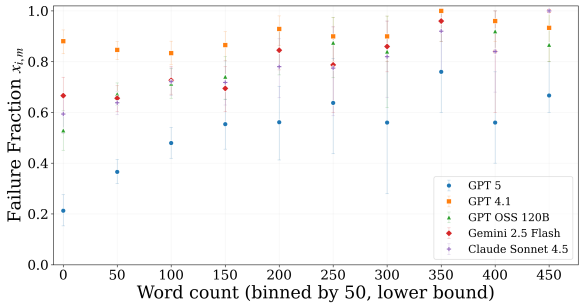
This quantity summarises cross-model disagreement on problem i. A value near zero indicates universal behaviour across models, that is, all models either fail, succeed, or exhibit the same failure rate on the given problem, whereas higher values indicate stronger separation in performance. However, because each $x_{i,m} \in [0, 1]$, the attainable magnitude of Var_i depends on μ_i and satisfies the bound

$$\mathrm{Var}_i \leq \mu_i(1 - \mu_i) \,.$$

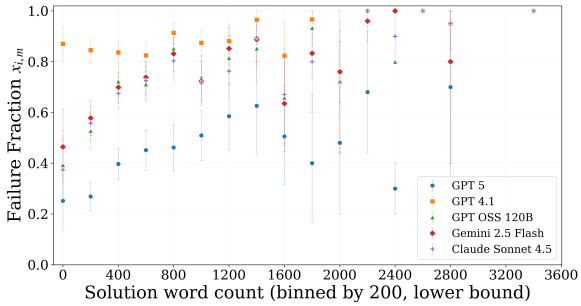
Consequently, the cross-model variance should be interpreted as a disagreement measure whose feasible range depends on mean empirical difficulty. Specifically, cross-model variance is bounded from above by the parabola $\mu_i(1 - \mu_i)$.

Because of this mechanical mean–variance coupling, raw correlations between structural length and cross-model variance are difficult to interpret directly. In a dataset concentrated toward harder problems, any variable that is positively associated with mean failure will tend to exhibit a downward-biased raw correlation with variance due to the variance being geometrically constrained so that the maximum attainable cross-model variance increases monotonically for $\mu_i < 0.5$, but then decreases monotonically for $\mu_i > 0.5$. In our particular dataset, the problems are heavily weighted to have higher mean failure fractions, as it contains a number of frontier-level problems that not even the most sophisticated reasoning models were able to solve. Consequently, there are more tasks in the interval $\mu_i > 0.5$ than $\mu_i < 0.5$, resulting in a negative correlation between cross-model variance and mean failure fraction.

As a more informative exploratory summary, we define a normalized variance score that reduces



(a) Prompt length vs. failure fraction.



(b) Solution length vs. failure fraction.

Figure 2: Failure fraction $x_{i,m}$ as a function of structural length variables. Error bars show 95% bootstrap confidence intervals.

the dependence of cross-model variance on mean failure fraction, and instead focuses on the variance achieved at a given difficulty level:

$$\widetilde{\mathrm{Var}}_i = \frac{\mathrm{Var}_i}{\mu_i(1-\mu_i)}\,. \tag{4}$$

This metric measures the fraction of the theoretical maximum variance achieved by problem i at its observed difficulty level. This quantity is only defined for tasks with $0 < \mu_i < 1$, so the normalised analysis excludes tasks that were empirically solved by all models or failed by all models. In our dataset, this leaves 517 tasks.

Under this normalisation, prompt length retains a weak negative association with difficulty-adjusted cross-model disagreement:

$$\rho(\text{prompt length}, \widetilde{\mathrm{Var}}) = -0.21, \qquad p \ll 0.001.$$

Solution length also retains a weak negative association, although the effect is somewhat smaller:

$$\rho(\text{solution length}, \widetilde{\mathrm{Var}}) = -0.17, \qquad p \ll 0.001.$$

While the normalisation by $\mu_i(1-\mu_i)$ removes the dominant outer bound linking mean failure and variance, it does not eliminate all finite-sample and geometric structure induced by the discreteness of the observed failure fractions and the small number of models. Accordingly, these correlations should be understood as exploratory descriptive summaries of this benchmark rather than as evidence of an independent structural compression effect.

3.1.1 Hierarchical modelling of structural effects

Our previous analyses so far reveal systematic relationships between structural length and model behavior. They do not, however, account for heterogeneity across model families. Because different LLMs exhibit distinct baseline failure rates and sensitivities to problem structure, we fit a hierarchical (mixed-effects) regression model to jointly capture global structural trends and model-specific deviations. We apply this framework separately to prompt length and solution length to determine whether the structural effects identified above persist when accounting for model-level variability.

Length Variable	$\rho(\mu)$	$\rho(\widetilde{\mathrm{Var}})$
Prompt	0.28	−0.21
Solution	0.32	−0.17

Table 1: Task-level Spearman correlations between structural length variables and model behaviour. The second column reports the primary result of the paper, namely the association with mean failure. The third column reports a secondary exploratory association with difficulty-adjusted cross-model disagreement, computed on the 517 tasks with $0 < \mu_i < 1$.

We model the observed failure fraction $x_{i,m}$ for problem i and model m as a function of a log-transformed measure of length L_i,

$$L_i = \log(1 + \text{word count}_i)\,. \tag{5}$$

The fitted model is

$$x_{i,m} = \beta_0 + \beta_1 L_i + u_m + v_m L_i + \varepsilon_{i,m}, \tag{6}$$

where β_0 is the global intercept (measuring average baseline failure fraction), β_1 is the global average effect of structural length, u_m is a model-specific random intercept which captures baseline performance differences between models, v_m is a model-specific random slope for length that captures model-specific length sensitivity, and $\varepsilon_{i,m}$ is residual noise. Throughout the analysis we assume u_m, v_m and $\varepsilon_{i,m}$ to be normally distributed,

$$\begin{pmatrix} u_m \\ v_m \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \right),$$

$$\varepsilon_{i,m} \sim \mathcal{N}(0, \sigma^2).$$

Prompt length

The model in (6) was estimated for $L_i^{(\text{prompt})}$ via REML with 3035 observations across the 5 LLMs under study. Table 2 presents a summary of results, and figure 3a represents the fitted hierarchical model by plotting the predicted failure trajectories for each LLM alongside the global fixed-effect trend.

Prompt Length Model			
Parameter	**Estimate**	**SE**	**p-value**
β_1	0.118	0.037	0.001
σ_u^2	0.281	–	–
σ_v^2	0.006	–	–
σ_{uv}	-0.042	–	–
σ^2	0.1153	–	–

Table 2: Mixed-effects model estimates for failure fraction $x_{i,m}$ as a function of log-transformed prompt length, $L_i^{(\text{prompt})}$.

Fixed effects The estimated effect of $L_i^{(\text{prompt})}$ on failure fraction was $\beta_1 = 0.118$ $(SE = 0.037\,, p = 0.001)$. This means a one-unit increase in $L_i^{(\text{prompt})}$ is associated with an average increase of approximately 0.118 in $x_{i,m}$, which confirms the earlier Spearman correlation analysis, i.e. longer prompts are associated with higher failure rates.

Random effects The random intercept variance, $\sigma_u^2 = 0.281$, is substantial, indicating meaningful baseline differences in failure rates across models. This heterogeneity is consistent with the vertical separation between models observed in the binned scatterplots (Figure 2a) and in the fitted trajectories shown in Figure 3a, and justifies the use of a hierarchical specification.

In contrast, the random slope variance, $\sigma_v^2 = 0.006$, is very small, indicating minimal variation across models in sensitivity to prompt length. This is visually reflected in the near-parallel fitted lines in Figure 3a, suggesting that models degrade at comparable rates as prompt length increases.

Finally, the residual variance $\sigma^2 = 0.1153$ remains substantial, indicating that additional problem-level factors beyond prompt length contribute to variation in $x_{i,m}$.

Overall, these results suggest that prompt length acts as a global structural stressor which affects all models similarly and does not selectively degrade weaker models.

Solution length

We estimated the same hierarchical model using $L_i^{(\text{solution})}$, the log-transformed word count of the reference solution. The model was fit using 3030 observations across the five LLMs. Table 3 reports the parameter estimates, and Figure 3b visualizes the corresponding fitted trajectories.

Solution Length Model			
Parameter	**Estimate**	**SE**	**p-value**
β_1	0.137	0.026	0.001
σ_u^2	0.225	–	–
σ_v^2	0.003	–	–
σ_{uv}	-0.025	–	–
σ^2	0.1128	–	–

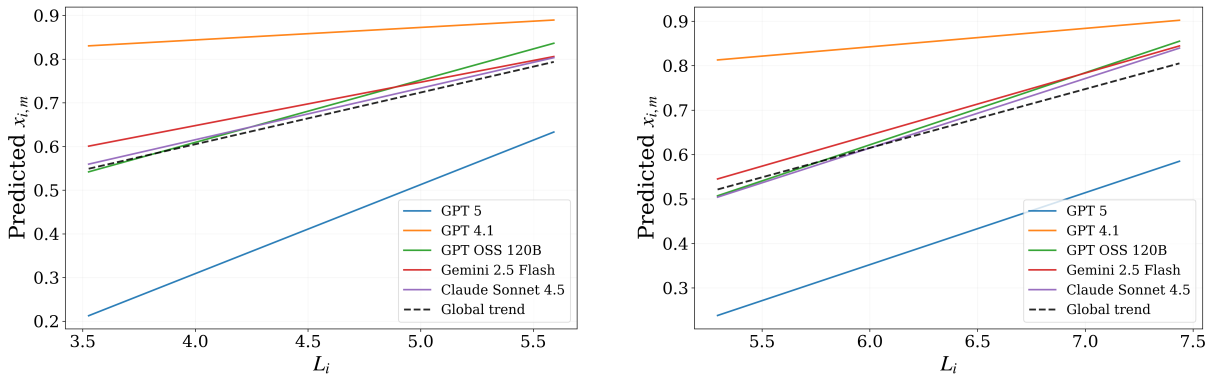
Table 3: Mixed-effects model estimates for failure fraction $x_{i,m}$ as a function of log-transformed solution length, $L_i^{(\text{solution})}$.

Fixed effects The fixed-effect coefficient for solution length is positive, $\beta_1 = 0.137$, indicating that problems with longer reference solutions tend to produce higher failure fractions. This result is consistent with the earlier correlation analysis and suggests that tasks requiring longer solutions are generally more difficult for models to solve reliably.

Random effects As in the prompt-length model, the random intercept variance $\sigma_u^2 = 0.225$ remains substantial, reflecting baseline differences in failure rates across models. However, variation in slopes across models is again limited ($\sigma_v^2 = 0.003$), indicating that models respond similarly to increases in solution length.

The residual variance $\sigma^2 = 0.1128$ remains non-negligible, indicating that solution length alone does not fully account for problem-level variation in failure rates.

Taken together, these results suggest that solution length is associated with increased task difficulty, but unlike prompt length, it likely reflects the intrinsic complexity of the underlying mathematics rather than a structural property of the prompt itself.



(a) Failure fraction as predicted by prompt length.

(b) Failure fraction as predicted by solution length.

Figure 3: Predicted failure fraction $x_{i,m}$ as a function of $L_i = \log(1 + \text{word count})$ under the fitted mixed-effects model. Solid lines show model-specific fits; the dashed line shows the global fixed effect.

4 Discussion

Our work studies how two objectively measurable structural variables, prompt length and reference-solution length, relate to model failure and model separation on an adversarially constructed mathematics benchmark.

At the prompt level, we find that prompt length is a consistent predictor of empirical difficulty across all evaluated models: longer prompts are associated with higher mean failure rates. This relationship is also reflected in the mixed-effects analysis, whose fitted trends suggest that all five models degrade in performance as prompt length increases. At the same time, the fitted slopes are broadly similar, indicating that the models do not appear to differ dramatically in their sensitivity to prompt length at the level captured by this analysis.

The mixed-effects analysis provides a useful descriptive summary of the length–failure relationship across models, but it should not be over-interpreted as a fully generative account of the data. In particular, the response variable is discrete and bounded, the number of models is small, and the present specification does not explicitly model problem-level random effects. We therefore regard the fitted trends mainly as evidence that the positive association between structural length and failure is broadly shared across the evaluated models.

At the level of problem–solution pairs, we likewise find that solution length is a significant predictor of model failure: longer reference solutions are associated with harder problems on average. This is consistent with prior work showing that mathematical reasoning difficulty often scales with the number of steps needed to arrive at a correct solution (Wei et al., 2022). As with prompt length, the fitted slopes in the mixed-effects analysis are broadly similar across models, suggesting that solution length functions mainly as a shared source of difficulty rather than a model-specific weakness.

The interpretation of cross-model disagreement is more delicate. Because disagreement measures based on variance are mechanically constrained by mean failure, raw variance correlations are difficult to interpret, especially in a dataset skewed toward harder tasks. For that reason, we treat the disagreement analysis as exploratory and focus on a simple normalised variance measure. Under this adjustment, both prompt length and solution length retain only weak negative associations with realised model separation. The prompt-length association is somewhat stronger than the solution-length one, but we interpret this difference as a preliminary indicator, as further investigation is needed.

Nevertheless, we do observe a qualitative asymmetry: while both prompt and solution length predict difficulty, only prompt length retains a residual association with reduced model separation after controlling for mean failure. This suggests that longer items may be somewhat less effective at separating models at a given difficulty level in this particular benchmark.

This difference may reflect the distinct roles of the two variables: prompt length is closer to a surface property of task presentation, whereas solution length may additionally reflect intrinsic mathematical complexity, authoring style, and the granularity with which reasoning is written down. More broadly, these results are specific to this adversarial dataset, this model set, and this evaluation protocol. We therefore view the disagreement analysis

as indicative rather than definitive.

All results taken together, the clearest conclusion is that there is a weak to moderate effect of structural length on failure rate: both prompt length and solution length are linked to empirical hardness in this benchmark. A secondary, tentative conclusion is that structural length may also be related to realised model separation. However, this part of the story is markedly less clear-cut, because disagreement is geometrically and statistically constrained. In the context of benchmark design, this reinforces the importance of analysing not only whether tasks are difficult, but also whether they remain informative for distinguishing model capabilities (Kiela et al., 2021; Singh et al., 2025).

These findings naturally motivate the study of these questions across a range of benchmarks, model families, and evaluation protocols, and the further development of richer disagreement measures that better account for the geometric and finite-sample constraints inherent in repeated-evaluation settings. This is particularly relevant in light of recent work on long-context evaluation (Bai et al., 2023, 2024), where structural length and reasoning burden may interact in complex ways.

4.1 Limitations and future analytical directions

Several of the analyses in this work are best understood as descriptive first-pass summaries rather than final statistical treatments of the underlying phenomena. In particular, our main response variable $x_{i,m}$ is discrete and bounded, since it is derived from only five attempts per model, and our disagreement measure is mechanically constrained by mean failure. For this reason, richer follow-up analyses would be valuable.

A natural next step would be to replace the linear mixed-effects analysis with a hierarchical binomial model. Instead of treating $x_{i,m}$ as approximately continuous, one could model the fail count $k_{i,m}$ directly as a binomial outcome with five trials, while allowing model-specific and problem-specific random effects. Such a formulation would better respect the discrete structure of the data, allow uncertainty to be propagated more appropriately, and provide a more principled estimate of how prompt length and solution length relate to failure probability.

A second improvement would be to model problem-level heterogeneity more explicitly. In the present analysis, structural length is used as a proxy for aspects of task complexity, but it is unlikely to capture all relevant variation across problems. Future work could introduce crossed random effects for both model and problem, or latent problem-difficulty parameters, in order to separate the contribution of structural length from unobserved mathematical difficulty more cleanly.

The analysis of cross-model disagreement could also be refined. Because variance is bounded by mean failure, raw variance correlations are geometrically constrained and difficult to interpret. Our normalised variance analysis removes the dominant outer bound, but it does not eliminate all finite-sample and combinatorial structure induced by the small number of models and the discreteness of the observed failure fractions. A more informative direction would therefore be to develop disagreement measures derived from an explicit probabilistic model, or to study latent between-model dispersion parameters jointly with mean difficulty rather than summarising disagreement through task-level variance alone.

More broadly, it would be useful to test the robustness of the present findings across additional datasets, model families, and evaluation protocols. Our benchmark is intentionally adversarial and the analysis is based on five contemporary models evaluated under a fixed repeated-attempt setup. Replicating the structural-length analyses on other mathematics benchmarks, as well as on future generations of reasoning models, would help determine which effects are benchmark-specific and which reflect more stable properties of LLM mathematical reasoning.

Finally, our prospective lines of research include investigating structural effects at a finer level than raw word count alone. For example, one could examine whether particular forms of verbosity, such as notational density, number of stated conditions, amount of irrelevant context, or solution branching depth, are more informative predictors than length itself. Such analyses may help distinguish whether the observed effects arise from surface form, intrinsic mathematical complexity, or an interaction between the two.

A final limitation is that the reported correlations are necessarily contingent on the benchmark and model family under study. Our dataset was intentionally curated to be adversarial and difficult, and the analysis is based on a fixed set of five contemporary models evaluated under a specific protocol. Accordingly, the quantitative relationships

reported here should be understood as descriptive of this setting, rather than as universally stable estimates of how structural length affects difficulty or cross-model variance across all mathematical benchmarks and all model classes.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. ArXiv:2308.14508.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. ArXiv:2412.15204.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. MathArena: Evaluating LLMs on uncontaminated math competitions. ArXiv:2505.23281.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. ArXiv:2110.14168.

Epoch AI. 2023. FrontierMath.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. ArXiv:2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. ArXiv:2103.03874.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? ArXiv:2404.06654.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. 2025. BIG-Bench Extra Hard. ArXiv:2502.19187.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. ArXiv:2104.14337.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. ArXiv:2307.03172.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. ArXiv:2310.02255.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. ArXiv:2410.05229.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. ArXiv:2401.00595.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad. ArXiv:2503.21934.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2025. The Leaderboard Illusion. ArXiv:2504.20879.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens. ArXiv:2402.13718.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. ArXiv:2304.06364.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. 2025. GSM-Infinite: How

Do Your LLMs Behave over Infinitely Increasing Context Length and Reasoning Complexity? ArXiv:2502.05252.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. ArXiv:2410.12405.

A Sample problem–solution pair

To illustrate the style of the benchmark, we include one representative original problem–solution pair below.

Sample Problem

A carpenter wants a rectangle with positive integer side lengths u and v such that its area is

$$uv = m^{19},$$

and its side lengths satisfy the extra constraint

$$v - 3u \equiv 0 \pmod{14}$$

for some m in the set $\{1, 2, \ldots, 30\}$.
For some values of m, no such rectangle exists. Find the sum of all such integers m.

Sample Solution

Step 1: Rewrite the congruence condition

We seek integers $m \in \{1, 2, \ldots, 30\}$ for which there do *not* exist positive integers u, v satisfying

$$uv = m^{19}$$

and

$$v - 3u \equiv 0 \pmod{14}.$$

The latter congruence is equivalent to the system

$$\begin{aligned} v - 3u &\equiv 0 \pmod{7}, \\ v - 3u &\equiv 0 \pmod{2}. \end{aligned}$$

Modulo 7, this becomes

$$v \equiv 3u \pmod{7},$$

while modulo 2, since $3 \equiv 1 \pmod{2}$, it becomes

$$v \equiv u \pmod{2}.$$

Thus we seek factor pairs (u, v) of m^{19} such that

$$uv = m^{19},$$

and also

$$\begin{aligned} v &\equiv 3u \pmod{7}, \\ v &\equiv u \pmod{2}. \end{aligned}$$

Equivalently, u and v must have the same parity, and their residues modulo 7 must satisfy $v \equiv 3u$.

Step 2: Handle the case where m is a multiple of 7

First consider m divisible by 7. In $\{1, 2, \ldots, 30\}$, these are

$$m \in \{7, 14, 21, 28\}.$$

Since $7 \mid m$, we also have $7 \mid m^{19}$. If we choose u and v both divisible by 7, then automatically

$$u \equiv v \equiv 0 \pmod{7},$$

so

$$v \equiv 3u \pmod{7}$$

holds.
It remains only to match parity.
If m is odd, namely 7 or 21, then m^{19} is odd, so choosing

$$u = 7, \qquad v = \frac{m^{19}}{7}$$

gives both u and v odd.
If m is even, namely 14 or 28, then choosing

$$u = 14, \qquad v = \frac{m^{19}}{14}$$

gives both u and v even.
Hence every multiple of 7 in $\{1, \ldots, 30\}$ is admissible. Therefore none of

$$7, 14, 21, 28$$

belongs to the set we seek.

Step 3: Assume $7 \nmid m$ and derive a necessary condition

Now assume $7 \nmid m$. Then $7 \nmid m^{19}$, so any admissible u and v are invertible modulo 7.
From

$$uv = m^{19}$$

and

$$v \equiv 3u \pmod{7},$$

multiplying the congruence by v gives

$$v^2 \equiv 3uv \equiv 3m^{19} \pmod{7}.$$

By Fermat's little theorem,

$$m^6 \equiv 1 \pmod{7},$$

hence

$$m^{19} = m^{18} \cdot m \equiv m \pmod{7}.$$

Therefore

$$v^2 \equiv 3m \pmod{7}.$$

So a necessary condition for the existence of a valid rectangle is that $3m$ be a quadratic residue modulo 7.
The quadratic residues modulo 7 are

$$0,\ 1,\ 2,\ 4.$$

Since $7 \nmid m$, we have $3m \not\equiv 0 \pmod{7}$, so we need

$$3m \equiv 1,\ 2,\ \text{or } 4 \pmod{7}.$$

The inverse of 3 modulo 7 is 5, so this is equivalent to

$$m \equiv 5,\ 3,\ \text{or } 6 \pmod{7}.$$

Thus if

$$m \equiv 1,\ 2,\ \text{or } 4 \pmod 7,$$

then no valid rectangle can exist.
The numbers in $\{1, \ldots, 30\}$ in these residue classes are

$$1, 8, 15, 22, 29,$$

$$2, 9, 16, 23, 30,$$

$$4, 11, 18, 25.$$

So the following 14 values definitely fail:

$$1, 2, 4, 8, 9, 11, 15, 16, 18, 22, 23, 25, 29, 30.$$

Step 4: Examine the remaining residue classes

We now examine the cases

$$m \equiv 3,\ 5,\ \text{or } 6 \pmod 7.$$

Case 4a: $m \equiv 3 \pmod 7$.
These are

$$3, 10, 17, 24.$$

Here

$$3m \equiv 2 \pmod 7,$$

so we need

$$v^2 \equiv 2 \pmod 7,$$

whose solutions are

$$v \equiv 3,\ 4 \pmod 7.$$

For each of these values, taking $v = m$ works:

$$3 \equiv 3, \qquad 10 \equiv 3, \qquad 17 \equiv 3, \qquad 24 \equiv 3 \pmod 7.$$

Also $u = m^{18}$ has the same parity as $v = m$. Hence all four values are admissible.

Case 4b: $m \equiv 5 \pmod 7$.
These are

$$5, 12, 19, 26.$$

Now

$$3m \equiv 1 \pmod 7,$$

so we need

$$v^2 \equiv 1 \pmod 7,$$

whose solutions are

$$v \equiv 1,\ 6 \pmod 7.$$

Taking $v = m^3$ works in each case, because

$$5^3 \equiv 6 \pmod 7,$$

and hence similarly

$$12^3 \equiv 19^3 \equiv 26^3 \equiv 6 \pmod 7.$$

Then $u = m^{16}$, so u and v have the same parity. Thus all four values are admissible.

Case 4c: $m \equiv 6 \pmod 7$.
These are

$$6, 13, 20, 27.$$

Now

$$3m \equiv 4 \pmod 7,$$

so we need

$$v^2 \equiv 4 \pmod 7,$$

whose solutions are

$$v \equiv 2,\ 5 \pmod 7.$$

For $m = 6$, the divisors of 6^{19} include 2, and

$$2^2 \equiv 4 \pmod 7.$$

Taking $v = 2$ works, and then

$$u = \frac{6^{19}}{2}$$

is even, so parity also matches.
For $m = 20$, the same choice $v = 2$ works, since $2 \mid 20^{19}$.
For $m = 27$, the divisors are powers of 3, and taking

$$v = 3^2 = 9$$

gives

$$v \equiv 2 \pmod 7, \qquad v^2 \equiv 4 \pmod 7.$$

Also both v and

$$u = \frac{27^{19}}{9} = 3^{55}$$

are odd, so parity matches.
For $m = 13$, however, every divisor is a power of 13, and

$$13 \equiv -1 \pmod 7.$$

So every divisor is congruent to either 1 or 6 modulo 7, and squaring either gives

$$1^2 \equiv 1, \qquad 6^2 \equiv 1 \pmod 7.$$

Thus no divisor v of 13^{19} can satisfy

$$v^2 \equiv 4 \pmod 7.$$

Therefore $m = 13$ is not admissible.

Step 5: Collect the failing values and sum them

The full set of $m \in \{1, \ldots, 30\}$ for which no such rectangle exists is

$$\{1, 2, 4, 8, 9, 11, 13, 15, 16, 18, 22, 23, 25, 29, 30\}.$$

Their sum is

$$\begin{aligned} &1 + 2 + 4 + 8 + 9 + 11 + 13 + 15 \\ &+ 16 + 18 + 22 + 23 + 25 + 29 + 30 \\ &= (1 + 29) + (2 + 30) + (4 + 25) + (8 + 22) \\ &\quad + (9 + 23) + (11 + 18) + (13 + 16) + 15 \\ &= 30 + 32 + 29 + 30 + 32 + 29 + 29 + 15 \\ &= 226. \end{aligned}$$

Therefore the answer is

$$\boxed{226}.$$