WHAT MAKES SOMETHING HARD(ER)? EXPLAINING QUESTION DIFFICULTY IN NATURAL LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Difficulty is one of the most fundamental properties of a question: it determines whether the question can meaningfully discriminate between models of differing ability. Although a variety of methods can now estimate or predict difficulty automatically, they yield only a single descriptive number, with no account of the underlying factors that make a question difficult in the first place. In this work, we propose a data-driven approach that automatically generates and validates natural-language hypotheses explaining what makes one question harder than another. We first estimate each item's difficulty from the responses of a large pool of LLMs using Item Response Theory. We then sample contrasting sets of easy and hard questions and prompt an LLM to propose candidate explanations of the difference, which are subsequently validated and selected on held-out questions. Experimental results across three datasets spanning mathematical, logical, and commonsense reasoning show that our method produces interpretable and predictive hypotheses. On their own, they predict the difficulty of unseen questions competitively with, or better than, advanced black-box difficulty regressors; used as additional features, they further improve those regressors, implying that they discover difficulty signals that existing models fail to capture. Moreover, we demonstrate that editing questions according to a hypothesis can shift their measured difficulty in the expected direction, indicating that the discovered hypotheses are causally valid difficulty factors rather than post-hoc descriptions. Our approach thus turns a purely descriptive difficulty score into actionable statements.

1 INTRODUCTION

The rapid growth of LLM capabilities has led to saturation on existing benchmarks (Akhtar et al., 2026), driving continuous benchmark iteration: harder, more challenging datasets are needed to keep discriminating between increasingly capable models (Rein et al., 2024; Wang et al., 2024; White et al., 2025). Yet what determines the difficulty of a question in the first place remains poorly understood. Prior work on question difficulty has largely focused on *measuring* difficulty and *predicting* difficulty. Measurement typically relies on the responses of test-takers. The representative approach in this field is the Item Response Theory (IRT) family of models (Hambleton & Swaminathan, 2013; Hambleton et al., 1991), which decomposes observed performance into item properties and test-taker ability. IRT has increasingly been adopted for LLM evaluation (Zouhar et al., 2025; Cui & Sachan, 2025). However, it only yields a descriptive number per item but does not reveal the underlying factors that make an item hard in the first place.

Some follow-up work (Fischer, 1973; Janssen et al., 2004; Boeck & Wilson, 2004) attempted to represent an item's difficulty as a linear combination of hand-specified item features, such as the cognitive operations it requires. This turns difficulty from a free parameter into an outcome of question properties, and thereby shifts the question from measuring difficulty to predicting and explaining it. A line of feature-based difficulty prediction followed across domains (Benedetto et al., 2023; AlKhuzaey et al., 2024). Such features, however, must be designed by domain experts and coded by hand, which makes these approaches hard to scale across domains or to large question pools. More recent work instead predicts difficulty with neural network or pre-trained LLMs that read the item difficulty directly (Huang et al., 2017; Benedetto et al., 2020; Loginova et al., 2021). While these methods usually enjoy stronger performance and greater flexibility they predict diffi-

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

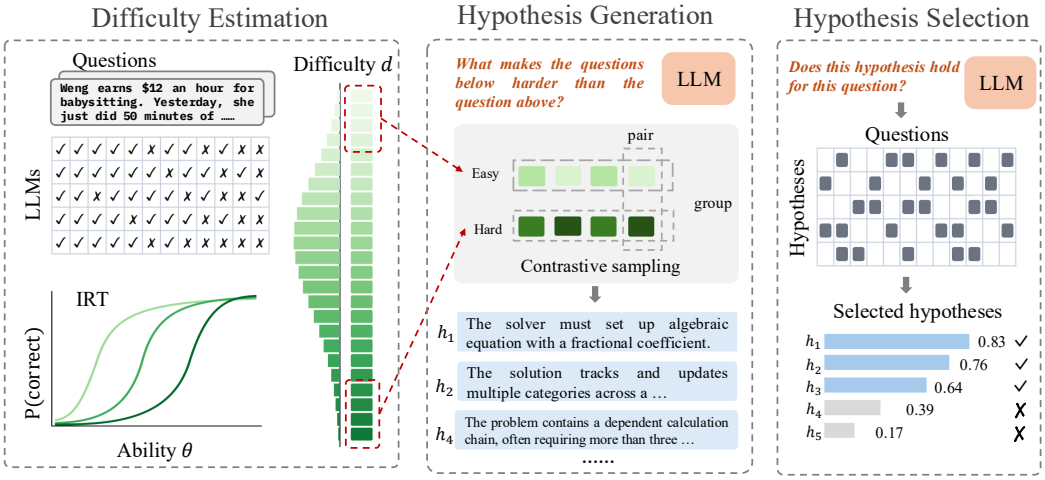

Figure 1: We first estimate question difficulty from LLMs' responses using an IRT model. We then present an LLM with contrasting pairs or groups of hard and easy questions and prompt it to propose candidate hypotheses that explain the difference between them. Finally, we retain the hypotheses that best explain the difficulty of held-out questions.

culty from representations that are themselves uninterpretable, and therefore provide little insight into what makes a question difficult.

On the other hand, a recent line of work uses LLMs to generate natural-language explanations for surface regularities in unstructured data (Zhou et al., 2024; Zhong et al., 2022; 2023; 2024; Hollmann et al., 2023). These methods typically group data that are likely to share a property, and then prompt an LLM to describe a pattern the group has in common. So far, these methods have mostly been applied to properties that are semantically explicit and directly observable in the text, for example, the topic, stance, or style (e.g., *"is about sports"*, *"is written in the first person"* (Zhong et al., 2022). Such properties are relatively easy to discover. In this work, we target *difficulty*, a property that is latent rather than annotated, continuous rather than categorical, and the product of many interacting factors, and that is correspondingly harder to explain. We first estimate item difficulty with an IRT model from the responses of a large pool of LLMs. We then repeatedly sample contrasting data of easy and hard questions and present them to an LLM to propose plausible explanations of the difference. We explore various sampling strategies to make the discovery more effective, and validate and select the most informative hypotheses on a held-out set of questions[1].

We apply our approach to generate hypotheses for various benchmarks, including mathematical reasoning, logical deduction, and commonsense reasoning. Experimental results show that our approach can generate higher-quality and valid explanations than other baseline generators. Individually, our generated explanations separate difficulty more sharply than explanations from baseline approaches, and collectively they predict difficulty of unseen questions as well as or better than finetuned difficulty predictors and few-shot LLM judges. Moreover, incorporating the generated explanations as features into these predictors further improves the performance, indicating that they capture difficulty-related information the existing models can not. Finally, we use the hypotheses to curate synthetic questions: we rewrite questions so that they satisfy or violate a given hypothesis and observe that their measured difficulty shifts in the expected direction. This demonstrates that the discovered hypotheses are not merely post-hoc descriptions but carry causal effect, which yields practical guidance for designing and iterating on benchmarks.

2 Method

Our method consists of three steps. First, we use IRT to estimate each question's difficulty. Second, we construct contrastive samples of hard and easy questions and prompt LLMs to generate

[1]We defer a detailed discussion of related work to Appendix A

candidate hypotheses explaining the difficulty gap. Third, we refine and select from these candidate hypotheses. The full pipeline is illustrated in Figure 1.

2.1 Question Difficulty Estimation

Let $\mathcal{D}$ be a dataset of N questions. To measure the difficulty of each question in $\mathcal{D}$, we first collect M LLMs' responses to the questions, forming a response matrix $R \in \{0,1\}^{M \cdot N}$, where each entry $R_{i,j} = 1$ indicates that the i-th model correctly answered the j-th question, and 0 otherwise.

We then use this matrix R to fit an IRT model, which provides an invariant and noise-aware measurement of item difficulty. Specifically, we use the one parameter logistic (1PL, Rasch) model (Rasch, 1960), where each item's difficulty and each test-taker's (LLM's) ability are learned jointly:

$$P(R_{ij} = 1) = \frac{1}{1 + \exp(-(\theta_i - b_j))}, \tag{1}$$

where θ_i is the i-th model's ability, b_j is the j-th question's difficulty.

2.2 Hypothesis Generation

We split the dataset into three disjoint sets: a generation set $\mathcal{G}$ used to generate candidate hypotheses, a verification set $\mathcal{V}$ to verify and select them, and a held-out set $\mathcal{T}$ for the final evaluation. Each set is constructed via difficulty-stratified sampling, ensuring coverage across the full difficulty range. Our core idea is to present LLMs with *contrastive* sets of difficult and easy items, and instruct an LLM to generate hypotheses explaining the difficulty gap between them. We repeat this sample–generate procedure for multiple rounds to collect a sufficient pool of candidate hypotheses.

Contrastive Sampling. We cut G into 10 difficulty-rank buckets $\mathcal{B}_{\{1\cdots10\}}$ of comparable size. For round t, let $n_{(\cdot)}^{<t}$ denote the number of times a bucket pair, a question, or a question pair has been sampled in earlier rounds. We first draw a bucket combination $(\mathcal{B}_h, \mathcal{B}_e)$ with:

$$w_t(\mathcal{B}_h, \mathcal{B}_e) \;\propto\; (h-e)^{\rho} \left(1 + n_{(\mathcal{B}_h,\mathcal{B}_e)}^{<t}\right)^{-\gamma_c}, \qquad h - e \geq \delta, \tag{2}$$

where $w(\cdot)$ is the sampling weight, h, e index the high and low difficulty buckets, and δ is the minimum gap. ρ tilts the draw toward larger difficulty gaps, which are otherwise underrepresented because few bucket combinations are far apart, and γ_c discourages combinations already used. We then draw k pairs $(q_h, q_e) \in \mathcal{B}_h \times \mathcal{B}_e$, sampled without replacement from all candidate pairs that have not been used in this round, with:

$$w_t(q_h, q_e) \;\propto\; \left(1 + n_{q_h}^{<t}\right)^{-\gamma_i} \left(1 + n_{q_e}^{<t}\right)^{-\gamma_i} \left(1 + n_{(q_h,q_e)}^{<t}\right)^{-\gamma_p} \tag{3}$$

Generation. We construct two forms of prompt, one by pair and one by group. For the pair form, we present the sampled question pairs (q_h, q_e) to the LLM directly and instruct it to explain the difficulty gap within each individual pair. For the group form, we reorganize the sampled pairs into a hard group and an easy group, and instruct the LLM to generate general hypotheses that explain the difference between the two groups and cover as many questions as possible. We combine the two strategies to discover difficulty factors at different granularities.

For the group form, we show 40 questions at a time (20 hard and 20 easy); for the pair form, we show 5 pairs at a time. For both prompting formats, we provide the question text, its reference answer, and its difficulty for every question shown. Although the two forms are intended to capture different granularities, we explicitly instruct the model in both cases to avoid generating overly specific hypotheses. See our prompts in Appendix B.

2.3 Hypothesis Selection

Refinement. We repeat the above sample-and-propose process for K rounds and obtain a list of candidate hypotheses. We compute pairwise similarities between hypotheses using cosine similarity over Qwen3-Embedding-4B embeddings. Hypotheses whose similarity exceeds 0.85 are grouped and passed to an LLM for deduplication and refinement: given a group of semantically similar or overlapping hypotheses, the LLM either retains those that are semantically distinct or merges near-equivalent ones into a single new hypothesis.

Match. We first prompt an LLM (GEMINI-3.7-FLASH) to annotate, for every refined hypothesis, whether each question satisfies it, yielding three binary matching matrices $U^{\mathcal{G}} \in \{0,1\}^{|\mathcal{G}|\times|\mathcal{H}|}, U^{\mathcal{V}} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{H}|}$ and $U^{\mathcal{T}} \in \{0,1\}^{|\mathcal{T}|\times|\mathcal{H}|}$, where $U_{i,j} = 1$ indicates that the i-th question satisfies the j-th hypothesis, and 0 otherwise. These matrices will be used for the subsequent selection and evaluation.

Selection. Finally, we select the most informative hypotheses from the refined candidates $\mathcal{H}$ on the validation set $\mathcal{V}$ using L_1-regularized regression. We fit a separate LASSO on each of 100 random subsamples containing 75% of $\mathcal{V}$, with the L1 penalty weight chosen by 5-fold cross-validation within each subsample, and retain the top hypotheses by selection frequency.

3 EXPERIMENTAL SETUP

3.1 DATASETS

We evaluate our method on three benchmarks across various tasks: (1) **GSM8K** (Cobbe et al., 2021) ($N = 1319$) is a widely used benchmark of grade-school math word problems that require multi-step arithmetic reasoning; (2) **BBH-structured** ($N = 1396$) comprises six subtasks in the BBH dataset (Suzgun et al., 2023) that require logical reasoning: *"logical deduction (three, five, and seven)"*, *"reasoning about colored objects"*, *temporal sequences*, and *"penguins in a table"*; (3) **Winogrande** (Sakaguchi et al., 2021) ($N = 1267$) is a dataset to test commonsense reasoning with pronoun resolution problems while minimizing the surface-level statistical biases present in the original Winograd Schema Challenge (Levesque et al., 2012). The difficulty of this dataset is known to be difficult to predict (Ding et al., 2024). We use it as a challenging testbed to test the limit of baselines and our method. We collect LLM response records from ROUTEREVAL (Huang et al., 2025), including $5,000$ LLMs for GSM8K and WinoGrande, and $3,811$ LLMs for BBH-structured. We fit a 1PL IRT model implemented in py-irt (Lalor & Rodriguez, 2023), which estimates the model by stochastic variational inference.

3.2 EVALUATION

We evaluate the quality of the generated hypotheses along two dimensions:

- *Predictive effect* – to what extent do they explain and predict the difficulty of unseen questions? We evaluate the selected hypotheses in two ways. The first is an *in-sample* evaluation, in which we use OLS to fit difficulty directly on the test pool T using its matching matrix U^T. This measures the goodness of fit of the hypotheses as features and their explanatory power on the data; we report the adjusted R^2 (Ezekiel, 1930), which corrects for the upward bias of R^2 as the number of predictors grows. The second is an *out-of-sample* evaluation, in which we fit an OLS model on $U^{\mathcal{G}\cup\mathcal{V}}$ and predict the difficulty of the questions in T. This setting tests whether the features can predict the difficulty of new questions.
- *Causal effect* – does editing a question according to a hypothesis change its difficulty in the expected direction? For a given hypothesis h, we edit questions according to h, modifying only the attribute that h describes while keeping the rest of the question unchanged. We check whether the measured difficulty shifts significantly in the direction predicted by h using *difficulty shift* (Δb_h) and *success rate* (SR_h), defined as:

$$\Delta b_h = \mathbb{E}_{q\sim\mathcal{D}_h}[b(q'_h) - b(q)], \quad \mathrm{SR}_h = \mathbb{E}_{q\sim\mathcal{D}_h}[\mathbf{1}[b(q'_h) > b(q)]], \tag{4}$$

  where $\mathcal{D}_h$ is the set of questions that can be meaningfully edited according to h, $b(\cdot)$ is a question's difficulty, q'_h is the rewritten question. While Δ_h captures the magnitude of the effect, SR_h measures the fraction of edits that move difficulty in the expected direction.

3.3 BASELINES

We compare our method with the following hypothesis generation baselines:

- **NLPARAM** (Zhong et al., 2024) learns natural-language predicates as model parameters by optimizing a continuous relaxation of each predicate in embedding space and then prompting an

Table 1: Top five hypotheses by our method on GSM8K, BBH-structured, and Winogrande. SEP is the difference in mean IRT difficulty between matched and unmatched test items (95% bootstrap CI); Cohen's d is the standardized effect size, where ‡ marks a large effect ($d \geq 0.8$) and † a medium effect ($d \geq 0.5$) (Cohen, 2013); Coverage is the matched-item percentage. The full set of hypotheses generated by each baseline on each dataset is provided in Appendix E.

#	Top-5 Generated Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	GSM8K			
1	*The problem requires multistep bookkeeping in which intermediate results feed subsequent calculations.*	$1.498_{[1.043,\ 1.959]}$	$0.899^{\ddagger}$	51.0
2	*The solver must derive a rate, ratio, or percentage rather than directly apply one.*	$2.035_{[1.341,\ 2.742]}$	$1.174^{\ddagger}$	9.0
3	*The solver must set up and solve an algebraic equation rather than perform only sequential arithmetic.*	$1.428_{[0.520,\ 2.320]}$	$0.790^{\dagger}$	4.0
4	*The reference answer contains a logical error or contradiction that conflicts with the correct solution.*	$3.696_{[3.043,\ 4.351]}$	$2.063^{\ddagger}$	1.0
5	*The solver must aggregate multiple time intervals or convert between time units to obtain a final duration.*	$1.057_{[-0.238,\ 2.519]}$	$0.584^{\dagger}$	6.5
	BBH-structured			
1	*The solver must order at least five entities from relational constraints or attributes.*	$1.429_{[1.157,\ 1.695]}$	$1.419^{\ddagger}$	23.6
2	*The solver must aggregate or transform multiple elements rather than retrieve one stated value.*	$0.449_{[0.192,\ 0.720]}$	0.387	27.5
3	*The answer choices contain many plausible alternatives rather than a small, easily separable set.*	$0.645_{[0.336,\ 0.953]}$	$0.566^{\dagger}$	29.3
4	*The requested relative position or rank depends on reconstructing the full ordering.*	$1.345_{[1.029,\ 1.643]}$	$1.264^{\ddagger}$	16.8
5	*The solver must filter a set and then aggregate or count the remaining elements.*	$0.768_{[0.332,\ 1.238]}$	$0.664^{\dagger}$	8.2
	Winogrande			
1	*Resolving the blank requires a multistep causal chain with an unstated intermediate state or action.*	$0.916_{[0.446,\ 1.386]}$	$0.559^{\dagger}$	26.4
2	*Resolving the blank requires implicit physical or spatial commonsense about object interactions.*	$0.739_{[0.221,\ 1.258]}$	0.445	20.9
3	*The sentence describes interpersonal causality and requires separating roles such as cause and experiencer.*	$0.742_{[0.199,\ 1.287]}$	0.447	19.7
4	*The sentence contains a contrast or concession whose resolution crosses the contrastive boundary.*	$0.596_{[0.062,\ 1.177]}$	0.356	17.3
5	*A contrastive or concessive marker requires inverting the expected relationship or attribute.*	$0.111_{[-0.507,\ 0.771]}$	0.066	13.8

LLM to verbalize it, iteratively replacing the least useful predicate. We use its classification variant with the logistic layer replaced by linear regression on IRT difficulty.

- **HYPOTHESESAES** (Movva et al., 2025) which applies SAEs to generate hypotheses for explaining data labels with the following pipeline: examples are first mapped to embeddings, an SAE turns the embeddings into sparse interpretable features, the features most predictive of the target are selected, and each selected feature is verbalized into a hypothesis. We obtain embeddings from the hidden-state activations of GEMMA-3-4B; because the evaluation set we use is not sufficient to finetune an SAE, we use the pretrained SAEs from Gemma-Scope-2, and have an LLM (GEMINI-3.7-FLASH) refine the verbal description of each selected feature using the questions from our dataset on which it activates.
- **HYPOGENIC** (Zhou et al., 2024) generates hypotheses from labeled examples with an LLM and improves them iteratively using a bandit-style reward. As HYPOGENIC is defined for classification, we adapt it to difficulty by discretizing IRT difficulty into 3 buckets and treating level prediction as the classification task.

In addition, to measure the predictive power of the generated hypotheses when used as features, we compare our method against a range of difficulty predictors: (1) **Fine-tuned models**, for which we use RoBERTa (Liu et al., 2019), a competitive model for this task as reported in Yaneva et al. (2024); (2) **Frozen representations** with a trained adapter, for which we take the last-token representation of GEMMA-3-4B (Team et al., 2025) and QWEN3-EMBEDDING-4B (Zhang et al., 2025) and train a Lasso regressor on top (a dense regressor would otherwise overfit as the dimension of embeddings is greater than the number of examples); (3) **Few-shot prompting** of frontier LLMs, for which we use GEMINI-3.1-PRO-PREVIEW and CLAUDE-SONNET-5. We split the difficulty range into 5 equal-frequency buckets and randomly draw 2 questions with their answers from each bucket

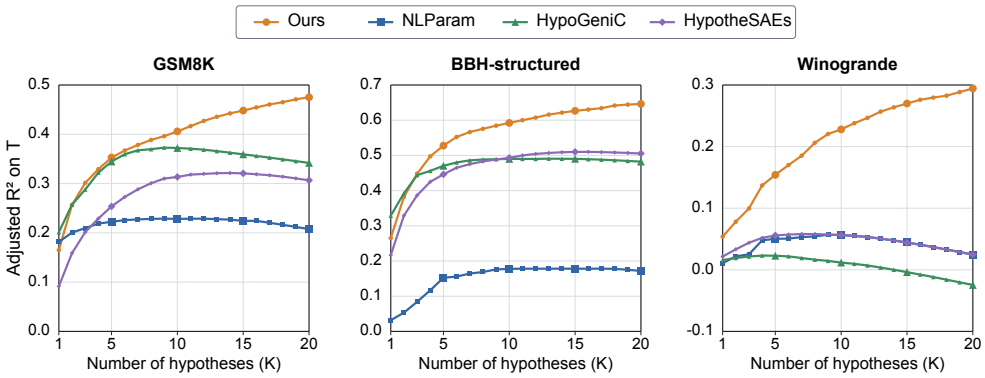

Figure 2: We fit the test set using hypotheses generated by different methods as features to test their explanatory power, and report the adjusted R^2 for varying numbers of hypotheses.

as demonstrations; (4) **Hypothesis-based predictors**, for which we first use a judge model to determine whether each question satisfies each hypothesis as in Section 2.3, and then train an OLS regressor on these binary features. The implementation details of our method and all baselines are provided in Appendix C.

4 Hypothesis Evaluation

4.1 Main Results

What makes a question hard? We present top hypotheses generated by our method in Table 1. Every hypothesis appears plausible and interpretable. Most of them concern the cognitive demands that the corresponding task actually imposes on the solver—the operations that have to be carried out and the information that has to be held and manipulated—rather than surface features of the text such as question length. On GSM8K, our method also captures issues of data quality: items whose reference answer is itself flawed, which affect only a small fraction of the data but carry a very large difficulty effect. In terms of the statistical measures, the 95% CI of SEP has a lower bound above zero for 13/15 hypotheses, showing that they genuinely discriminate between hard and easy items. Except on Winogrande, whose difficulty is known to be hard to perceive even for humans (Ding et al., 2024), 9/10 hypotheses reach at least a medium effect†, and half of them reach a large effect‡.

Our approach discovers more effective hypotheses. To measure how much of the difficulty of unseen questions our generated hypotheses can explain, we use them as features to fit the difficulty scores on the test set. We evaluate the goodness of fit using adjusted R^2. The results of our approach and baselines are presented in Figure 2. We rank the features by greedily picking, at each step, the hypothesis that yields the largest incremental R^2, so that at any given number of features we can compare how the top-k hypotheses perform collectively. On three datasets, our method achieves the best fit at most hypothesis budgets. More importantly, the baselines saturate quickly: they yield only 5–10 effective hypotheses, beyond which adjusted R^2 declines, indicating that the additional hypotheses carry noise rather than real signal. Our method, in contrast, keeps producing informative hypotheses. We attribute this to the fact that the baselines generate patterns directly from the data and therefore discover only the most salient signals, whereas our sampling strategy allows the model to explore new difficulty factors by repeatedly contrasting difficult and easy questions.

The generated hypotheses predict the difficulty on par with black-box predictors. We now examine how well H predicts the difficulty of unseen questions. We train an OLS regressor on $\mathcal{G}$ and $\mathcal{V}$ using the hypotheses as features and evaluate its R^2 on the test set, comparing our method against the baseline difficulty predictors introduced in Section 3.3. The results are presented in Table 2. Overall, our model achieves the best performance on two datasets, GSM8K and WinoGrande, and the second best on BBH-structured, showing that these features alone already constitute a strong difficulty predictor. We note that all models perform poorly on WinoGrande. This again echoes the

Table 2: Performance of different difficulty prediction methods on three datasets. All methods are trained on $\mathcal{G} \cup \mathcal{V}$, and we report R^2 and RMSE on $\mathcal{T}$. Results of ROTBERTA-BASE are averaged over 5 seeds, and prompting of LLMs over 3 set of demonstrations. The number of hypotheses $|H|$ for the baselines is selected on the verification pool over a sweep of $\{10, 20, 50\}$. Subscripts denote standard deviations across seeds. Lasso and OLS are deterministic given the data, so no s.d. is reported for them. Best results are in bold and second-best are underlined.

Approaches	GSM8K		BBH-structured		Winogrande	
	$R^2 \uparrow$	RMSE $\downarrow$	$R^2 \uparrow$	RMSE $\downarrow$	$R^2 \uparrow$	RMSE $\downarrow$
ROBERTA-BASE$_{\text{finetuned}}$	<u>0.362</u> $_{0.02}$	<u>1.454</u> $_{0.02}$	**0.646** $_{0.01}$	**0.698** $_{0.01}$	-0.038	1.711
GEMMA-3-4B$_{\text{last-token}}$	0.191	1.637	0.554	0.783	-0.019	1.697
QWEN3-EMBEDDING-4B	0.200	1.629	0.533	0.802	-0.000	1.681
GEMINI-3.1-PRO$_{\text{10-shot}}$	0.346 $_{0.03}$	1.472 $_{0.04}$	0.139 $_{0.172}$	1.085 $_{0.107}$	-0.134 $_{0.14}$	1.788 $_{0.11}$
CLAUDE-SONNET-5$_{\text{10-shot}}$	0.041 $_{0.06}$	1.782 $_{0.05}$	0.043 $_{0.284}$	1.139 $_{0.165}$	<u>0.080</u> $_{0.03}$	<u>1.618</u> $_{0.03}$
HYPOTHESAE $_{\lvert\mathcal{H}\rvert=\{50,50,50\}}$	0.351	1.467	0.477	0.848	-0.160	1.810
HYPOGENIC $_{\lvert\mathcal{H}\rvert=\{20,20,20\}}$	0.327	1.494	0.465	0.861	-0.004	1.684
NLPARAM$_{\lvert\mathcal{H}\rvert=\{20,20,20\}}$	0.240	1.587	0.088	1.120	0.015	1.668
Ours $_{\lvert\mathcal{H}\rvert=\{40,12,44\}}$	**0.373**	**1.441**	<u>0.580</u>	<u>0.760</u>	**0.090**	**1.603**

finding that the difficulty of WinoGrande is hard to predict (Ding et al., 2024)[2]. Nevertheless, our method still extracts some signal on this dataset, while all baselines except CLAUDE-SONNET-5 fail completely, attaining R^2 values around or even below 0.

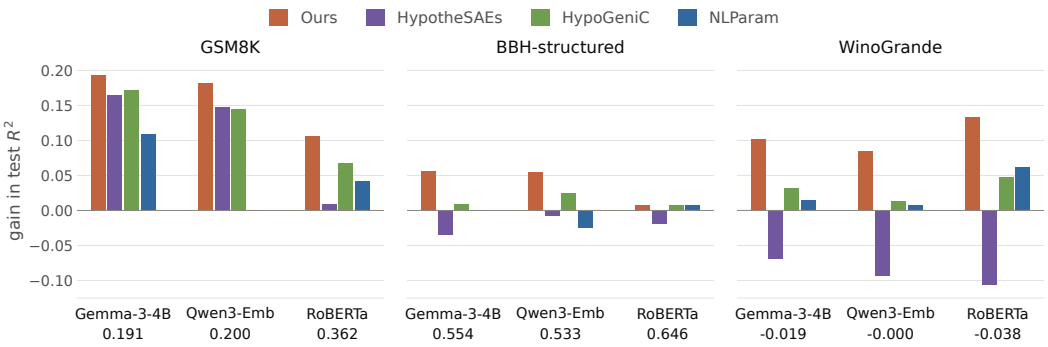

Figure 3: Comparison of R^2 gains from incorporating hypotheses generated by different methods into different models.

The hypotheses capture difficulty signals the baselines miss. So far we have compared our hypothesis-based difficulty predictor against state-of-the-art models. In practice, however, our method is complementary to them rather than competing with them: the hypotheses it discovers can be supplied to any predictor as additional features. We test this on three semantic predictors–fine-tuned ROBERTA-BASE, GEMMA-3-4B and QWEN3-EMBEDDING-4B—by adding each method's hypothesis match vector as a second feature block. Figure 3 reports the gain in test R^2. Our hypotheses give the largest gain of any hypothesis source in all settings. The gains are largest on GSM8K: $+0.19$ and $+0.18$ over the two frozen embeddings, and $+0.11$ even over the fine-tuned RoBERTa ($0.362 \rightarrow 0.468$), by far the strongest baseline there. On BBH-structured the gain over the strongest model is small: RoBERTa-base already reaches $R^2 = 0.646$ there, and our hypotheses add only $+0.008$ on top of it (they still add $+0.05$ to the two frozen embeddings), implying that on this dataset the difficulty signal is largely already captured by the encoder itself. On WinoGrande, every black-box predictor sits at zero (R^2 between -0.038 and -0.000), yet our hypotheses lift each of them to R^2 between 0.08 and 0.10.

[2]On Winogrande, human annotators agree with the IRT-estimated difficulty on only 73.4% of item pairs. Under a bivariate-normal assumption, this agreement level implies $R^2 \approx 0.45$ as a ceiling.

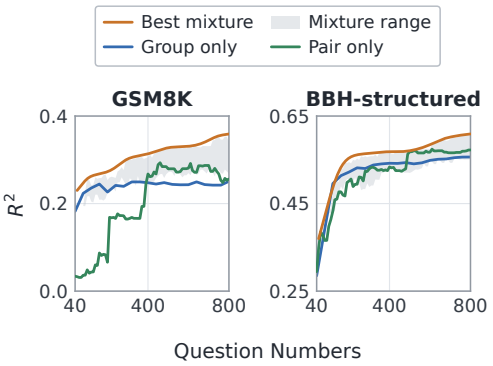

Figure 4: Group and pair sampling on GSM8K and BBH-structured. Shading shows the range across mixed allocations; orange marks the best mixture at each question budget.

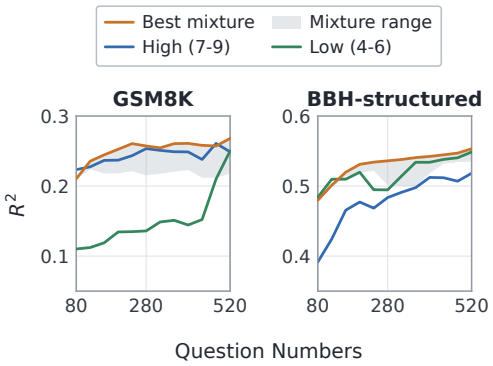

Figure 5: Contrast-strength ablation on GSM8K and BBH-structured. High and low use bucket gaps of 7–9 and 4–6. Shading shows the mixture range; orange marks its upper bound.

4.2 Impact of Sampling Strategy

Combining group and pair contrasts outperforms either alone. Figure 4 compares three prompting strategies: group-only, pair-only, and a mixture of group and pair prompts. The first two are designed to discover hypotheses at different granularities. For the mixture, we keep the total number of questions fixed and vary the proportion of group versus pair prompts for fair comparison, with the range of results across these proportions shown in the shaded region. On GSM8K, group-only converges faster, although pair-only reaches comparable or better performance with a larger budget. On BBH, the two perform similarly. On both datasets, however, interpolating between them under the same budget yields better results in most cases, scales better as the budget grows, and reaches a higher upper bound (orange line). This indicates that mixing the two strategies indeed discovers complementary hypotheses.

Larger gaps are not always better. Figure 5 compares the effect of contrast strength: sampling high-contrast questions (bucket gap $\delta \geq 7$; Eq. 2), sampling low-contrast questions (bucket gap $4 \leq \delta \leq 6$), and mixing the two. The former is designed to capture more pronounced difficulty patterns, the latter more nuanced signals. The two strategies behave in opposite ways on the two datasets: high contrast is the stronger choice on GSM8K, while low contrast wins on BBH-structured. The mixed strategy serves as a more stable option: it does not beat the better single strategy at every budget, but it stays close to it in most cases. We therefore recommend the mixed strategy as the more reliable option.

Summary. Overall, with a suitable sampling and prompting strategy, our method reaches a good level of performance within 200 questions (a proxy for the annotation budget) and retains the potential to scale further when the budget allows, as none of the mixed-strategy curves in Figure 4 has fully converged. This echoes the results in Figure 2, and again indicates the method's ability to keep discovering diverse and effective hypotheses.

5 Causal Analysis

Counterfactual question editing. We test whether a hypothesis captures a difficulty-bearing property through targeted question editing (see examples in Appendix Table 5). Because the difficulty signal in WinoGrande is inherently weak, we focus here on GSM8K and BBH-structured. We randomly sample 50 test questions, determine which hypotheses apply to each of them, and prompt an LLM to edit and verify the questions, changing only what the hypothesis describes while leaving the rest intact. Details are given in Appendix C.4.

Table 3: Hypothesis-guided rewriting results on both datasets. Orig. and Rewrite are mean accuracies (%) across pairs and models; $\Delta\mathrm{Acc}_{\mathcal{H}}$ and $\mathrm{SR}_{\mathcal{H}}$ are the accuracy shift and success rate (counts in parentheses) averaged over all hypotheses, respectively (Eq. 4).

	GSM8K				BBH-structured			
Intervention	Orig.	Rewrite	$\Delta\mathrm{Acc}_{\mathcal{H}}$	$\mathrm{SR}_{\mathcal{H}}$	Orig.	Rewrite	$\Delta\mathrm{Acc}_{\mathcal{H}}$	$\mathrm{SR}_{\mathcal{H}}$
Increase difficulty	45.01	29.75	-15.27	70.9% (100/141)	60.63	36.75	-23.88	77.6% (52/67)
Decrease difficulty	25.61	57.79	$+32.18$	92.2% (47/51)	36.82	65.34	$+28.52$	85.5% (47/55)

Evaluation protocol. We evaluate the causal effect using the difficulty shift and the success rate defined in Section 3.2. Measuring the difficulty of a rewritten question would require re-evaluating it on thousands of LLMs, which is prohibitively costly. We therefore re-run only a subset of LLMs, chosen to cover a range of performance levels (from below 10% to above 80% accuracy) and limited to models under 30B due to the restriction of our compute resource. A reliable difficulty re-estimation is not attainable from this subset, so throughout this section we use accuracy as a proxy for difficulty[3]. Because each hypothesis applies to a different set of questions (those it can actually rewrite), and induces a different kind of edit, comparing hypotheses against one another in a controlled, fair setting is difficult, which in turn makes a comparison against the baselines difficult. In this section we therefore focus on validating the effectiveness of our own hypotheses.

Results. In Table 3 we report the difficulty shift and success rate averaged over all hypotheses, which shows consistent bidirectional changes on both datasets. Increase-difficulty rewrites reduce mean accuracy by 15.27 and 23.88 percentage points on GSM8K and BBH-structured, respectively, while decrease-difficulty rewrites improve it by 32.18 and 28.52 points. Question-level alignment reaches 70.9% and 92.2% for increase and decrease rewrites on GSM8K, compared with 77.6% and 85.5% on BBH-structured. Thus, the aggregate changes are accompanied by directional consistency across most rewritten questions, rather than only a few large shifts.

We present the full model list and per-model results in Appendix Table 6). It shows that these patterns extend across model families: 14/17 GSM8K models and $15/16$ BBH-structured models lose accuracy after increase rewrites, whereas $16/17$ and $16/16$ gain accuracy after decrease rewrites. The changes are not confined to weak models. On BBH-structured, YI-1.5-9B-CHAT falls from 88.1% to 37.3% after increase rewrites; on GSM8K, LLAMA-3-8B-INSTRUCT rises from 43.1% to 94.1% after decrease rewrites. Conversely, all three models with reversed increase-direction changes start below 15% accuracy on the corresponding original questions. On the one hand, their poor performance leaves little room for a further drop. On the other hand, most of them probably lack genuine reasoning ability and perform close to random guessing, so they may be insensitive to changes in difficulty. Finally, Table 7 provides the per-hypothesis results, where we can see that most individual hypotheses also achieve a solid difficulty shift and success rate. Together, these results support the relevance of the targeted properties to difficulty across models, while showing that individual models need not respond uniformly.

6 CONCLUSION

In this work, we presented a hypothesis generation method for explaining what makes a question difficult. We measure difficulty with IRT from the responses of thousands of LLMs, sample contrasting sets of easy and hard questions, and prompt an LLM to propose natural-language hypotheses. Experiments on three datasets show that the resulting hypotheses are interpretable and of high quality. They capture difficulty signals that semantic encoders do not, predict the difficulty of unseen questions as well as or better than black-box predictors, and further improve those predictors when added as features. Beyond prediction, they also guide the editing of questions to manipulate their difficulty in a controlled direction, which suggests that they identify genuine sources of difficulty rather than post-hoc correlations. We hope this offers a step toward benchmarks whose difficulty can be understood and designed, rather than only measured.

[3]Under the 1PL model an item's mean accuracy over a fixed respondent set is a strictly monotone function of its difficulty, so the direction of a within-panel accuracy change matches the direction of the difficulty shift.

AI USE STATEMENT

In this work, we used generative models for part of the code implementation and execution, as well as for language polishing of some sentences and paragraphs. We have carefully reviewed all AI-generated content: the code was tested and its outputs were checked against the experiment logs, and all AI-polished text was thoroughly reviewed. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the prompts used in our study in Appendix B and all the implementation details in Appendix C. These appendices document the instructions and experimental configurations underlying our results.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Suppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Q Knight, Yiyang Nan, Jyoutir Raj, Yu Fan, Shubham Singh, Subramanyam Sahoo, Eliya Habba, Usman Gohar, Siddhesh Milind Pawar, Robert Scholz, Arjun Subramonian, Jingwei Ni, Mykel Kochenderfer, Sanmi Koyejo, Mrinmaya Sachan, Stella Biderman, Zeerak Talat, Avijit Ghosh, and Irene Solaiman. When AI benchmarks plateau: A systematic study of benchmark saturation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YC1OtscjbS.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pp. 412–421, 2020.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37, 2023.

Paul Boeck and Mark Wilson. *Explanatory item response models: A generalized linear and nonlinear approach*. Springer, 2004.

Matthew Byrd and Shashank Srivastava. Predicting difficulty and discrimination of natural language questions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 119–130, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.15. URL https://aclanthology.org/2022.acl-short.15/.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1550–1553. IEEE, 2019.

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Peng Cui and Mrinmaya Sachan. Adaptive and personalized exercise generation for online language learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10184–10198, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.567. URL https://aclanthology.org/2023.acl-long.567/.

Peng Cui and Mrinmaya Sachan. Investigating the zone of proximal development of language models for in-context learning. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6485–6498, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.362. URL https://aclanthology.org/2025.findings-naacl.362/.

Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1962–1981, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533240. URL https://doi.org/10.1145/3531146.3533240.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, et al. Easy2hard-bench: Standardized difficulty labels for profiling llm performance and generalization. *Advances in Neural Information Processing Systems*, 37:44323–44365, 2024.

Andreea Dutulescu, Stefan Ruseti, Mihai Dascalu, and Danielle Mcnamara. How hard can this question be? an exploratory analysis of features assessing question difficulty using llms. In Benjamin Paaßen and Carrie Demmans Epp (eds.), *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 802–808, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society. ISBN 978-1-7336736-5-5. doi: 10.5281/zenodo.12729956.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Mordecai Ezekiel. *Methods of correlation analysis*. J. Wiley & Sons, Incorporated, 1930.

Gerhard H Fischer. The linear logistic test model as an instrument in educational research. *Acta psychologica*, 37(6):359–374, 1973.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ronald K Hambleton and Hariharan Swaminathan. *Item response theory: Principles and applications*. Springer Science & Business Media, 2013.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of item response theory*, volume 2. Sage, 1991.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=mxcCg9YRqj.

Noah Hollmann, Samuel Müller, and Frank Hutter. Large language models for automated data science: Introducing CAAFE for context-aware automated feature engineering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9WSxQZ9mG7.

Xingjian Hu, Ziqian Zhang, Yue Huang, Kai Zhang, Ruoxi Chen, Yixin Liu, Qingsong Wen, Kaidi Xu, Xiangliang Zhang, Neil Zhenqiang Gong, and Lichao Sun. EIP: Weighted ranking of LLMs by quantifying question difficulty. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jnX5GJIoYt.

Wenyue Hua, Lifeng Jin, Linfeng Song, Haitao Mi, Yongfeng Zhang, and Dong Yu. Discover, explain, improve: An automatic slice detection benchmark for natural language processing. *Transactions of the Association for Computational Linguistics*, 11:1537–1552, 2023.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 1352–1359. AAAI Press, 2017.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. RouterEval: A comprehensive benchmark for routing LLMs to explore model-level scaling up in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3860–3887, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.208. URL https://aclanthology.org/2025.findings-emnlp.208/.

Rianne Janssen, Jan Schepers, and Deborah Peres. Models with item and item group predictors. In *Explanatory item response models: A generalized linear and nonlinear approach*, pp. 189–212. Springer, 2004.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Ying Jiao, Kumar Shridhar, Peng Cui, Wangchunshu Zhou, and Mrinmaya Sachan. Automatic educational question generation with difficulty level controls. In *International Conference on Artificial Intelligence in Education*, pp. 476–488. Springer, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://mlanthology.org/iclr/2015/kingma2015iclr-adam/.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench - a sparse benchmark of reasoning and knowledge in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4T33izzFpK.

John Patrick Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13, January 2023. ISSN 1526-5528. doi: 10.1287/ijoc.2022.1250. URL http://dx.doi.org/10.1287/ijoc.2022.1250.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pp. 552–561. AAAI Press, 2012. ISBN 9781577355601.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 846–855, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

William Lugoloobi and Chris Russell. Llms encode how difficult problems are. *arXiv preprint arXiv:2510.18147*, 2025.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44997–45023. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/movva25a.html.

Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. Question difficulty estimation for large language models via answer plausibility scoring. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11124–11151, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.510. URL https://aclanthology.org/2026.acl-long.510/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating LLMs with fewer examples. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=qAml3FpfhG.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, 1960.

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

Svetlana Sagadeeva and Matthias Boehm. Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In *Proceedings of the 2021 international conference on management of data*, pp. 2290–2299, 2021.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.71. URL https://aclanthology.org/2023.findings-acl.71/.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Wim J Van der Linden, Cees AW Glas, et al. *Computerized adaptive testing: Theory and practice*, volume 13. Springer, 2000.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 470–482, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.bea-1.39/.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27099–27116. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhong22a.html.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PnbCA4ylIc.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=u5BkOgWWZW.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart (eds.), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://aclanthology.org/2024.nlp4science-1.10/.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. How to select datapoints for efficient human evaluation of NLG models? *Transactions of the Association for Computational Linguistics*, 13: 1789–1811, 2025. doi: 10.1162/tacl.a.60. URL https://aclanthology.org/2025.tacl-1.80/.

A Related Work

Difficulty Modeling. Question difficulty is a central quantity wherever questions are used to measure ability. In educational testing it determines which items enter an exam, how scores are compared across test forms, and which item an adaptive test presents next (Van der Linden et al., 2000); it also has to be known for newly written questions before they can be administered, which has motivated predicting and controlling the difficulty of generated items (Cui & Sachan, 2023; Jiao et al., 2023). In LLM evaluation it has become equally consequential: benchmarks are filtered or rebuilt around the items that remain hard (Kipnis et al., 2025; Zouhar et al., 2025), and item-level difficulty is used to pick informative subsets, to test models adaptively at a fraction of the cost (Hofmann et al., 2025; Polo et al., 2024), or to weight items when ranking models by competency (Hu et al., 2026). How difficulty is obtained falls into two families. They first estimates it from responses: descriptive IRT models treat each item's difficulty as a free parameter, while explanatory item response models constrain it to be a function of item covariates, classically a linear combination of the cognitive operations an item requires (Fischer, 1973; Janssen et al., 2004). These covariates are specified by experts, so the explanation is only as complete as the expert's list; our approach can be seen as an explanatory model whose covariates are discovered from data. The second family predicts difficulty from the item text alone. Early work relied on hand-designed linguistic and cognitive features (Benedetto et al., 2023; AlKhuzaey et al., 2024), which were then replaced by neural encoders of the item, the approach that dominated the recent shared task on medical exam items (Huang et al., 2017; Loginova et al., 2021; Yaneva et al., 2024); Byrd & Srivastava (2022) extend this line beyond difficulty to discrimination, relating both parameters to textual and contextual properties of the question. With the arrival of LLMs, difficulty has been estimated by prompting them to judge items against IRT-derived levels, by scoring the plausibility of the candidate answers (Mozafari et al., 2026), and by reading it off the hidden states of the answering model itself (Lugoloobi & Russell, 2025). Closest to our goal, Dutulescu et al. (2024) use LLMs to analyze which features are associated with question difficulty; there the features are drawn from a predefined inventory and examined post hoc, whereas we let the model propose candidate explanations, validate each of them on held-out items, and select among them by their joint predictive power.

Automated Error Analysis. A related area, slice discovery, seeks to discover the data on which an AI model is prone to err and to explain the patterns behind those errors, which is in essence also a form of explanation of difficulty. Early work assumed structured inputs and searched over conjunctions of feature values for slices on which a model performs significantly worse than average, with the slice definition itself serving as the explanation (Chung et al., 2019; Sagadeeva & Boehm, 2021); Later methods instead look for coherent, underperforming regions of a model's representation space and attach natural-language descriptions to the regions they find, by matching them to captions in a shared embedding space or by prompting an LLM with their members (d'Eon et al., 2022; Eyuboglu et al., 2022; Hua et al., 2023). What these methods explain, however, is the difficulty an item poses to one particular model. The errors of a single model, however, are likely to confound the difficulty of an item with the weaknesses and biases of that particular model, rather than reflecting an intrinsic property of the item. While effective for model debugging, the weaknesses these methods uncover are therefore unlikely to transfer to other models, and offer limited insight into the broader questions of how a dataset should be curated and iterated on.

Hypothesis generation with LLMs. A growing line of work uses LLMs to describe, in natural language, the patterns shared by a group of data. These methods first partition the data into groups that are likely to share a pattern or behavior, for example, by taking samples with different labels or from different corpora (Zhong et al., 2022; 2023), or by clustering or otherwise exploiting structure in embedding space (Zhong et al., 2024; Movva et al., 2025). They then prompt an LLM with samples from these groups to generate descriptions of what the samples have in common, typically followed by some form of validation of the generated descriptions. Across previous approaches, the property being explained is usually one that is observable in the text or supplied as a label: a topic, stance, or writing style, the aspects that drive a product rating, membership of a given group, or an annotated outcome such as engagement or persuasiveness (Zhong et al., 2022). Our work keeps the propose-and-verify structure of this line but changes the target and the evaluation: the property is a latent, continuous quantity inferred from model responses rather than annotated, the contrasting groups are constructed by sampling from that quantity, and hypotheses are scored by how much of its variance they explain, individually and jointly, on items never used for their discovery.

B PROMPTS

Prompt 1: Group Proposer Prompt

System:

You study why some questions in a benchmark are harder for language models to answer correctly than others. "Harder" is measured, not guessed: every question has an IRT (Item Response Theory) difficulty parameter fit from many models' responses. Each question below is annotated with that parameter: larger values mean harder questions and smaller values mean easier questions. The values are positions on this benchmark's relative IRT scale, not probabilities or percentages.
{Dataset-specific context} You will see two GROUPS of questions from the same benchmark: a HARDER group, whose questions all have much higher measured difficulty, and an EASIER group. Your task is to propose hypotheses about WHAT MAKES A QUESTION HARD in this benchmark: short, testable natural-language descriptions of a property that the harder questions tend to have and the easier ones tend not to, or the reverse.
How to work:

- First read BOTH groups in full, before writing anything. Look for patterns that recur across several questions in the HARDER group and are absent or rarer in the EASIER group, rather than what is special about any one question.
- Use the individual IRT values as graded evidence. Prefer a pattern when questions exhibiting it tend to have larger values across the examples, and do not let one extreme-valued outlier stand in for a recurring pattern.
- Write one hypothesis for each pattern actually found. There is no target count. Do not pad the list.

Guidelines:

1. Stay on difficulty. Each hypothesis must describe something that plausibly makes a question harder or easier to answer correctly. Avoid topic labels and properties specific to one question.
2. Consider, but are not limited to: {Dataset-specific aspects}.
3. Generalize. A hypothesis should be a property that many unseen questions in the benchmark could exhibit and that can be checked without knowing the question's difficulty.
4. Prefer breadth. A pattern shared by several HARDER questions is preferable to one that fits only a single question.
5. Propose distinct hypotheses covering different plausible explanations, not rewordings of one another.
6. State each hypothesis as a specific, checkable yes/no criterion, rather than a vague label such as "is conceptually difficult" or "requires careful reasoning."
7. Keep each hypothesis concise and atomic. If two properties can be judged separately, state them as separate hypotheses rather than joining them with "and" or "or."

Respond with ONLY a JSON object:
`{"hypotheses": ["...", "...", ...]}`

User:

HARDER group (much higher measured difficulty):
{Numbered harder questions with IRT difficulties and reference answers/explanations}
EASIER group (much lower measured difficulty):
{Numbered easier questions with IRT difficulties and reference answers/explanations}
Read both groups first. Then propose the NEW hypotheses you find, one per pattern actually observed—as many or as few as the groups genuinely support, with no fixed count.

Prompt 2: Pair-Moderate Proposer Prompt

System:

You study why some benchmark questions are harder for language models than others. Difficulty is an IRT parameter estimated from many models' responses; larger values mean harder questions.
{Dataset-specific context}
You will see exactly 5 randomly sampled pairs. Within each pair, HARDER has an IRT difficulty at least {minimum gap} logits above EASIER. Pairs were not matched by topic and were not drawn from difficulty buckets.
Read all pairs before answering. Propose short, testable hypotheses about properties that may make questions harder or easier. For each hypothesis, list every pair number that supports it. A hypothesis may be supported by a single pair, but it must describe a reusable property that could naturally occur in unseen questions.
IMPORTANT: Include both broad recurring factors and narrower reusable subtypes when the pairs support them; a subtype does not need to recur across many shown pairs. Prefer concrete yes/no criteria, but do not copy names, objects, answers, incidental constants, distinctive wording, or a complete solution path from a shown question. Do not create a long conjunction whose main purpose is to identify one displayed item. Numerical thresholds are allowed only when they define a reusable complexity class.
Consider, but are not limited to: {Dataset-specific aspects}. Return as many well-supported, non-redundant hypotheses as the evidence justifies, up to 10. Do not pad the list.
Return ONLY valid JSON:

```
{"hypotheses": [{"hypothesis": "...", "pairs": [1, 4]},
                {"hypothesis": "...", "pairs": [2]}]}
```

User:

5 random hard/easy pairs (each IRT gap is at least {minimum gap}):
Pair 1
  HARDER [IRT difficulty: {value}]: {Question with reference answer/explanation}
  EASIER [IRT difficulty: {value}]: {Question with reference answer/explanation}
⋮
Pair 5
  HARDER [IRT difficulty: {value}]: {Question}
  EASIER [IRT difficulty: {value}]: {Question}
Read all pairs first, then return only the required JSON object.

Prompt 3: Match Judge Prompt

System:

You judge whether a benchmark question exhibits each of several candidate properties.
Candidate properties:
1. {Hypothesis 1}
2. {Hypothesis 2}
⋮
K. {Hypothesis K}
For the question given by the user, answer whether it exhibits each candidate property above. When uncertain, answer 0 (do not guess yes).
Respond with ONLY a JSON object:

```
{"scores": [0 or 1, 0 or 1, ...]}
```

Return exactly one integer per candidate property, in the same order they were listed.

User:

Question and reference answer/explanation:
{Item text}
Scores:

C Implementation Details

C.1 Hypothesis Discovery

All generation, feature selection and model fitting take place inside the generation and verification pools; the test pool is used once, for the reported metrics. Two models recur throughout: GEMINI-3.1-PRO-PREVIEW proposes hypotheses, and GEMINI-3.7-FLASH labels items against them. Both run at temperature 0 with JSON-constrained output and the model's lowest thinking setting.

Our method. For each dataset we run 30 rounds of group-based generation and 30 rounds of pair-based generation. Each group sample contains 40 questions, 20 hard and 20 easy, and each pair prompt presents 10 questions in 5 pairs. Figure 4 shows how performance changes with the question budget. The candidate pool is reduced by stability selection with a Lasso base learner: we draw 100 subsamples of 75% the training items, record how often each candidate receives a non-zero coefficient, and keep the most frequently selected ones.

HypoGeniC. The method presumes a classification target, so we discretize difficulty into three equal-frequency levels and map a predicted level back to that level's mean difficulty when regression metrics are needed. Apart from these adaptations, we follow the original implementations as closely as possible. 10 initial items seed ten initial hypotheses. A bank of 20 hypotheses is then maintained by an upper-confidence-bound rule with exploration weight 0.5, the 10 highest-scoring hypotheses vote on each new item, and a pool of ten mispredicted items triggers a generation of five new hypotheses whose rewards are initialized on the most recently scored items. We also ran a stricter variant in which generation only ever sees the generation pool $\mathcal{G}$ and rewards only ever come from the validation pool $\mathcal{T}$. We stop generation once the model has seen a fixed number of questions, matching the number seen by our method, as a proxy for the budget to ensure a fair comparison.

NLParam. We follow the reference implementation closely. Questions are embedded with INSTRUCTOR-XL (Su et al., 2023) exactly as that code embeds them. Predicate directions are initialized by k-means++ and refined by 10 alternations between the regression weights and the relaxed predicate representation, each solved with Adam (Kingma & Ba, 2015) at learning rate 10^{-2} for at most a thousand steps. Refinement then repeats 10 times, each time ablating the least useful predicate, refitting that direction and replacing it with the best of five candidates under the full-data loss. We report the results of 20 predicates.

HypotheSAEs. We use the Gemma-Scope-2[4] autoencoders for GOOGLE/GEMMA-3-4B-PT at the final layer, width 16 thousand, running them over every token of an item and max-pooling per item; features active on fewer than 20 training questions are discarded. Following the original, a Lasso with cross-validated penalty selects features, the non-zero ones are ranked by coefficient magnitude.

C.2 Learned and Frozen Representations

Fine-tuned encoder. We fine-tune ROBERTA-BASE end-to-end with a scalar regression head on the item text (question and answer), optimizing squared error against the standardized difficulty and mapping predictions back to the logit scale before scoring. We use AdamW (Loshchilov & Hutter, 2019) at learning rate 2×10^{-5} with weight decay 0.01, 10% linear warm-up, batch size 16 and a maximum sequence length of 512 tokens, for at most 10 epochs with early stopping on a per-seed tenth of the training pool held out for that purpose, with patience three. Every reported number is the mean over 5 seeds, since end-to-end fine-tuning on roughly a thousand items is seed-sensitive. Training runs on a single NVIDIA RTX 4090 GPU and takes a few minutes per seed.

Frozen representations. GEMMA-3-4B is run in `bfloat16` with a maximum length of $2,176$ tokens and tail truncation, and we take the last-token hidden state of the final layer. QWEN3-EMBEDDING-4B is used through its sentence-embedding interface with ℓ_2-normalised outputs and no instruction prefix. For both models, features are standardized using training-pool statistics, a

[4] `https://huggingface.co/google/gemma-scope-2`

Table 4: Demonstrations used by the prompted judges, listed in increasing order of measured difficulty.

Dataset	Seed	Demonstration item identifiers
GSM8K	0	3344, 3268, 2478, 3542, 3198, 2561, 2761, 3407, 3132, 2894
	1	2823, 2876, 3727, 3336, 3078, 3406, 2555, 2866, 2806, 3139
	2	2946, 2602, 3689, 2950, 3273, 3149, 3272, 3306, 3005, 2880
BBH-structured	0	36661, 34883, 36010, 36121, 34851, 36773, 35797, 36779, 35888, 35503
	1	36235, 36068, 35911, 37867, 36206, 36670, 37800, 37794, 35495, 34780
	2	36664, 37919, 36648, 35527, 35981, 35567, 34958, 36780, 35475, 34813
WinoGrande	0	860, 422, 1006, 894, 139, 1146, 1225, 10, 581, 273
	1	1139, 1054, 918, 744, 100, 1031, 1019, 699, 1124, 135
	2	338, 1083, 881, 499, 292, 255, 215, 1239, 80, 1100

Lasso penalty is chosen by five-fold cross-validation inside the training pool, and the resulting model is evaluated once on the test pool.

C.3 Prompted Judges

The prompted baselines receive no training set. A context of 10 labelled demonstrations is drawn once from the training pool by splitting it into five equal-frequency difficulty buckets and sampling 2 questions from each, and the same context precedes every test item. The system message explains that difficulty is an IRT parameter on a logit scale, gives the benchmark's observed range, mean and standard deviation, and lists the demonstrations with their measured values; the model replies with a single number, which we use unchanged. GEMINI-3.1-PRO-PREVIEW runs at temperature 0 with its lowest thinking setting and JSON-constrained output. The CLAUDE-SONNET-5 API rejects an explicit temperature for this model, so the request omits it and the model decodes at its default setting, with extended thinking disabled; the shared system block is marked for prompt caching. Both judges see item text truncated to $1,500$ characters and are run with three demonstration draws, seeds 0 through 2, over which we report mean and standard deviation. Because the draw depends only on the seed and the pools, both judges see identical contexts. To ensure reproducibility, we provide the identifiers of the sampled questions in Table 4.

C.4 Question Rewriting

We construct hypothesis-guided question edits in three stages. First, we determine whether the original question satisfies the target hypothesis using its binary match label. If the property is present, we request its removal or simplification. If it is absent, an applicability judge assesses whether it can be introduced naturally while preserving the question's broad setting and answer format; inapplicable interventions are skipped. Second, a rewriter produces a separate candidate for each applicable intervention (including questions and reference answers), targeting the specified property while preserving the question's broad setting and answer format. Third, GEMINI and ASTRA verify the candidates in sequence, each solving the rewritten question from scratch and checking its solvability, intended property change, and preservation of the original setting. Malformed or incomplete responses are retried, and only rewrites accepted by both verifiers enter the evaluation.

D MORE RESULTS

Table 5: Bidirectional rewriting for a hypothesis with an explicit entity-count threshold. **Red text** marks additions or replacements; the decrease rewrite also removes the third dog's stated distance. The increase rewrite expands three clothing categories to four, whereas the decrease rewrite reduces four dogs to three. Both preserve the task and answer units.

Hypothesis	*Requires tracking and calculating values for four or more distinct entities or categories.*
Increase: original	Mishka bought 3 pairs of shorts, 3 pairs of pants, and 3 pairs of shoes. One pair of shorts costs $16.50. One pair of pants costs $22.50 and one pair of shoes costs $42. How many dollars did Mishka spend on all the clothing items? *Ans.* $243
Increase: rewrite	Mishka bought 3 pairs of shorts, 3 pairs of pants, 3 pairs of shoes, **and 3 shirts**. One pair of shorts costs $16.50. One pair of pants costs $22.50, one pair of shoes costs $42, **and one shirt costs $15**. How many dollars did Mishka spend on all the clothing items? *Ans.* $288
Decrease: original	Bernie has 4 dogs. They each need a certain amount of exercise per day. The first needs to walk 1 mile. The second needs to walk 4 miles. The third needs to walk 3 miles. On average, they need to walk 3 miles per day. How many miles does the last dog need? *Ans.* 4 miles
Decrease: rewrite	Bernie has **3 dogs**. They each need a certain amount of exercise per day. The first needs to walk 1 mile. The second needs to walk 4 miles. On average, they need to walk 3 miles per day. How many miles does the last dog need? *Ans.* 4 miles

Table 6: Counterfactual evaluation of hypothesis-guided question rewrites on GSM8K and BBH-structured across open-weight language models. Values are accuracies (%); Δ is rewritten-item accuracy minus original-item accuracy in percentage points. For increase-difficulty rewrites, a negative Δ is the intended outcome; for decrease-difficulty rewrites, a positive Δ is intended. GSM8K contains 141 increase and 51 decrease rewrites; BBH-structured contains 67 increase and 55 decrease rewrites.

		Increase-difficulty rewrites			Decrease-difficulty rewrites		
#	Model	Orig.	Rewrite	Δ	Orig.	Rewrite	Δ
	GSM8K						
1	TinyLlama-1.1B (Zhang et al., 2024)	1.4	1.4	+0.0	0.0	0.0	+0.0
2	Gemma-2B-IT (Team et al., 2024)	4.3	6.4	+2.1	2.0	25.5	+23.5
3	Qwen1.5-0.5B-Chat (Bai et al., 2023)	10.6	2.8	−7.8	0.0	13.7	+13.7
4	Phi-1.5 (Li et al., 2023)	13.5	19.9	+6.4	3.9	43.1	+39.2
5	Qwen1.5-1.8B-Chat (Bai et al., 2023)	22.0	2.1	−19.9	7.8	9.8	+2.0
6	Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)	22.7	17.0	−5.7	11.8	60.8	+49.0
7	Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	39.7	19.9	−19.9	7.8	60.8	+52.9
8	Qwen1.5-MoE-A2.7B-Chat (Bai et al., 2023)	41.8	31.2	−10.6	29.4	66.7	+37.3
9	Gemma-1.1-7B-IT (Team et al., 2024)	47.5	38.3	−9.2	19.6	74.5	+54.9
10	Qwen2-0.5B (Yang et al., 2024)	51.1	19.1	−31.9	25.5	64.7	+39.2
11	Yi-9B (Young et al., 2024)	61.7	34.8	−27.0	31.4	82.4	+51.0
12	Qwen1.5-4B (Bai et al., 2023)	61.7	34.0	−27.7	39.2	60.8	+21.6
13	Llama-3-8B-Instruct (Grattafiori et al., 2024)	73.0	57.4	−15.6	43.1	94.1	+51.0
14	Qwen2-1.5B (Yang et al., 2024)	73.8	41.1	−32.6	43.1	74.5	+31.4
15	Yi-1.5-9B (Young et al., 2024)	77.3	51.1	−26.2	60.8	80.4	+19.6
16	Yi-1.5-9B-Chat (Young et al., 2024)	80.9	58.9	−22.0	56.9	82.4	+25.5
17	Phi-3-mini-4k-instruct (Abdin et al., 2024)	82.3	70.2	−12.1	52.9	88.2	+35.3
	Mean across models	—	—	**−15.3**	—	—	**+32.2**
	BBH-structured						
1	Qwen2.5-0.5B-Instruct (Qwen et al., 2025)	13.4	17.9	+4.5	29.1	38.2	+9.1
2	TinyLlama-1.1B (Zhang et al., 2024)	14.9	9.0	−6.0	10.9	27.3	+16.4
3	Gemma-2B-IT (Team et al., 2024)	38.8	17.9	−20.9	9.1	25.5	+16.4
4	Qwen1.5-4B (Bai et al., 2023)	43.3	16.4	−26.9	12.7	45.5	+32.7
5	Gemma-2-2B-IT (Team et al., 2024)	44.8	17.9	−26.9	23.6	50.9	+27.3
6	Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	56.7	37.3	−19.4	32.7	69.1	+36.4
7	Qwen1.5-MoE-A2.7B-Chat (Bai et al., 2023)	59.7	25.4	−34.3	14.5	38.2	+23.6
8	Qwen2.5-3B-Instruct (Qwen et al., 2025)	61.2	26.9	−34.3	27.3	65.5	+38.2
9	Llama-3-8B-Instruct (Grattafiori et al., 2024)	65.7	38.8	−26.9	38.2	80.0	+41.8
10	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	68.7	47.8	−20.9	43.6	72.7	+29.1
11	Phi-3.5-mini-instruct (Abdin et al., 2024)	77.6	59.7	−17.9	45.5	80.0	+34.5
12	Phi-3-medium-4k-instruct (Abdin et al., 2024)	80.6	68.7	−11.9	63.6	94.5	+30.9
13	Qwen2.5-7B-Instruct (Qwen et al., 2025)	85.1	53.7	−31.3	45.5	87.3	+41.8
14	Gemma-2-9B-IT (Team et al., 2024)	85.1	53.7	−31.3	60.0	92.7	+32.7
15	Qwen2.5-14B-Instruct (Qwen et al., 2025)	86.6	59.7	−26.9	74.5	90.9	+16.4
16	Yi-1.5-9B-Chat (Young et al., 2024)	88.1	37.3	−50.7	58.2	87.3	+29.1
	Mean across models	—	—	**−23.9**	—	—	**+28.5**

Table 7: Hypothesis-level rewriting outcomes. Q counts accepted rewrites; transition counts pool 17 models on GSM8K and 16 on BBH-structured. Let n_{CC}, n_{CW}, n_{WW}, and n_{WC} denote correct-to-correct, correct-to-incorrect, incorrect-to-incorrect, and incorrect-to-correct counts, respectively. Success rate is $(n_{CW} + n_{WW})/N$ for Increase and $(n_{WC} + n_{CC})/N$ for Decrease, where $N = n_{CC} + n_{CW} + n_{WW} + n_{WC}$ includes unchanged outcomes. A dash denotes no accepted rewrites. Original outcomes come from historical RouterEval evaluations.

	Increase				Decrease			
#	**Q**	**✓→✗**	**✗→✓**	**Success rate (%)**	**Q**	**✓→✗**	**✗→✓**	**Success rate (%)**
				GSM8K				
1	11	46	9	75.9	0	0	0	—
2	11	48	5	83.4	0	0	0	—
3	9	37	11	68.0	2	0	10	29.4
4	28	97	40	68.3	5	2	25	36.5
5	3	22	1	90.2	3	1	18	56.9
6	32	102	55	61.2	8	4	60	66.9
7	5	39	0	80.0	25	22	156	63.8
8	1	5	1	76.5	3	3	13	39.2
9	10	37	9	84.7	0	0	0	—
10	31	103	39	67.0	5	3	32	57.6
Overall	**141**	**536**	**170**	**70.3**	**51**	**35**	**314**	**57.8**
				BBH-structured				
1	8	43	10	60.2	8	10	42	50.8
2	11	36	13	55.1	2	0	15	56.2
3	5	29	5	62.5	7	5	31	65.2
4	9	38	8	68.1	12	9	59	67.2
5	15	87	23	66.7	8	12	41	61.7
6	3	16	7	52.1	2	4	15	62.5
7	14	71	15	64.3	3	0	15	64.6
8	2	18	1	84.4	13	4	77	76.9
Overall	**67**	**338**	**82**	**63.2**	**55**	**44**	**295**	**65.3**

E ALL GENERATED HYPOTHESES

Table 8: All 20 final HypoGenic hypotheses on GSM8K, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
	GSM8K			
1	*Questions involving counter-intuitive operations or implicit variable relationships, such as returned items or equating frequency of work to earnings, are more likely to be rated as difficulty level 3.*	$2.036_{[1.079,\ 2.975]}$	1.154	6.0
2	*Questions that require tracking and aggregating multiple entities or sub-totals over several steps tend to be rated as moderate to high difficulty.*	$1.442_{[0.957,\ 1.905]}$	0.855	40.5
3	*The inclusion of negative numbers, unit conversions, or complex ratios significantly increases the difficulty level of a question.*	$1.052_{[0.263,\ 1.883]}$	0.584	10.5
4	*Questions that require tracking and combining multiple sub-totals before reaching the final answer are more difficult than those with a single linear calculation path.*	$1.291_{[0.798,\ 1.774]}$	0.753	41.0
5	*Problems that involve implicit conversions or understanding of units, such as 'dozen' or 'a week', often have a higher difficulty level.*	$0.718_{[-0.130,\ 1.607]}$	0.396	11.5
6	*Questions requiring more than three distinct mathematical operations tend to have a higher difficulty level.*	$1.520_{[1.058,\ 1.996]}$	0.912	42.5
7	*Questions with a larger number of distinct steps or operations required to reach the final answer generally have a higher difficulty level.*	$1.762_{[1.282,\ 2.259]}$	1.080	32.5
8	*Questions that require tracking multiple overlapping categories or states, such as different genders and species of fish, tend to have the highest difficulty level.*	$1.502_{[0.174,\ 2.718]}$	0.828	2.5
9	*Questions that require simple, sequential addition and subtraction operations without complex conversions or intermediate reasoning steps tend to have a lower difficulty level.*	$-1.465_{[-1.939,\ -1.008]}$	-0.851	23.0
10	*Questions with a larger number of distinct arithmetic operations or intermediate steps generally have a higher difficulty level.*	$1.584_{[1.125,\ 2.053]}$	0.960	44.5
11	*Questions involving four or more distinct mathematical operations or intermediate steps are more likely to be rated as difficulty level 3.*	$1.572_{[1.118,\ 2.046]}$	0.951	44.0
12	*Questions with more than three distinct calculation steps tend to be more difficult.*	$1.479_{[0.995,\ 1.934]}$	0.884	46.5
13	*Questions with a larger number of sequential arithmetic steps (e.g., calculating values for five consecutive days) tend to be more difficult.*	$2.000_{[1.354,\ 2.704]}$	1.182	14.0
14	*Questions with multiple sequential steps involving different operations (e.g., addition followed by multiplication and then subtraction) are more difficult.*	$1.203_{[0.731,\ 1.677]}$	0.696	54.5
15	*Questions involving percentages or interest rates are more likely to be rated as difficulty level 3.*	$0.340_{[-0.339,\ 1.033]}$	0.186	15.0
16	*Questions that require setting up algebraic equations to solve for unknown variables, particularly involving time shifts like future ages, are associated with higher difficulty levels.*	$1.175_{[0.522,\ 1.833]}$	0.649	5.5
17	*The presence of conditional logic or threshold-based rules, such as applying an operation only if a certain value is exceeded, increases the difficulty level of the question.*	$0.366_{[-1.520,\ 1.669]}$	0.200	1.5
18	*Questions that require multiple unit conversions or rate calculations across different time scales tend to have a higher difficulty level.*	$1.060_{[-0.113,\ 2.348]}$	0.586	6.5
19	*Questions that require more than three distinct mathematical operations to reach the final answer are generally harder.*	$1.328_{[0.853,\ 1.816]}$	0.779	43.5
20	*Problems that can be solved with a direct, single-layer sum of products are generally the easiest and receive a difficulty level of 1.*	$-1.178_{[-1.753,\ -0.611]}$	-0.659	12.0

Table 9: All 20 final HypoGenic hypotheses on BBH-structured, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Statistics exclude missing labels (275 of 280 test items have usable labels). Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
	BBH-structured			
1	*Questions where the final answer is explicitly stated verbatim in the prompt text are easier than those requiring implicit logical deduction.*	$-0.949_{[-1.219,\ -0.672]}$	-0.865	30.2
2	*Questions that require tracking and ordering five or more entities are generally harder than those involving fewer entities, unless the answer is explicitly stated.*	$0.824_{[0.555,\ 1.096]}$	0.743	44.0
3	*Logical deduction tasks that require combining multiple relative constraints to deduce a complete sequence are typically rated at difficulty level 3.*	$0.946_{[0.634,\ 1.253]}$	0.859	28.4
4	*Questions that require sorting numerical values or combining multiple relational constraints to find an intermediate position tend to be rated at the highest difficulty level.*	$1.322_{[1.029,\ 1.616]}$	1.261	21.5
5	*Questions that require deducing the relative order of objects from multiple constraints are generally more difficult than questions that explicitly state the answer or involve simple counting.*	$0.170_{[-0.112,\ 0.450]}$	0.144	48.4
6	*Questions involving a larger number of objects or constraints (e.g., five or seven objects) tend to have a higher difficulty level than those with fewer objects (e.g., three objects), unless the answer is explicitly stated.*	$0.692_{[0.427,\ 0.960]}$	0.613	47.6
7	*Logical deduction tasks involving a larger number of objects or more complex relative positioning constraints are associated with higher difficulty levels.*	$1.060_{[0.750,\ 1.365]}$	0.978	26.9
8	*Logical deduction tasks that require chaining multiple relative position statements (e.g., 'A is left of B', 'B is left of C') to find an unanchored object's position are more difficult (level 3) than those with fewer steps.*	$0.937_{[0.614,\ 1.260]}$	0.849	27.6
9	*Questions with a larger number of objects to order (e.g., 5 or 7) are generally more difficult than those with fewer objects (e.g., 3), unless the answer is explicitly given.*	$0.677_{[0.400,\ 0.951]}$	0.598	45.8
10	*Questions that require combining multiple constraints to deduce a full sequence or count items are more difficult than those that ask for a single explicit fact.*	$0.452_{[0.169,\ 0.722]}$	0.390	53.8
11	*Questions involving a larger number of objects or events to track tend to have a higher difficulty level.*	$1.365_{[1.139,\ 1.587]}$	1.413	56.0
12	*Questions where the answer is explicitly stated in the text without requiring any deductive reasoning are typically assigned a lower difficulty level.*	$-1.295_{[-1.546,\ -1.040]}$	-1.225	20.7
13	*Questions where the answer is explicitly stated in the prompt (e.g., 'The cardinal is the fourth from the left') are easier, but might still be rated as difficulty 2 if they are embedded in a complex paragraph.*	$-1.121_{[-1.388,\ -0.853]}$	-1.015	16.7
14	*Questions where the answer is explicitly stated in the prompt without requiring any logical deduction or inference tend to have lower difficulty levels (1-2).*	$-1.295_{[-1.552,\ -1.045]}$	-1.225	20.7
15	*Questions requiring the model to chain multiple logical constraints together to deduce an unstated sequence are harder than those where the answer is explicitly stated in the text.*	$0.110_{[-0.167,\ 0.384]}$	0.093	53.8
16	*Questions where the correct answer is explicitly stated verbatim in the premise are the easiest and have a difficulty level of 1.*	$-1.273_{[-1.526,\ -1.025]}$	-1.193	19.6
17	*Questions that require identifying an object based on a relative position (e.g., 'second from the right') and then deducing the remaining positions are harder than those with direct positional statements.*	$1.037_{[0.725,\ 1.351]}$	0.943	21.8
18	*Questions that require combining multiple constraints to find an unstated missing interval are harder than those that ask for a direct lookup.*	$0.201_{[-0.032,\ 0.433]}$	0.170	22.9
19	*Tasks that require deducing the position of an object through a chain of relative spatial or temporal constraints are more likely to be rated at the highest difficulty level (level 3).*	$0.872_{[0.560,\ 1.184]}$	0.784	29.5
20	*Questions with a large number of distractors or irrelevant information to filter through before counting tend to be more difficult.*	$0.562_{[-0.023,\ 1.199]}$	0.479	6.2

Table 10: All 20 final HypoGenic hypotheses on Winogrande, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	SEP$_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	Winogrande			
1	*Questions requiring the model to infer physical constraints or spatial relationships, such as a bulky object breaking a band or a minuscule ring not fitting a gem, are typically harder.*	$0.258_{[-0.400,\ 0.933]}$	0.153	10.6
2	*Increased syntactic distance or the use of complex clause structures between the target blank and the contextual clues increases the difficulty level.*	$0.263_{[—]}$	—	0.4
3	*Questions containing explicit causal conjunctions like 'because' or 'so' that directly link an action to its immediate outcome are more likely to be rated as difficulty level 1.*	$0.058_{[-0.385,\ 0.486]}$	0.034	65.4
4	*The presence of negation, sarcasm, or counter-intuitive scenarios increases the difficulty level of the question.*	$0.070_{[-0.748,\ 0.947]}$	0.041	4.7
5	*Questions that rely on direct, common-sense word associations, such as 'chainsaw' and 'tree cutter', are easier (level 1) than those requiring multi-step logical deductions.*	$-0.919_{[-1.573,\ -0.138]}$	-0.550	7.1
6	*Questions where the pronoun resolution depends on understanding basic emotional reactions to direct actions, such as lying to avoid someone's anger, are generally easier.*	$0.287_{[-0.614,\ 1.277]}$	0.170	4.7
7	*Questions where the causal link is explicitly stated using simple, direct adjectives (e.g., 'too large', 'tasteless', 'too awkward') are generally the easiest (level 1).*	$0.205_{[-0.499,\ 0.935]}$	0.122	11.0
8	*Questions that explicitly contrast two entities with well-known attributes, like clear versus dark liquors or heavy versus light objects, are easier to resolve (level 2).*	$-0.475_{[-0.942,\ 0.032]}$	-0.283	15.7
9	*Questions where the blank can be resolved by matching a simple descriptive trait to a stated outcome are typically of moderate difficulty.*	$-0.246_{[-0.704,\ 0.224]}$	-0.146	18.9
10	*Questions involving human names and straightforward cause-and-effect relationships regarding personal actions tend to be of moderate difficulty (level 2).*	$-0.236_{[-0.684,\ 0.208]}$	-0.140	64.2
11	*Questions that rely on specific external commonsense knowledge about objects, such as stencils or seeing-eye dogs, usually have moderate difficulty (level 2).*	$-0.152_{[-0.720,\ 0.423]}$	-0.090	15.0
12	*Questions where the correct referent can be resolved through simple syntactic parallelism or explicit contrast in the sentence structure tend to be easier (level 1).*	$-0.073_{[-0.574,\ 0.451]}$	-0.043	13.4
13	*Questions involving direct physical properties or simple cause-and-effect relationships, such as ice being hard or stencils making neat letters, are generally easier (level 2).*	$-0.109_{[-0.588,\ 0.390]}$	-0.065	24.8
14	*Questions involving the spatial tracking of objects transferred between multiple actors tend to be harder.*	$0.263_{[—]}$	—	0.4
15	*Questions that rely on understanding implicit social norms or specific background knowledge, like the purpose of an etiquette school, tend to be harder (level 3).*	$-0.145_{[-0.887,\ 0.618]}$	-0.086	4.3
16	*Questions that rely on direct, explicit causal links stated in the text, such as a garment falling off because it is loose, are the easiest (level 1).*	$0.006_{[-0.460,\ 0.483]}$	0.004	27.2
17	*Questions that require reasoning about physical constraints or highly context-dependent attributes of inanimate objects tend to be the hardest (level 3).*	$0.308_{[-0.392,\ 1.029]}$	0.183	7.5
18	*Questions where the correct answer relies on understanding the negation of a typical expectation, such as a curable illness eliciting less sympathy, tend to be harder.*	$-0.670_{[—]}$	—	0.4
19	*Questions that involve straightforward, direct cause-and-effect relationships without complex contrasting conditions are generally easier and have a difficulty level of 1.*	$-0.243_{[-0.677,\ 0.197]}$	-0.144	37.0
20	*Questions that require multi-step logical deduction involving contrast words like 'although' or 'while' are more likely to have a difficulty level of 3.*	$-0.284_{[-1.069,\ 0.481]}$	-0.169	5.5

Table 11: The 20 explained HypotheSAEs features on GSM8K. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Matching means SAE activation > 0, not LLM-judged agreement with the explanation. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	GSM8K			
1	*no coherent item-level property* Trigger: the final digit of the last number in an arithmetic expression, immediately preceding the equals sign.	—	—	100.0
2	*no coherent item-level property* Trigger: The space token immediately following the '####' marker before the final answer in GSM8K solutions.	$-0.579_{[-1.576,\ 0.434]}$	-0.318	94.0
3	*The question involves specific times of day formatted with a colon.* Trigger: The colon ':' used in time formats (e.g., 5:00, 10:00).	$1.523_{[0.546,\ 2.600]}$	0.855	8.0
4	*no coherent item-level property* Trigger: A space character immediately preceding a number.	$1.072_{[0.561,\ 1.576]}$	0.612	39.0
5	*no coherent item-level property* Trigger: The equals sign (' =') immediately preceding a calculator annotation (' ¡¡') in the reference solution.	$0.136_{[-0.356,\ 0.624]}$	0.074	53.0
6	*no coherent item-level property* Trigger: the first digit of a multi-digit number, particularly in the context of mathematical equations or quantities	$0.533_{[0.037,\ 1.033]}$	0.295	56.0
7	*no coherent item-level property* Trigger: the comma token ',' in lists or clauses	$0.410_{[-0.173,\ 0.992]}$	0.226	69.5
8	*no coherent item-level property* Trigger: digits inside the result part of a calculator annotation (¡¡...=...¿¿)	$1.074_{[0.411,\ 1.748]}$	0.603	19.0
9	*The reference solution contains a mathematical expression followed by '=$¡¡' or '=¡¡' where the expression ends in a multi-digit number or a number with a decimal.* Trigger: The last digit of a number immediately preceding the '=$¡¡' or '=¡¡' calculator annotation in GSM8K solutions.	$0.690_{[-0.340,\ 1.725]}$	0.379	93.0
10	*The reference solution contains mathematical expressions with parentheses or brackets.* Trigger: the closing bracket or parenthesis in a mathematical expression, or the final character before it	$0.970_{[0.489,\ 1.467]}$	0.550	52.0
11	*The reference solution contains an intermediate or final calculation resulting in a number starting with 18 or 19.* Trigger: The '=' token inside the ¡¡ ¿¿ calculator annotation when the result starts with '18' or '19'.	$0.822_{[0.318,\ 1.326]}$	0.460	63.5
12	*The question or reference solution involves numbers with zeros, such as decimals (e.g., 0.03, 1.04) or hundreds (e.g., 105, 106, 2000).* Trigger: The digit '0' or a decimal point '.' when it is part of a number containing a zero, especially in the context of decimals or hundreds.	$1.258_{[—]}$	—	99.5
13	*no coherent item-level property* Trigger: Tokens that are part of the narrative or descriptive text in math word problems, often nouns or prepositions following numbers or quantities.	$0.136_{[-0.768,\ 0.972]}$	0.074	88.5
14	*no coherent item-level property* Trigger: Sentence-ending punctuation (periods) and subsequent whitespace (spaces or newlines) in the problem description or solution text.	$0.653_{[0.123,\ 1.187]}$	0.362	36.0
15	*The question involves financial loss, negative values, or calculating a difference/change in money.* Trigger: Tokens related to financial loss, negative values, or the start of the answer section in problems involving money or negative numbers.	$0.695_{[-0.163,\ 1.501]}$	0.382	89.5
16	*The question involves a character planning, intending, or being able to do something in the future.* Trigger: The tokens ' to', ' will', and ' can' when used as auxiliary verbs or infinitives indicating future action or intent.	$0.973_{[0.394,\ 1.559]}$	0.547	25.0
17	*The question involves calculating time, dates, or durations.* Trigger: Prepositions and temporal markers indicating time or sequence, such as 'at', 'before', 'on', 'from', and 'late'.	$1.331_{[0.498,\ 2.202]}$	0.752	13.5
18	*The reference solution involves setting up and solving algebraic equations or performing arithmetic calculations with explicit operators.* Trigger: mathematical operators like +, -, and * in algebraic expressions or arithmetic calculations	—	—	100.0
19	*no coherent item-level property* Trigger: words introducing a specific subset, condition, or variable assignment in a math word problem, such as 'with', 'total', 'of', 'the', 'weigh', 'has', 'red', 'charges', 'equal'	$0.710_{[-0.277,\ 1.627]}$	0.390	93.5
20	*no coherent item-level property* Trigger: The comma token ',' used as a separator in lists or as a thousands separator in numbers.	$1.711_{[1.393,\ 2.051]}$	0.942	98.5

Table 12: The 20 explained HypotheSAEs features on BBH-structured. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Matching means SAE activation > 0, not LLM-judged agreement with the explanation. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	BBH-structured			
1	*The reference solution involves calculating a sum, average, or identifying a specific numerical quantity.* Trigger: The space or first digit of a number that represents the result of an arithmetic calculation or a specific numerical value in the reasoning process.	$0.522_{[0.277,\ 0.769]}$	0.453	31.1
2	*The reference solution contains a list of relative ordering constraints or properties separated by commas, culminating in an 'and'.* Trigger: The comma separating the penultimate and final items in a list of sequential logical deductions or conditions, typically right before 'and'.	—	—	100.0
3	*The question asks to sort items by alphabetic order or find the second heaviest item.* Trigger: The token 'abetic' in the phrase 'alphabetic order' or the token ' second' in the phrase 'The second heaviest'.	$0.869_{[0.109,\ 1.707]}$	0.745	3.6
4	*The question has at least six multiple-choice options (reaching option F or beyond).* Trigger: The open parenthesis '(' for option F, or the name 'Gwen' in option D.	$0.520_{[0.234,\ 0.816]}$	0.450	27.9
5	*The question involves logical deduction over ordered objects, specifically identifying the position of an object in a sequence.* Trigger: the word 'the' in phrases like 'from the left' or 'from the right' within multiple-choice options	$0.326_{[-0.185,\ 0.828]}$	0.278	93.2
6	*The reference solution directly quotes or refers to a specific statement in the prompt that explicitly gives the answer.* Trigger: The word 'explicitly' in the phrase 'explicitly states' or 'explicitly establishes' within the explanation of a logical deduction problem.	$0.346_{[-0.521,\ 1.168]}$	0.294	2.5
7	*The question involves identifying or filtering objects based on their attributes (e.g., name, color, age) from a given set or table.* Trigger: The word 'named' or 'is' when identifying a specific object or entity from a table or description.	—	—	100.0
8	*no coherent item-level property* Trigger: The tokens ' alph' (in 'alphabetic') and ' electronic' (in 'electronic store').	$0.282_{[-0.185,\ 0.893]}$	0.240	5.4
9	*The reference solution explicitly states the answer using a phrase like 'is the [superlative]' or 'is [name]'* Trigger: the word 'the' or 'is' immediately preceding the superlative or name that answers the question in the reference solution	$0.352_{[-0.526,\ 1.199]}$	0.300	5.4
10	*The question asks 'How many [adjective] things do you see' or the reference solution evaluates conditions with 'does not meet'.* Trigger: The token ' you' in the phrase 'How many [adjective] things do you see' or the token ' not' in 'does not meet'.	$0.689_{[0.392,\ 0.981]}$	0.608	68.9
11	*no coherent item-level property* Trigger: Tokens like '-', 'have', 'the', and ')' that appear in the middle of a sentence or phrase in the reference solution or question.	$-0.296_{[-0.760,\ 0.171]}$	-0.252	7.5
12	*The reference solution involves identifying the leftmost object in a sequence.* Trigger: the word 'the' immediately preceding 'leftmost' or 'first' in the context of identifying the leftmost object in a sequence	$-1.482_{[-1.995,\ -0.698]}$	-1.281	2.1
13	*no coherent item-level property* Trigger: punctuation or short words immediately following the first sentence or phrase of the 'Answer:' section in multiple-choice reasoning tasks	$-0.720_{[-1.475,\ 0.071]}$	-0.613	1.1
14	*The reference solution uses the phrase 'establishes that' to explain the logical deductions.* Trigger: The word 'that' following the word 'establishes' in the explanation of a logical deduction problem.	$-0.637_{[-1.018,\ -0.250]}$	-0.552	84.3
15	*The question asks 'how many' objects remain or meet a condition, resulting in a long list of numerical options (A through Q).* Trigger: The newline character '\n' immediately preceding a multiple-choice option letter in parentheses, specifically in long lists of options (e.g., A through Q).	$2.290_{[1.768,\ 2.815]}$	1.972	99.3
16	*no coherent item-level property* Trigger: The hyphen in hyphenated words like 'non-black' or 'fourth-newest', or the 'ie' in 'scrunchiephone'.	$0.715_{[0.444,\ 0.986]}$	0.635	66.4
17	*The reference solution uses relative clauses to filter or describe items from a table or list.* Trigger: relative pronouns like 'which' or 'that', or past participles like 'listed', introducing a relative clause in the explanation of a reasoning problem	$0.552_{[-0.731,\ 1.624]}$	0.471	2.5

Continued on next page

HypotheSAEs – BBH-structured (continued)

#	Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
18	*The reference solution involves determining a sorted order or sequence of items.* Trigger: The period at the end of a sentence that establishes a sorted order or sequence of items.	—	—	100.0
19	*The reference solution involves multi-step logical deduction to determine the relative order or position of objects.* Trigger: The word 'must' or 'to' in the context of logical deduction steps in the reference solution.	$1.636_{[1.335,\ 1.936]}$	1.400	99.3
20	*no coherent item-level property* Trigger: Tokens like 'the', 'green', 'as', 'for', 'fidget', 'mauve', 'at', 'saw' in the context of listing objects, schedules, or attributes in the reference solution or question.	$-0.028_{[-0.421,\ 0.375]}$	-0.024	3.6

Table 13: The 20 explained HypotheSAEs features on Winogrande. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Matching means SAE activation > 0, not LLM-judged agreement with the explanation. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	Winogrande			
1	*The reference solution explains the reasoning by attributing a state of being or property to one of the characters (e.g., being blind, Jewish, hungry, skilled, poor).* Trigger: The copula verbs 'is', 'are', 'was', 'be' or the noun 'kids' when describing a property or state of a person in the explanation of a Winograd schema.	$0.297_{[-0.528,\ 1.064]}$	0.176	93.3
2	*The question involves a scenario set in a workplace, business, or commercial context.* Trigger: Tokens related to businesses, workplaces, or commercial entities (like 'factory', 'plant', 'supermarket', 'company', 'companies', 'merchant', 'office') and their preceding determiners/modifiers.	$-0.208_{[-1.070,\ 0.651]}$	-0.123	2.4
3	*no coherent item-level property* Trigger: The feature fires on nouns or names that are the object of a preposition or part of a noun phrase in the explanation section of a Winograd schema problem.	$-0.383_{[-1.182,\ 0.563]}$	-0.227	2.4
4	*The reference solution uses a complex sentence structure with relative clauses, modal verbs, or infinitive phrases to explain the reasoning.* Trigger: Tokens like 'would', 'which', 'to', 'that', 'can', or commas that introduce a relative clause, modal verb, or infinitive phrase explaining a causal or conditional relationship in the reasoning.	$0.263_{[-0.774,\ 1.286]}$	0.156	4.3
5	*no coherent item-level property* Trigger: Tokens related to quantity, degree, or specific actions (e.g., 'any', 'his', 'a', 'many', 'less', 'cure', 'pose', 'of') in the context of reasoning explanations.	$-0.020_{[-1.275,\ 1.636]}$	-0.012	2.0
6	*no coherent item-level property* Trigger: The closing quote mark (") immediately following the correct answer word in the explanation.	$-0.235_{[-0.640,\ 0.178]}$	-0.139	41.3
7	*no coherent item-level property* Trigger: Various function words and verbs in the middle of the sentence, such as 'had', 'ordered', 'into', 'it', 'arrested', 'and', 'for', 'of', 'ran', 'but', 'even'.	$-0.961_{[-2.068,\ 0.118]}$	-0.571	1.6
8	*The reference solution contains a conjunction ('and' or 'or') linking two items or concepts.* Trigger: The word 'and' or 'or' used as a conjunction between two nouns or noun phrases.	$-0.502_{[-0.983,\ 0.007]}$	-0.299	17.7
9	*no coherent item-level property* Trigger: Tokens related to specific objects or actions in the sentence, such as 'rolls', 'roller', 'toilet', 'the', 'a', 's', 'but', 'cinnamon'.	$-0.997_{[-1.603,\ -0.366]}$	-0.595	4.7
10	*The question involves reasoning about quantities, amounts, or volumes of substances or items.* Trigger: words related to quantity, volume, or amount, such as 'amount', 'volume', 'portions', and 'much'	$0.304_{[-0.655,\ 1.225]}$	0.180	3.1
11	*The question involves a scenario where someone wants, agrees, or intends to do something.* Trigger: Tokens related to wanting, agreeing, or intending to do something (e.g., 'to', 'for', 'their', 'willing', 'favored', 'wanted').	$0.109_{[-1.304,\ 1.792]}$	0.065	2.8
12	*The question has the blank at the very beginning of the second sentence or clause, immediately following a period.* Trigger: The underscore token ('_' or ' _') representing the blank in a Winograd-schema sentence.	$0.123_{[-0.300,\ 0.548]}$	0.073	27.6
13	*no coherent item-level property* Trigger: Tokens that are parts of words related to specific concepts like 'jew' (Jewish/jeweled), 'fever', 'pun', 'finishing', or 'ugl' (uglier).	$-0.588_{[—]}$	—	0.4
14	*The reference solution explicitly refers to the 'first' or 'second' clause, part, or sentence.* Trigger: The words 'first' and 'second', typically referring to parts of a sentence or sequence.	$-0.528_{[-0.944,\ -0.099]}$	-0.316	34.6
15	*no coherent item-level property* Trigger: Pronouns or names referring to the subject of the sentence, often in the explanation section of a Winograd schema problem.	$-0.608_{[-1.304,\ 0.167]}$	-0.362	5.5
16	*no coherent item-level property* Trigger: Tokens related to cooking, food, or specific ingredients, as well as contrastive conjunctions like 'whereas'.	$1.282_{[-1.143,\ 3.701]}$	0.761	0.8
17	*The question involves a scenario related to work, employment, commuting, or performing a specific task/job.* Trigger: the word 'to' or other words related to working, commuting, or performing an action, often in the context of employment or tasks	$0.197_{[-0.234,\ 0.643]}$	0.117	37.4

Continued on next page

HypotheSAEs – Winogrande (continued)

#	Hypotheses	$\mathbf{SEP}_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
18	*no coherent item-level property* Trigger: Tokens like 'was', 'is', 'which', and commas in the explanation section of a Winograd schema problem.	$-0.303_{[-0.877,\ 0.304]}$	-0.180	15.4
19	*The question sentence ends with a blank immediately followed by a period.* Trigger: The period token '.' at the end of a sentence with a blank at the end.	$3.053_{[2.749,\ 3.353]}$	1.833	0.8
20	*no coherent item-level property* Trigger: Tokens that are part of the reasoning chain explaining the correct answer, often comparative or descriptive words like 'more', 'less', 'maximizing', 'the', 'is', 'now', 'an', 'his'.	$0.008_{[-0.421,\ 0.452]}$	0.005	36.2

Table 14: All 20 final NLParam hypotheses on GSM8K, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
	GSM8K			
1	*requires calculating an average; specifically, the question asks for the mean or average of a set of values*	$1.514_{[0.413,\ 2.500]}$	0.835	2.5
2	*involves unit conversions; specifically, the problem requires converting between different units of measurement (e.g., minutes to hours, kb to Mb, yards to miles)*	$0.781_{[-0.225,\ 1.926]}$	0.430	8.0
3	*contains a narrative with physical actions causing subsequent events; specifically, the text describes an action (like throwing an object) that directly causes another event (like more objects falling)*	$-1.026_{[—]}$	—	0.5
4	*involves multi-step arithmetic with more than three operations; specifically, the solution requires performing four or more distinct mathematical operations to reach the final answer*	$1.574_{[1.114,\ 2.040]}$	0.954	46.0
5	*uses decimal numbers; specifically, the text contains non-integer numerical values such as 1.5, 1.25, or 2.50 in the problem description*	$0.045_{[-0.630,\ 0.765]}$	0.025	9.0
6	*features water or liquid volume as the primary subject; specifically, the problem involves measuring, transferring, or calculating amounts of water in gallons or liters*	$-0.260_{[-1.000,\ 0.482]}$	-0.142	1.0
7	*has a shorter problem description; specifically, the question text is relatively brief and contains fewer sentences compared to other problems*	$-1.296_{[-1.990,\ -0.623]}$	-0.730	87.0
8	*involves multiple people or entities with comparative relationships; specifically, the problem defines one person's amount or time in terms of another's (e.g., twice as long, $0.50 more)*	$0.026_{[-0.581,\ 0.645]}$	0.014	20.5
9	*requires tracking multiple distinct items with different prices; specifically, the text involves calculating the total cost of several different goods or services, each with its own unit price*	$0.084_{[-0.739,\ 0.930]}$	0.046	8.0
10	*involves multiple entities with relative values; specifically, the problem defines the value of one entity in terms of another (e.g., '800 pounds more than', '4 times what Grace weighs')*	$-0.112_{[-0.653,\ 0.434]}$	-0.061	27.5
11	*involves geometric or spatial concepts; specifically, the problem mentions shapes, angles, or dimensions like length, width, or height*	$0.469_{[-0.283,\ 1.235]}$	0.257	7.5
12	*requires tracking time or schedules; specifically, the text involves calculating durations, start times, or end times across a timeline*	$0.446_{[-0.525,\ 1.468]}$	0.244	9.0
13	*involves multiple distinct entities or items; specifically, the problem requires tracking quantities for three or more different types of objects or people (e.g., MTB, BMX, and Trekking bikes; bicycles, tricycles, and unicycles)*	$-0.110_{[-0.688,\ 0.475]}$	-0.060	20.0
14	*requires calculating rates or unit prices; specifically, the text involves quantities expressed as 'per hour', 'per week', 'apiece', or 'each'*	$0.089_{[-0.417,\ 0.600]}$	0.049	47.0
15	*involves fractions or percentages; specifically, the question requires calculating or reasoning with fractions, ratios, or percentages*	$0.617_{[0.092,\ 1.138]}$	0.342	40.0
16	*requires converting units or rates; specifically, the problem involves translating between different units of measurement, time, or percentages to solve the problem*	$0.781_{[-0.234,\ 1.929]}$	0.430	8.0
17	*requires categorizing a single type of item into multiple sub-types; specifically, the problem involves tracking a generic item (like apples or trees) that is split into specific categories (e.g., large/small, tallest/shortest/average)*	$0.012_{[-0.631,\ 0.711]}$	0.007	12.0
18	*requires tracking items that are both added and removed; specifically, the problem involves a scenario where items are gained and then some are lost, given away, or thrown*	$0.589_{[-0.704,\ 1.953]}$	0.323	5.5
19	*requires multi-step reasoning with intermediate variables; specifically, the problem cannot be solved in one or two simple arithmetic steps and requires keeping track of multiple intermediate quantities*	$1.406_{[0.944,\ 1.872]}$	0.831	58.0
20	*involves large numbers; specifically, the problem requires arithmetic operations on numbers in the thousands*	$1.214_{[0.507,\ 1.974]}$	0.684	16.0

Table 15: All 20 final NLParam hypotheses on BBH-structured, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	SEP$_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	BBH-structured			
1	*requires determining the position of an object relative to others; specifically, the question asks for the position of an object (e.g., 'third from the left', 'rightmost') based on a set of relative ordering constraints*	$0.309_{[0.009,\ 0.609]}$	0.264	27.5
2	*involves logical deduction of a sequence; specifically, the text requires ordering a set of objects based on a series of relative constraints*	$-0.048_{[-0.315,\ 0.223]}$	-0.041	54.3
3	*requires multi-step logical deduction to determine a sequence; specifically, the text provides a set of constraints about the relative order or age of objects and asks to identify the position of a specific object*	$0.134_{[-0.142,\ 0.411]}$	0.114	49.3
4	*includes an example sentence explaining how to read the data; specifically, the text contains a phrase like '*	$-0.021_{[-0.627,\ 0.624]}$	-0.018	6.4
5	*has a multiple-choice format with more than five options; specifically, the text presents a question followed by six or more lettered options*	$0.454_{[0.158,\ 0.737]}$	0.393	37.5
6	*requires determining the oldest or newest item in a sequence; specifically, the question asks to identify the item at the extreme end of an ordered list*	$-0.596_{[-0.984,\ -0.177]}$	-0.513	12.5
7	*contains distractor objects; specifically, the text lists items that are irrelevant to the final question being asked, requiring the model to ignore them*	$-0.019_{[-0.377,\ 0.341]}$	-0.016	22.9
8	*requires deducing the order of a sequence of objects; specifically, the task involves arranging items based on relative position clues*	$-0.048_{[-0.318,\ 0.218]}$	-0.041	54.3
9	*contains the phrase 'extremely simple questions'; specifically, the task description explicitly states it is about answering extremely simple questions about objects*	$-0.069_{[-0.456,\ 0.318]}$	-0.058	18.6
10	*involves spatial reasoning with relative positions; specifically, the text requires deducing the exact order of objects based on clues like 'left of', 'rightmost', or 'second from the left'*	$-0.242_{[-0.553,\ 0.060]}$	-0.207	24.3
11	*involves multiple relative constraints; specifically, the text provides several statements comparing the positions or ages of objects (e.g., 'older than', 'to the right of') that must be combined*	$-0.037_{[-0.310,\ 0.242]}$	-0.032	53.9
12	*options are complete sentences; specifically, the multiple-choice options provided are full sentences rather than single words, numbers, or short phrases*	$-0.048_{[-0.313,\ 0.218]}$	-0.041	54.3
13	*requires identifying a gap in a schedule; specifically, the text requires finding an unaccounted time period between known activities to determine when an event could have occurred*	$0.144_{[-0.076,\ 0.373]}$	0.122	20.7
14	*describes a fixed arrangement task; specifically, the text explicitly states that objects are 'arranged in a fixed order' or asks to deduce the position of an item in a sequence*	$-0.048_{[-0.318,\ 0.221]}$	-0.041	54.3
15	*describes a logical deduction task; specifically, the text explicitly introduces the problem as 'A logical deduction task which requires deducing the order of a sequence of objects'*	$-0.048_{[-0.323,\ 0.221]}$	-0.041	54.3
16	*requires reasoning about prices or costs; specifically, the text involves comparing the expense of different items, such as fruits at a stand*	$-0.212_{[-0.718,\ 0.306]}$	-0.180	7.5
17	*describes objects on a surface; specifically, the text begins by listing items located on a 'table', 'floor', or 'nightstand'*	$-0.093_{[-0.494,\ 0.301]}$	-0.079	17.9
18	*requires tracking multiple attributes of objects simultaneously; specifically, the text lists objects with both color and type attributes and asks a question that depends on filtering by one or both attributes*	$0.016_{[-0.373,\ 0.407]}$	0.014	19.3
19	*contains multiple objects to order; specifically, the text describes a set of three, five, or seven objects that must be arranged in a fixed sequence*	$-0.048_{[-0.321,\ 0.216]}$	-0.041	54.3
20	*uses relative positional language; specifically, the text contains relational phrases such as 'to the left of', 'to the right of', 'above', or 'below' to establish relationships between entities*	$-0.101_{[-0.381,\ 0.169]}$	-0.086	40.7

Table 16: All 20 final NLParam hypotheses on Winogrande, in exported order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	SEP$_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	Winogrande			
1	*requires reasoning about inverse relationships; specifically, the text requires understanding that an increase in one quantity (e.g., eating more) implies a decrease or lack in another (e.g., having a smaller lunch or being hungry)*	$0.096_{[-0.864,\ 1.203]}$	0.057	3.9
2	*involves negation in the premise; specifically, the text contains words like 'not', 'didn't', or 'unlike' to establish a negative condition for one of the subjects*	$-0.225_{[-0.666,\ 0.233]}$	-0.134	18.5
3	*uses dimensional or spatial adjectives; specifically, the sentence containing the blank uses words like 'small', 'big', 'broad', or 'confined' to describe the missing entity*	$0.300_{[-0.399,\ 1.014]}$	0.178	10.2
4	*requires inferring the cause of a state or action from a contrasting pair of individuals; specifically, the text contrasts two people and requires deducing which one is responsible for a specific outcome or state based on their described characteristics*	$-0.229_{[-0.635,\ 0.200]}$	-0.136	28.7
5	*involves contrasting actions or states between two individuals; specifically, the text describes two people doing different things or having different characteristics, and the blank must be resolved by matching the correct person to the concluding action or state*	$-0.115_{[-0.525,\ 0.308]}$	-0.068	36.2
6	*involves a scenario where one person is acting upon or communicating with another person*	$0.407_{[-0.052,\ 0.880]}$	0.242	27.2
7	*involves inanimate objects rather than human actors; specifically, the options provided for the blank are non-living things, physical items, or spaces rather than people's names*	$0.024_{[-0.408,\ 0.468]}$	0.014	39.0
8	*requires understanding environmental or hygiene concepts; specifically, the text involves reasoning about cleanliness, sterility, pollution, or recycling to resolve the blank*	$0.368_{[-0.405,\ 1.198]}$	0.218	5.5
9	*involves comparing two inanimate objects or locations; specifically, the text asks to resolve a pronoun or blank referring to one of two non-human entities (e.g., rooms, vehicles, containers)*	$0.062_{[-0.370,\ 0.502]}$	0.037	37.8
10	*involves interpersonal relationships or social dynamics; specifically, the text describes interactions between two named individuals, such as giving gifts, breaking hearts, or doing deeds in someone else's name*	$0.452_{[-0.028,\ 0.939]}$	0.270	28.3
11	*requires resolving a pronoun or blank to a person's name; specifically, the options provided are names of people rather than inanimate objects or concepts*	$-0.097_{[-0.525,\ 0.331]}$	-0.058	59.1
12	*contains a subordinate clause starting with 'because' or 'as' that explains a preference or outcome; specifically, the sentence structure relies on a causal conjunction to link the blank to the correct entity*	$0.065_{[-0.344,\ 0.474]}$	0.039	54.7
13	*contains a contrastive conjunction; specifically, the text uses words like 'although', 'but', or 'rather than' to set up the logical relationship between the clauses*	$0.039_{[-0.380,\ 0.477]}$	0.023	33.9
14	*involves reasoning about interpersonal relationships or actions; specifically, the sentence describes an interaction, comparison, or event involving two distinct individuals*	$-0.056_{[-0.492,\ 0.377]}$	-0.033	60.2
15	*requires resolving a pronoun or blank to a subject based on contrasting preferences or behaviors; specifically, the text contrasts two individuals' actions or preferences and requires identifying which one aligns with a given trait or subsequent action*	$-0.000_{[-0.414,\ 0.432]}$	-0.000	30.7
16	*features a 'but not' construction; specifically, the text contrasts two individuals using the phrase 'but not' to indicate that an action or state applies to one but not the other*	$0.110_{[-0.883,\ 1.145]}$	0.065	3.5
17	*requires reasoning about the physical properties or conditions of inanimate objects; specifically, the text asks the reader to infer which object possesses a certain physical trait (e.g., being used, brand new, having faulty screws, being soft, lasting longer) based on the context provided*	$0.061_{[-0.417,\ 0.550]}$	0.036	28.0
18	*requires resolving pronoun or blank reference based on physical properties; specifically, the text requires identifying the correct noun based on physical attributes like 'big', 'thin', or 'soft'*	$0.296_{[-0.221,\ 0.836]}$	0.176	22.4
19	*requires resolving a pronoun or blank to a person based on social roles or interpersonal dynamics; specifically, the text requires inferring who is the employer, who is offering a gift, or who is asking for help based on the context of the interaction*	$0.392_{[-0.097,\ 0.888]}$	0.234	26.0
20	*involves multiple objects with distinct attributes; specifically, the text describes two inanimate objects and assigns different physical attributes or actions to them, requiring the model to map the correct attribute to the correct object*	$0.189_{[-0.277,\ 0.655]}$	0.112	29.9

Table 17: The 40 hypotheses selected by our main-table model on GSM8K, in selection order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
	GSM8K			
1	*The question contains implicit information or requires the solver to infer a missing value from a combination of given constraints (e.g., working backwards from a total or a difference).*	$1.112_{[0.546,\ 1.703]}$	0.623	15.0
2	*The question involves multi-step bookkeeping where intermediate quantities must be calculated and then used as inputs for subsequent calculations, rather than a simple linear sequence of operations.*	$1.498_{[1.043,\ 1.954]}$	0.899	51.0
3	*The reference answer contains a logical error or contradiction, such as ignoring a stated value and summing the wrong numbers, making it difficult for a correct model to match the expected output.*	$3.696_{[3.037,\ 4.361]}$	2.063	1.0
4	*The question requires more than three distinct arithmetic operations to reach the final answer.*	$0.735_{[0.185,\ 1.311]}$	0.408	26.0
5	*Requires at least five distinct arithmetic operations to reach the final answer.*	$1.147_{[0.567,\ 1.732]}$	0.648	20.0
6	*The question requires setting up and solving an algebraic equation with an unknown variable, rather than just performing a sequence of arithmetic operations.*	$1.428_{[0.482,\ 2.338]}$	0.790	4.0
7	*The question requires tracking and comparing multiple entities across different timeframes or conditions (e.g., Jon vs. James, James vs. cousin, super ticket vs. regular ticket).*	$0.101_{[-1.119,\ 1.284]}$	0.055	5.0
8	*The question contains complex sentence structures with multiple conditions, nested clauses, or information presented out of chronological order.*	$0.507_{[—]}$	—	0.5
9	*The question involves rates or unit conversions that are interrupted or require combining multiple different units of measure (e.g., gigabytes per minute with a time interruption, or feet per second combined with spacing in feet).*	$2.216_{[1.559,\ 2.877]}$	1.220	1.0
10	*The question involves multi-step unit conversions or rate calculations (e.g., converting between hours and minutes, or calculating rates based on other rates).*	$1.284_{[0.451,\ 2.212]}$	0.713	6.5
11	*The question requires calculating a rate, ratio, or percentage from given quantities, rather than just applying a given rate or percentage.*	$2.035_{[1.331,\ 2.776]}$	1.174	9.0
12	*Requires calculating a rate or unit cost from a given quantity and time/amount before applying it to a new quantity.*	$2.078_{[1.095,\ 3.109]}$	1.176	5.5
13	*Requires spatial reasoning or understanding the physical layout of an object (e.g., pages on a folded sheet of paper).*	$1.029_{[0.045,\ 2.003]}$	0.563	1.0
14	*The question contains a distractor number or a number that is not directly used in the primary calculation but defines a constraint (e.g., 'six groups of equal size' where the size itself isn't needed, just the number of groups).*	$1.433_{[0.493,\ 2.385]}$	0.803	8.5
15	*The question requires tracking and updating multiple distinct categories or sub-categories of items across a sequence of events.*	$2.096_{[0.422,\ 4.568]}$	1.157	1.5
16	*The question involves calculating a rate of change or unit rate based on a comparative relationship (e.g., '4 weeks to lose the same amount... in a single week').*	$1.909_{[0.813,\ 3.090]}$	1.075	5.5
17	*Requires calculating a difference between an estimated/expected value and a true/actual value.*	$4.605_{[—]}$	—	0.5
18	*The question involves unit conversions or geometric calculations (e.g., volume, area) before the main arithmetic can be completed.*	$-0.011_{[-0.716,\ 0.743]}$	-0.006	3.5
19	*Requires calculating total cost and total revenue to find profit.*	$0.966_{[0.431,\ 1.581]}$	0.529	2.0
20	*Requires calculating percentages of a total amount.*	$0.166_{[-0.561,\ 0.930]}$	0.091	11.5
21	*Requires converting between minutes and hours to aggregate total time across different activities.*	—	—	0.0
22	*The question requires calculating a final remaining amount or difference from a stated initial capacity, total, or budget.*	$-0.036_{[-0.747,\ 0.744]}$	-0.019	14.5
23	*The question contains distractor information or complex conditional logic that must be parsed to determine which numbers are actually relevant to the final calculation (e.g., a lawyer donating three times the sum of specific other donations).*	$0.549_{[-0.418,\ 1.515]}$	0.301	3.0
24	*The question requires working backwards from a final state or a target average to find an initial value or a missing component (e.g., finding the initial number of fries, the initial points, or the required score on a final test).*	$1.155_{[0.411,\ 1.957]}$	0.637	4.5
25	*The question involves calculating a final value based on a combination of fixed costs/quantities and variable costs/quantities (e.g., studio rental fee plus per-student fee, base salary plus per-mile rate and bonus).*	$1.388_{[0.285,\ 2.723]}$	0.763	2.0
26	*The question requires calculating a net profit or rate by subtracting an ongoing cost from an ongoing revenue before dividing a fixed cost by the net rate.*	—	—	0.0
27	*Requires converting between different units of time (such as seconds and minutes) to calculate a total duration.*	$-1.456_{[—]}$	—	0.5

Continued on next page

Ours – GSM8K (continued)

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
28	*Requires tracking and calculating values for four or more distinct entities or categories.*	$1.408_{[0.589,\ 2.254]}$	0.787	8.0
29	*The question requires calculating a percentage of a quantity.*	$0.232_{[-0.449,\ 0.933]}$	0.127	14.5
30	*Requires adding an implicit +1 to account for the host or the original subject in a count of people.*	$1.305_{[—]}$	—	0.5
31	*The question requires working backwards from a total or a final state to find an intermediate or initial value, or solving an implicit algebraic equation where the unknown is not just the final result of a forward calculation.*	$0.884_{[0.273,\ 1.516]}$	0.488	9.5
32	*The question requires setting up an algebraic equation with a fractional coefficient (e.g., 2/3 * x = 12) to solve for an unknown.*	$2.797_{[2.376,\ 3.168]}$	1.556	1.5
33	*Requires calculating a baseline allowance or threshold, then determining the quantity of items that exceed it, and applying a rate only to the excess.*	$0.705_{[0.089,\ 1.350]}$	0.389	14.5
34	*Requires calculating a value based on a comparative phrase involving both a multiplier and an offset (e.g., 'more than twice the amount' or 'more than 5 times as much').*	$-0.023_{[-0.912,\ 0.902]}$	-0.013	4.0
35	*Requires arithmetic operations with negative numbers.*	$0.894_{[—]}$	—	0.5
36	*The question requires calculating an intermediate value that is not explicitly asked for but is necessary to find the final answer, often involving a sequence of dependent calculations where the output of one step is the input to the next (e.g., calculating the number of people at different times of day to find the total, calculating the time taken for different parts of a triathlon to find the final run time).*	$0.038_{[-0.912,\ 0.954]}$	0.021	93.5
37	*The question requires calculating a percentage or fraction of a previously calculated or updated value, rather than a static initial value.*	$0.120_{[-0.632,\ 0.951]}$	0.066	13.5
38	*Requires calculating a remaining budget after subtracting costs for four or more distinct line items.*	$3.561_{[—]}$	—	0.5
39	*Requires interpreting a rate applied to a fractional or partial unit (e.g., 'or part thereof').*	—	—	0.0
40	*The question involves a multi-step chain of reasoning where the result of one calculation is used as the input for the next calculation, often involving more than three distinct steps.*	$1.651_{[0.984,\ 2.315]}$	0.957	16.0

Table 18: The 12 hypotheses selected by our main-table model on BBH-structured, in selection order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
	BBH-structured			
1	*The question asks for a fact, property, or position that is explicitly stated in the text (e.g., 'The plums are the cheapest', 'What color is the textbook?'), requiring only direct lookup rather than logical deduction.*	$-1.278_{[-1.521,\ -1.032]}$	-1.206	20.0
2	*The question asks to identify an entity at a specific non-extreme ordinal position (e.g., 'second-to-last', 'third from the right') rather than at an extreme end (e.g., 'leftmost', 'cheapest').*	$0.565_{[0.257,\ 0.875]}$	0.492	28.2
3	*The question requires sorting a list of strings (e.g., names) alphabetically to determine the answer, rather than performing a simple numerical minimum/maximum operation or direct lookup.*	$2.214_{[0.757,\ 3.670]}$	1.905	0.7
4	*The question requires identifying an object based on its relative position to another specific object (e.g., 'directly to the right of', 'furthest from') rather than its absolute position in the sequence (e.g., 'left-most', 'right-most').*	$0.505_{[0.118,\ 0.888]}$	0.432	6.1
5	*The question requires reasoning about relative spatial or temporal positions (e.g., 'to the right of', 'between what times') rather than absolute attributes (e.g., 'is the jug mauve?').*	$0.738_{[0.462,\ 1.009]}$	0.658	62.1
6	*The question requires performing an operation over multiple elements, such as finding an unaccounted time gap in a schedule, summing values in a table, or counting rows after a deletion, rather than retrieving a single stated value.*	$0.449_{[0.187,\ 0.718]}$	0.387	27.5
7	*The problem involves a larger number of entities (e.g., seven objects) to order or track, increasing the state space compared to problems with fewer entities (e.g., three objects).*	$1.297_{[0.944,\ 1.650]}$	1.189	13.2
8	*The question requires counting or performing arithmetic operations on objects after applying a filtering condition.*	$0.521_{[0.107,\ 0.954]}$	0.447	10.7
9	*The question requires sorting items alphabetically or numerically before identifying a specific item, rather than retrieving an item based on its original order or explicit value.*	$2.214_{[0.758,\ 3.671]}$	1.905	0.7
10	*The question requires deducing the answer by combining multiple constraints or steps, whereas the easier question asks for a fact that is explicitly stated in the text.*	$-0.179_{[-1.071,\ 0.705]}$	-0.152	3.6
11	*Questions are harder when the answer choices require distinguishing between many plausible options (e.g., 7 to 17 options) rather than a small set of options (e.g., 2 to 5 options).*	$0.645_{[0.331,\ 0.951]}$	0.566	29.3
12	*The logical deduction task requires ordering a sequence of five or seven objects, rather than a simple sequence of only three objects.*	$1.366_{[1.081,\ 1.635]}$	1.335	23.6

Table 19: The 44 hypotheses selected by our main-table model on Winogrande, in selection order. SEP is the mean IRT difficulty of matched minus unmatched test items (95% bootstrap CI); Cohen's d is the standardized difference, and coverage is the matched-item percentage. Undefined statistics are shown as —; CI and d are omitted when either group has fewer than two items.

#	Hypotheses	SEP$_{[95\% \text{ CI}]}$	Cohen's d	Cov. (%)
	Winogrande			
1	*The sentence structure involves a complex causal chain where the blank refers to an entity that is the cause of a state described earlier in the sentence, requiring the model to infer the cause from the effect.*	$0.676_{[-0.262,\ 1.729]}$	0.402	4.3
2	*The easier question can be solved using strong lexical associations, common collocations, or direct causal links (e.g., 'tough stain', 'IOS' to 'iPhone', 'lazy' to 'failed').*	$-0.889_{[-1.583,\ -0.069]}$	-0.532	7.5
3	*The question requires multi-step causal reasoning to connect the premise to the resolution, where the solver must infer an unstated intermediate state or action (e.g., hating a food means eating less of it, which leaves more to harvest; or accepting a challenge implies believing one can win), rather than relying on direct semantic associations or attribute matching (e.g., ice is cold, or an animal rights activist dislikes leather).*	$0.916_{[0.449,\ 1.393]}$	0.559	26.4
4	*The question can be resolved using basic selectional restrictions or inherent physical properties of the entities (e.g., a cloth is soft, a video is boring, a color is bright) rather than requiring causal reasoning about the events described.*	$-0.872_{[-1.456,\ -0.220]}$	-0.524	11.8
5	*The harder question involves resolving a pronoun based on an action that caused regret or a negative emotional reaction (e.g., feeling bad and vowing not to do it again).*	—	—	0.0
6	*The harder question requires reasoning about implicit physical, biological, or social mechanisms (e.g., relative hardness for scratching, evergreen vs. deciduous traits, professional vs. amateur competence, or the implications of acting in someone else's name), whereas the easier question relies on direct lexical associations or common sense (e.g., library for a paper, productive means hard worker).*	—	—	0.0
7	*The question requires multi-step causal reasoning where the blank is resolved by inferring an implicit intermediate state (e.g., fewer items means they take up more space, or asking about a wedding implies the other person is the groom).*	$0.386_{[-0.038,\ 0.817]}$	0.230	31.9
8	*Questions are harder when the causal relationship requires inferring a missing intermediate step or implicit consequence (e.g., 'cows hated eating the corn' -¿ 'farmer had more corn to harvest', 'sunglasses were poor' -¿ 'eyes could not see clearly') rather than a direct, explicit cause-and-effect relationship.*	$0.427_{[0.026,\ 0.835]}$	0.255	46.9
9	*The question requires complex social or interpersonal reasoning (such as inferring roles, intentions, or dynamics) rather than relying on direct physical properties or straightforward cause-and-effect.*	$0.881_{[0.257,\ 1.503]}$	0.531	13.8
10	*The sentence contains a semantic contradiction or counter-intuitive resolution where the correct referent violates typical commonsense expectations or requires strict adherence to syntactic structure over semantic plausibility.*	—	—	0.0
11	*The question requires resolving the pronoun or blank based on a counter-intuitive, anti-stereotypical, or unexpected causal relationship (e.g., having more of a crop because animals hated eating it, or an item failing to fit because the inner layer was too large rather than the outer layer being too small).*	$2.640_{[0.569,\ 4.557]}$	1.588	1.2
12	*The harder question requires understanding implicit social, economic, or legal relationships (e.g., the cost implications of a gift vs. a card, or the distinction between a homeowner and a tenant).*	$-0.448_{[-1.329,\ 0.426]}$	-0.266	0.8
13	*The question can be answered using direct lexical or semantic associations between the clue and the correct option (e.g., 'chainsaw' and 'tree cutter', 'wood' and 'rotten', 'video' and 'boring'), which makes it easier, whereas harder questions require reasoning about the specific situational dynamics or causal relationships described in the sentence.*	$-0.571_{[-1.043,\ -0.083]}$	-0.342	22.8
14	*The harder question requires resolving a reference based on explicit possession or attribution in the text rather than general causal reasoning about traits.*	—	—	0.0
15	*The easier question can be solved by directly matching a straightforward descriptive trait, profession, or physical state to the person exhibiting the corresponding behavior, without needing to infer complex interactions.*	$-0.604_{[-1.176,\ 0.069]}$	-0.360	7.5
16	*The easier question uses explicit contrast structures (e.g., 'instead of', 'unlike') where the blank is resolved by mapping a straightforward positive or negative trait directly to the accepted or rejected entity.*	$0.398_{[-0.825,\ 2.399]}$	0.236	1.2
17	*The question requires multi-step causal or physical reasoning about the implicit properties of objects or actions (e.g., that bronzer is a fast way to tan, or that soaking an object transfers liquid from the container) rather than relying on direct lexical associations.*	$0.266_{[-0.178,\ 0.720]}$	0.158	30.7
18	*The question requires complex situational or social commonsense reasoning to infer roles and goals (e.g., understanding what a 'ride along' implies or the implications of doing something 'in place of' someone else).*	$0.236_{[-0.486,\ 0.998]}$	0.140	9.4

Continued on next page

Ours – Winogrande (continued)

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
19	*The question requires inferring an implicit goal or intention behind an action involving multiple agents, locations, or domains, rather than relying on explicit causal markers.*	$-0.516_{[-1.117,\ 0.067]}$	-0.306	2.4
20	*The question requires multi-step reasoning or specific domain knowledge to connect the clue to the correct entity (e.g., knowing that 'saccharine' means sweet and fruits are sweet, or that snowboarding requires going up a hill, or that commodities are bought when priced low), rather than relying on direct, everyday commonsense associations (e.g., returning an item because it is the wrong size, or changing a filter because it is dirty).*	$0.319_{[-0.345,\ 1.003]}$	0.189	8.7
21	*The question requires reasoning about physical, spatial, or quantitative relationships where the correct answer depends on an inverse correlation (e.g., having more of one item implies the other item takes up more space, or transferring a substance to one object means the source object has less).*	$0.832_{[0.145,\ 1.530]}$	0.498	10.2
22	*The easier question can be resolved largely through direct semantic selectional restrictions or strong lexical associations between the predicate and the correct entity (e.g., a video is 'boring', a street is 'long', birds fly to 'bird seed'), without needing to integrate the full causal logic of the sentence.*	$-0.998_{[-1.857,\ 0.096]}$	-0.597	5.9
23	*The harder questions often require the model to infer the antecedent based on the logical consequence of an action that is not explicitly stated but must be deduced from world knowledge (e.g., a person who fled a raid avoids arrest, a person who asks a question lacks knowledge).*	$0.627_{[0.214,\ 1.036]}$	0.378	52.4
24	*The easier question relies on direct physical constraints, basic definitions, or highly conventional event scripts (e.g., a door being too small for a bed, a cashier making less money, returning an item to the person who loaned it).*	$-0.762_{[-2.075,\ 1.423]}$	-0.452	1.6
25	*The sentence structure involves a comparison where the blank refers to the entity that is less prominent or requires reasoning about the negative/lesser side of the comparison.*	$0.170_{[-0.493,\ 0.865]}$	0.101	9.4
26	*The correct resolution of the blank violates typical semantic or physical expectations, such as a person being 'too short to climb over' a fence rather than the fence being too short, forcing the model to override strong lexical priors.*	$2.640_{[0.567,\ 4.539]}$	1.588	1.2
27	*The question requires a multi-step inference combining a stated premise with implicit background knowledge about a role or action (e.g., inferring that a worse baker makes a mess, or that a cashier accepts real bills and real bills have sharp lines).*	$0.445_{[0.023,\ 0.859]}$	0.266	36.6
28	*The easier question contains explicit causal conjunctions ('because', 'since', 'as') that directly link a stated preference, action, or outcome to an attribute of the correct answer.*	$-0.659_{[-1.349,\ 0.088]}$	-0.393	9.8
29	*The harder questions frequently involve implicit causal reasoning where the blank refers to the cause of a negative outcome or failure (e.g., 'was not enough', 'lacked the aptitude', 'was boring'), requiring the model to infer which entity possesses the negative attribute.*	$0.205_{[-0.264,\ 0.699]}$	0.121	23.2
30	*The correct resolution contradicts a strong local syntactic or semantic expectation, such as a counter-intuitive physical description ('the thief was too short' rather than the fence) or an unexpected source of a trait.*	$4.577_{[—]}$	—	0.4
31	*The harder questions require resolving a pronoun or blank to an entity that is the object or patient of the preceding clause, whereas easier questions often resolve to the subject or agent.*	$0.110_{[-0.311,\ 0.518]}$	0.065	39.4
32	*The question requires resolving a reference based on implicit social roles, physical constraints, or metonymy rather than direct descriptions (e.g., knowing that a 'best man' implies the other person is the groom, or describing a 'sprain' as 'fat'), which increases the difficulty.*	$0.442_{[-0.069,\ 0.982]}$	0.263	18.1
33	*The question requires inverse or counter-intuitive causal reasoning, where the correct answer depends on an indirect relationship (e.g., inferring that having more of one item means the other item is bigger, or that socks being 'too large' prevents boots from fitting, rather than the boots being too small).*	$1.029_{[0.460,\ 1.613]}$	0.626	16.1
34	*The correct answer requires resolving a pronoun or implicit reference to the entity that was NOT chosen, NOT affected, or NOT the primary subject of the main clause (e.g., the cheaper item, the person who didn't get sick, the vehicle not driven).*	$-0.225_{[-0.744,\ 0.289]}$	-0.133	15.0
35	*The easier question contains strong, direct lexical or causal cues linking the blank to the correct entity (e.g., 'returned' strongly implies the other person 'loaned' it; 'scary' directly explains why someone went on a ride less).*	$-1.950_{[-2.239,\ -1.667]}$	-1.165	1.2
36	*The harder question involves a potentially counter-intuitive or highly specific causal attribution that defies typical expectations, making the reference resolution much more difficult for models to predict correctly.*	—	—	0.0
37	*The correct answer requires inferring a contrast or counter-intuitive relationship, such as a concession or an indirect consequence, rather than a direct causal link.*	$0.737_{[-0.098,\ 1.683]}$	0.438	2.8
38	*The sentence requires reasoning about a comparative degree or inverse relationship (e.g., 'more than', 'less') to determine which entity possesses the trait that resolves the blank.*	$-0.106_{[-0.545,\ 0.346]}$	-0.063	25.6

Continued on next page

Ours – Winogrande (continued)

#	Hypotheses	$\mathbf{SEP}_{[95\%\ \mathrm{CI}]}$	Cohen's d	Cov. (%)
39	*The question requires inverse reasoning or understanding a negative constraint, where the correct referent is identified by its association with an undesirable outcome, a lack of a trait, or an avoided option (e.g., 'instead of', 'unlike', 'sounded quieter', 'too short to climb over').*	$0.266_{[-0.193,\ 0.728]}$	0.158	25.6
40	*The sentence involves complex interpersonal dynamics where one person's action or trait directly causes the other person's reaction, requiring the reader to carefully assign distinct roles (e.g., cause vs. experiencer) to resolve the pronoun.*	$0.742_{[0.199,\ 1.292]}$	0.447	19.7
41	*The easier question involves resolving a reference based on physical properties or direct physical actions (e.g., melting, falling), whereas the harder question involves abstract states, intentions, or social situations.*	$-0.249_{[-1.539,\ 0.963]}$	-0.148	1.6
42	*The correct answer requires resolving a pronoun or blank to an entity that is the subject of a subordinate clause, where the causal relationship is counterintuitive or requires overriding a strong local association (e.g., 'cows hated eating the corn' leading to more corn, or 'sprain is too fat' instead of 'foot').*	$4.577_{[—]}$	—	0.4
43	*The question involves complex causal reasoning about constraints and degrees of difficulty (e.g., deducing that having less money makes a budget harder to follow).*	$2.197_{[—]}$	—	0.4
44	*The easier question relies on direct lexical associations or straightforward commonsense properties (e.g., cashier = less money, lack of light = dark room, sneaky = coming up behind).*	$-1.340_{[-1.629,\ -1.022]}$	-0.798	1.6