Icelandic Math Eval: A Competitive Mathematics Benchmark for Large Language Models

Hafsteinn Einarsson*, Jökull Ari Haraldsson†, Ívar Armin Derayat†, Sigrún Helga Lund†, Benedikt Steinar Magnússon†

*Department of Computer Science, University of Iceland, Reykjavík, Iceland
hafsteinne@hi.is

†Department of Mathematics, University of Iceland, Reykjavík, Iceland
{jah39, iad3, sigrunhl, bsm}@hi.is

Abstract

We introduce Icelandic Math Eval, the first comprehensive benchmark for evaluating large language models (LLMs) on competitive mathematics problems in Icelandic. Our dataset comprises 1,027 problems from Icelandic mathematics competitions spanning from 1984 to 2025, covering algebra, geometry, number theory, and combinatorics across ten difficulty levels. We evaluate three state-of-the-art models, Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-5, using a dual evaluation methodology that tests both with and without multiple-choice options. Our results reveal several key findings: (1) models achieve 81-93% overall accuracy, demonstrating substantial cross-lingual transfer of mathematical reasoning capabilities; (2) a dramatic 17.5 percentage point performance drop on problems containing images highlights persistent challenges in multimodal mathematical reasoning; (3) a 6.7 percentage point gap between evaluation modes suggests that multiple-choice formats may overestimate genuine reasoning capabilities; and (4) systematic performance degradation with increasing difficulty, dropping to 43% on the most challenging problems. Using an LLM-as-judge evaluation approach, we provide detailed analysis across problem types, difficulty levels, and model capabilities. This work contributes to multilingual AI evaluation and demonstrates the importance of developing rigorous benchmarks for diverse languages to ensure comprehensive assessment of AI capabilities.

Keywords: mathematical reasoning, multilingual evaluation, Icelandic, benchmark dataset, large language models

1. Introduction

The evaluation of large language models (LLMs) on mathematical reasoning tasks has emerged as a critical benchmark for assessing their problem-solving capabilities (Cobbe et al., 2021; Hendrycks et al., 2021b). Recent years have seen the development of increasingly challenging benchmarks to test the limits of AI reasoning, including ARC-AGI (Chollet, 2019), which evaluates abstract reasoning through visual puzzles, GPQA (Rein et al., 2024), a graduate-level question-answering benchmark designed to be "Google-proof," and Humanity's Last Exam (Phan et al., 2025), which features expert-level questions across multiple domains. In mathematics specifically, competitions like AIME (Hendrycks et al., 2021b) provide rigorous tests of advanced problem-solving abilities. While significant progress has been made in developing comprehensive mathematical benchmarks for English, the landscape for other languages remains sparse, particularly for low-resource languages. This disparity raises important questions about the generalizability of LLM capabilities across linguistic boundaries and the potential for inequitable access to AI-powered mathematical assistance across different language communities.

Icelandic, spoken by approximately 350,000 people, represents an interesting case study for multilingual mathematical reasoning. Despite its relatively small speaker population, Iceland has a rich tradition of mathematics education and competitive mathematics, with organized competitions dating back several decades. However, no standardized benchmark exists for evaluating LLM performance on Icelandic mathematical problems, limiting our understanding of how these models handle mathematical reasoning in this language.

In this paper, we introduce **Icelandic Math Eval**, the first comprehensive benchmark for evaluating LLMs on competitive mathematics problems in Icelandic. Our dataset[1] comprises problems from Icelandic mathematics competitions spanning from 1984 to the present, covering various difficulty levels and mathematical domains.

We evaluate several state-of-the-art LLMs on our benchmark, including Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-5. Our experiments reveal significant performance disparities across models. These findings highlight the challenges that current models face when processing mathematical concepts expressed in languages with limited training data.

Our contributions are threefold:

[1] https://github.com/Haffi112/icelandic-math-eval

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 393–406
11–16 May 2026.

- We present the first competitive mathematics benchmark for Icelandic, comprising problems from national competitions spanning four decades.
- We provide a comprehensive evaluation of state-of-the-art LLMs on this benchmark.
- We release our dataset and evaluation framework to facilitate future research in multilingual mathematical reasoning.

2. Related Work

Our work builds upon three main research areas: mathematical reasoning benchmarks, multilingual evaluation of LLMs, and Icelandic natural language processing.

2.1. Mathematical Reasoning Benchmarks

The evaluation of mathematical reasoning in LLMs has evolved rapidly with the introduction of increasingly sophisticated benchmarks. Cobbe et al. (2021) introduced GSM8K, a dataset of 8,500 grade school math word problems that has become a standard benchmark for evaluating basic mathematical reasoning. The problems in GSM8K require 2-8 steps of reasoning and cover fundamental arithmetic operations, making it an accessible entry point for mathematical evaluation.

For more advanced mathematical reasoning, Hendrycks et al. (2021b) developed the MATH dataset, comprising 12,500 challenging competition mathematics problems sourced from contests such as AMC 10/12 and AIME. This benchmark includes problems requiring sophisticated mathematical knowledge across algebra, geometry, number theory, and probability, with detailed step-by-step solutions that enable fine-grained evaluation of reasoning processes.

Hendrycks et al. (2021a) introduced MMLU (Massive Multitask Language Understanding), which includes mathematical subtasks spanning elementary to college-level mathematics. While MMLU provides broad coverage across multiple domains, recent analyses (Gema et al., 2025) have identified quality issues, with approximately 6.5% of ground truth answers found to contain errors, highlighting the challenges in creating reliable evaluation benchmarks.

The emergence of chain-of-thought prompting, instructing models to produce intermediate reasoning steps before arriving at a final answer (Wei et al., 2022), has significantly improved LLM performance on mathematical tasks by encouraging models to generate intermediate reasoning steps. For instance, Wei et al. demonstrated that chain-of-thought prompting improved performance on GSM8K from 17.9% to 58.1% using PaLM 540B, with further improvements to 74% when combined with self-consistency. Recent meta-analyses (Sprague et al., 2025) confirm that chain-of-thought helps mainly on math and symbolic reasoning tasks. However, modern LLMs have been incorporated with thought processes that replace the need for explicit chain-of-thought prompting during inference.

2.2. Multilingual Mathematical Evaluation

The extension of mathematical benchmarks to multilingual settings has revealed important insights about the language-dependence of mathematical reasoning. Shi et al. (2023) introduced MGSM (Multilingual Grade School Math), a manually translated version of 250 GSM8K problems into 10 typologically diverse languages. Their work demonstrated that while LLMs can perform mathematical reasoning in multiple languages, performance typically degrades compared to English, with the degradation varying by language and model architecture.

Building on this work, Luo et al. (2025) developed MMATH, addressing the limitation that MGSM had become too easy for contemporary models. MMATH comprises 374 high-quality problems across 10 languages, specifically designed to be challenging for current state-of-the-art models while maintaining cultural and linguistic appropriateness across languages.

Recent work has also explored cross-lingual transfer in mathematical reasoning, investigating whether mathematical reasoning capabilities learned in one language can transfer to others. These studies suggest that while some transfer occurs, language-specific training data remains crucial for optimal performance, particularly for morphologically rich languages.

2.3. Icelandic Natural Language Processing

The Icelandic Language Technology Programme 2019-2023 (Nikulásdóttir et al., 2020) has been instrumental in advancing NLP resources for Icelandic, establishing comprehensive infrastructure and tools for this morphologically rich language. A significant development in this effort was the creation of Natural Questions in Icelandic (NQiI) (Snæbjarnarson and Einarsson, 2022), which provides 18,000 labeled question-answer pairs adapted for Icelandic. This work emphasizes the typological diversity of Icelandic compared to languages in existing multilingual benchmarks.

Notably, while the TyDi QA dataset (Clark et al., 2020) covers 11 typologically diverse languages, it does not include Icelandic, highlighting a gap in

coverage for Nordic languages with complex morphology. This omission is particularly relevant when considering that although MGSM (Shi et al., 2023) addressed ten typologically diverse languages, the unique linguistic features of Icelandic (including its rich inflectional system and preserved Germanic case structure) may present distinct challenges for mathematical reasoning that are not captured in existing multilingual benchmarks.

The development of IceBERT (Snæbjarnarson et al., 2022) has served as a foundation for various downstream tasks including question answering (Snæbjarnarson, 2021). However, the field has increasingly transitioned towards generative models in recent years, following global trends in NLP. This shift presents both opportunities and challenges for low-resource languages like Icelandic, where the benefits of language-specific pretraining must be balanced against the computational costs and data requirements of large generative models.

The broader Nordic language community has also developed evaluation resources. Nielsen (2023) introduced ScandEval, a comprehensive benchmark for Scandinavian languages including Icelandic. While ScandEval includes various NLP tasks, it does not specifically address mathematical reasoning, highlighting the gap that our work aims to fill.

3. Dataset Construction

3.1. Data Collection

We collected mathematical problems from various Icelandic mathematics competitions spanning from 1984 to 2025. Our sources include:

- National high school mathematics competitions (Stærðfræðikeppni framhaldsskólanema)
- Regional mathematics olympiads
- Historical competition archives maintained by the Icelandic Mathematical Society

Problems were digitized from printed materials when necessary, with careful attention to preserving mathematical notation and problem statements exactly as originally presented. Each problem was verified by at least two native Icelandic speakers with mathematics backgrounds to ensure accuracy. The dataset spans 41 competition files collected over four decades, providing a comprehensive view of Icelandic mathematical competition problems.

3.2. Dataset Statistics

Our final dataset comprises **1,027 problems** categorized by multiple dimensions:

Difficulty levels: Problems are classified into ten difficulty tiers (Level 1-10) based on their original competition level and expected solution complexity. The distribution shows concentration in mid-range difficulties, with 99 problems at Level 1, 206 at Level 2, 494 at Level 3, and 242 at Level 4, declining to just 1 problem at Level 10. This distribution reflects the natural difficulty progression in competitive mathematics.

Mathematical domains: Following standard competition mathematics categorization (Hendrycks et al., 2021b), problems are classified into four primary types:

- *Algebra*: 395 problems (38.5%)
- *Number Theory*: 209 problems (20.3%)
- *Geometry*: 308 problems (30.0%)
- *Combinatorics*: 117 problems (11.4%)

Answer formats: Problems feature two answer types: multiple-choice questions (847 problems, 82.5%) with four options each, and numeric answer questions (180 problems, 17.5%) requiring precise numerical responses.

Multimodal content: A significant portion of our dataset includes visual elements. Of the total problems, 273 (26.6%) contain images such as geometric diagrams, graphs, or visual problem representations, while 754 (73.4%) are text-only. This multimodal aspect allows us to evaluate models' capabilities in visual mathematical reasoning (Lu et al., 2024), which has emerged as a critical challenge for contemporary LLMs.

Each problem in our dataset includes the correct answer and source information, which are released alongside the problems to facilitate reproducible evaluation.

3.3. Quality Assurance

To ensure dataset quality, we implemented a multi-stage verification process:

1. **Solution verification**: Each problem's solution was independently verified by at least two reviewers.
2. **Answer format validation**: We validated that multiple-choice options were properly formatted and that numeric answers were specified with appropriate precision.
3. **Difficulty**: Each problem was rated on a scale from 1 to 10 by one undergraduate mathematics student and two students with a B.Sc. degree in mathematics and the assigned difficulty level was based on their consensus.

4. Evaluation Methodology

4.1. Model Selection

We evaluate three state-of-the-art large language models representing different architectural approaches and training paradigms:

- **Anthropic Claude Sonnet 4.5**: A recent flagship model from Anthropic known for strong reasoning capabilities.
- **Google Gemini 2.5 Pro**: Google's multimodal model with enhanced image understanding and mathematical reasoning.
- **OpenAI GPT-5**: The latest generation model from OpenAI, demonstrating significant advances in mathematical problem-solving.

Claude and Gemini were accessed via OpenRouter and GPT-5 via its respective API. We used temperature = 1 in the evaluation.

4.2. Evaluation Protocol

To comprehensively assess model capabilities, we employ a dual evaluation mode strategy that tests both constrained and open-ended reasoning:

With Choices Mode: Models are presented with the complete problem including all four multiple-choice options (for multiple-choice questions) or the full problem context (for numeric questions). This mode simulates standardized test-taking scenarios where models can leverage elimination strategies and pattern matching alongside mathematical reasoning.

Without Choices Mode: Models receive only the problem statement without multiple-choice options, requiring them to generate answers independently. For multiple-choice questions, models must produce the answer without seeing the choices, which is then matched against the correct option. For numeric questions, the evaluation is identical to the with-choices mode. This mode provides a more stringent test of genuine mathematical understanding, as models cannot rely on option elimination or recognition (Zhang et al., 2024).

4.3. LLM-as-Judge Evaluation

Given the diversity of answer formats and the potential for valid alternative formulations, we employ an LLM-as-judge evaluation methodology (Li et al., 2024a) to assess answer correctness. This approach has become increasingly prevalent in evaluating open-ended mathematical reasoning, offering advantages over strict string-matching methods.

Our primary judge (GPT-5) evaluates each model response by comparing it against the ground truth answer while accounting for:

- **Numerical equivalence**: Different representations of the same numerical value (e.g., fractions vs. decimals, simplified vs. unsimplified forms)
- **Mathematical notation**: Various valid ways to express mathematical concepts
- **Extraction from reasoning chains**: Identifying the final answer within chain-of-thought reasoning when models provide detailed solution steps

The judge extracts and validates the final answer from the complete response. All accuracy figures reported in this paper use GPT-5 judgments.

Because GPT-5 also serves as one of the evaluated models, its role as judge raises a potential self-assessment bias. To address this concern, we conducted a multi-judge validation study using two additional independent judges: Gemini 3 Flash and Claude Sonnet 4.6. Each judge independently evaluated all model responses under identical instructions. Table 1 reports pairwise and three-way agreement.

Agreement rates range from 98.8% to 98.9%, with pairwise Cohen's κ between 0.939 and 0.947, and Fleiss' $\kappa = 0.944$ across all three judges. These values indicate near-perfect consistency in correctness judgments, regardless of which model serves as judge. The high agreement between GPT-5 and the two independent judges suggests that self-assessment bias does not materially affect the reported results.

5. Results

We present a comprehensive analysis of model performance across multiple dimensions: overall accuracy, evaluation mode effects, difficulty scaling, problem type variations, and multimodal reasoning capabilities.

5.1. Impact of Evaluation Mode

The evaluation mode significantly affects measured performance. As shown in Figure 1, all three models demonstrate higher accuracy when provided with multiple-choice options. GPT-5 achieves 96.01% accuracy with choices but drops to 89.19% without them (6.82 percentage point decrease). Gemini 2.5 Pro shows the smallest gap, declining from 93.48% to 88.61% (4.87 points), while Claude Sonnet 4.5 exhibits the largest drop from 85.20% to 76.92% (8.28 points).

This substantial gap aligns with recent findings on multiple-choice evaluation biases (Zhang et al., 2024), suggesting that models benefit from seeing answer options through elimination strategies

Judge Pair	N	Agreement (%)	Cohen's κ
GPT-5 vs. Gemini 3 Flash	6161	98.8	0.939
GPT-5 vs. Claude Sonnet 4.6	6153	98.9	0.947
Gemini 3 Flash vs. Claude Sonnet 4.6	6152	98.9	0.946
Fleiss' κ (all 3 judges, $N = 6152$): 0.944			

Table 1: Inter-judge agreement between three LLM judges (GPT-5, Gemini 3 Flash, and Claude Sonnet 4.6) on correctness judgments. High agreement rates and substantial Cohen's κ values indicate consistency across judge models.

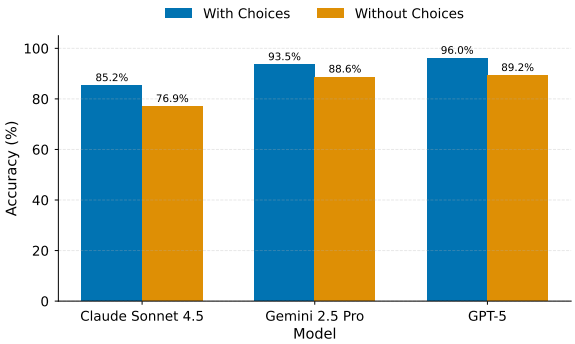

Figure 1: Performance comparison across evaluation modes for each model. All models show decreased accuracy when multiple-choice options are not provided, but the magnitude varies by model.

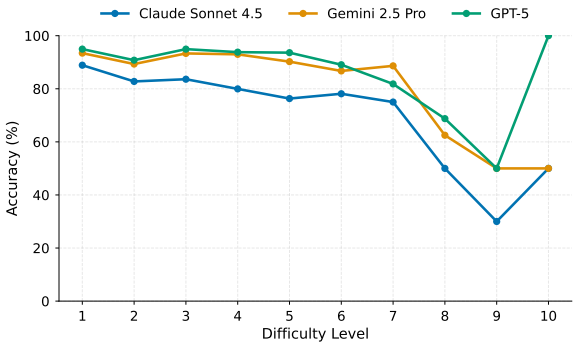

Figure 2: Performance degradation across difficulty levels for all three models. While all models handle easier problems well, performance drops substantially for the most challenging problems (Levels 8-10).

and pattern recognition beyond pure mathematical reasoning.

5.2. Performance by Difficulty Level

Performance degrades systematically with increasing problem difficulty, as shown in Table 2 and Figure 2. The dataset contains problems ranging from Level 1 (99 problems) to Level 10 (1 problem), with the majority concentrated in Levels 2-4 (206, 247, and 242 problems respectively).

Examining individual models reveals distinct scaling behaviors across the difficulty spectrum. GPT-5 maintains the strongest performance, achieving 94.9% at Level 1 and declining more gradually to 68.8% at Level 8, while remarkably achieving 100% on the single Level 10 problem. Gemini 2.5 Pro demonstrates consistent strong performance through mid-level difficulties (93.4% at Level 1, maintaining above 90% through Level 5), but drops to 62.5% at Level 8. Claude Sonnet 4.5 shows the steepest decline, from 88.9% at Level 1 to 50.0% at Level 8. Notably, all models struggle significantly with Level 9 problems (5 problems, 30-50% accuracy), indicating that the most challenging competition problems remain beyond current model capabilities.

5.3. Performance by Problem Type

Mathematical domain significantly influences performance, as detailed in Figure 3. The dataset contains 395 algebra problems, 209 number theory problems, 308 geometry problems, and 117 combinatorics problems. Performance varies substantially across these domains.

All models perform best on algebra and number theory. GPT-5 achieves 95.7% on algebra and 94.5% on number theory, with Gemini 2.5 Pro close behind at 95.2% and 92.6% respectively. Claude Sonnet 4.5 maintains 89.4% and 89.2% on these domains. The performance gaps widen considerably for geometry and combinatorics: GPT-5 reaches 89.6% on geometry and 86.8% on combinatorics, while Claude Sonnet 4.5 drops to 70.3% and 67.1% respectively. Gemini 2.5 Pro demonstrates balanced performance across all domains, maintaining above 86% accuracy even on the more challenging problem types.

5.4. Multimodal Reasoning: Impact of Images

The presence of visual elements dramatically affects model performance, as shown in Figure 4. The dataset contains 273 problems with images and 754 text-only problems, allowing for compre-

Model	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Total	99	206	247	242	133	64	22	8	5	1
Claude Sonnet 4.5	88.9%	82.8%	83.6%	80.0%	76.3%	78.1%	75.0%	50.0%	30.0%	50.0%
Gemini 2.5 Pro	93.4%	89.3%	93.3%	93.0%	90.2%	86.7%	88.6%	62.5%	50.0%	50.0%
GPT-5	94.9%	90.8%	94.9%	93.8%	93.6%	89.1%	81.8%	68.8%	50.0%	100.0%

Table 2: Accuracy (%) by difficulty level for each model. Top row shows total problem count per level.

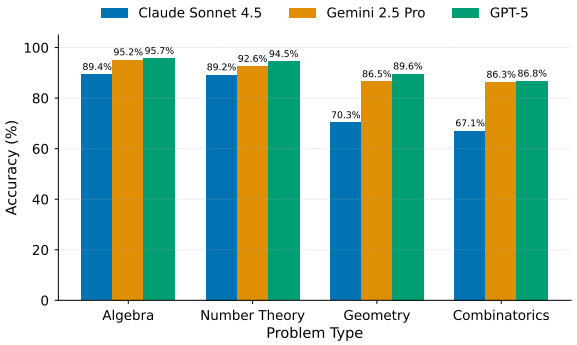

Figure 3: Performance variation across mathematical domains. All models perform best on algebra and number theory, with greater challenges in geometry and combinatorics.

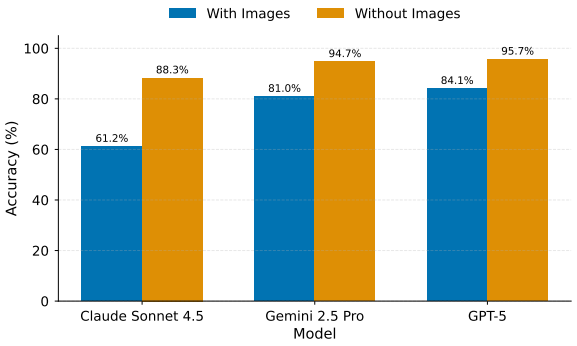

Figure 4: Impact of visual elements on model performance. All models show substantial performance degradation on problems containing images, with Claude Sonnet 4.5 experiencing the largest drop.

hensive evaluation of multimodal reasoning capabilities.

All models struggle with image-based problems, but to varying degrees. Claude Sonnet 4.5 shows the largest performance gap: 61.2% accuracy with images versus 88.3% without (27.1 percentage point difference). Gemini 2.5 Pro performs better but still shows a substantial gap: 81.0% with images versus 94.7% without (13.7 points). GPT-5 demonstrates the strongest multimodal capabilities at 84.1% with images versus 95.7% without (11.6 points), though even this represents a significant performance degradation.

These findings align with recent work on multimodal mathematical reasoning (Lu et al., 2024; Zhang et al., 2025; Sun et al., 2024), which has identified visual understanding as a persistent challenge for LLMs. The geometry problems in our dataset, which most frequently include diagrams, show correspondingly lower accuracy (82.14%) compared to algebra and number theory (93.42% and 92.11%), suggesting that visual reasoning difficulties contribute substantially to domain-specific performance variations.

5.5. Performance by Answer Type

Both answer types achieve similar overall accuracy across all evaluations: multiple-choice questions at 88.33% (4,489/5,082 correct) and numeric questions at 87.78% (948/1,080 correct). However, this similarity masks important differences across models. GPT-5 actually performs *better* on numeric questions (95.28%) than multiple-choice (92.03%), suggesting stronger capabilities in generating precise numerical answers. In contrast, Claude Sonnet 4.5 shows slightly better performance on multiple-choice (81.46%) versus numeric questions (79.17%), and the same applies to Gemin 2.5 Pro (91.50% vs. 88.89%).

6. Discussion

Our evaluation of state-of-the-art LLMs on Icelandic mathematical problems reveals several important findings with implications for multilingual AI development, evaluation methodology, and the fundamental capabilities of current models.

6.1. The Multimodal Reasoning Challenge

Perhaps our most striking finding is the dramatic 17.48 percentage point performance drop on problems containing images compared to text-only problems. This gap persists across all models, though with varying magnitudes: Claude Sonnet 4.5 shows a 27.09 point drop, Gemini 2.5 Pro 13.74 points, and GPT-5 11.62 points. These results suggest that despite significant advances in multimodal architectures, visual mathematical reasoning remains a fundamental challenge.

The difficulty manifests particularly in geometry problems, which achieve only 82.14% accuracy compared to 93.42% for algebra. Given that geometry problems most frequently include diagrams showing spatial relationships, angles, and geometric constructions, this performance gap likely reflects difficulties in accurately extracting and reasoning about visual information. Recent work on multimodal mathematical benchmarks (Lu et al., 2024; Zhang et al., 2025) has identified similar challenges, noting that models struggle with diagram interpretation, spatial reasoning, and integrating visual and textual information.

6.2. Evaluation Methodology Matters

The 6.65 percentage point performance difference between evaluation modes raises important questions about what mathematical reasoning benchmarks actually measure. When models see multiple-choice options, they achieve 91.56% accuracy, but this drops to 84.91% when required to generate answers independently. This gap aligns with recent findings (Zhang et al., 2024) on multiple-choice evaluation biases and suggests that some of the measured performance reflects test-taking strategies such as option elimination, pattern recognition, and educated guessing rather than pure mathematical reasoning.

The variation across models is revealing. Gemini 2.5 Pro shows the smallest gap (4.87 points), suggesting more consistent reasoning capabilities regardless of answer format. Claude Sonnet 4.5's larger gap (8.28 points) indicates greater reliance on seeing answer options. These differences have practical implications: benchmarks using only multiple-choice evaluation may overestimate model capabilities on open-ended mathematical problem-solving.

We recommend that future mathematical reasoning benchmarks employ dual evaluation modes, as our methodology does, to provide a more complete picture of model capabilities. The without-choices mode offers a more stringent test of genuine mathematical understanding, while the with-choices mode better reflects performance on standardized tests and structured assessments.

6.3. Difficulty Scaling and Model Capabilities

The systematic performance degradation with increasing difficulty reveals important insights about model capabilities and limitations. While models achieve over 90% accuracy on easier problems (Levels 1-3), performance drops substantially at higher levels, reaching just 43.33% on Level 9 problems and 60.42% on Level 8 problems. We note that Levels 8–10 contain only 8, 5, and 1 problem(s) respectively, so these figures should be interpreted with caution. This scaling behavior suggests that current models have largely mastered routine competitive mathematics problems but struggle with problems requiring deeper insight, creative problem-solving approaches, or multiple sophisticated reasoning steps.

GPT-5's more gradual decline (94.95% at Level 1 to 68.75% at Level 8) compared to Claude Sonnet 4.5 (88.89% to 50.00%) suggests architectural or training differences that better support complex mathematical reasoning. However, even the best-performing model shows substantial degradation at the highest difficulty levels, indicating that the most challenging competition problems remain beyond current capabilities.

This finding has implications for AI capabilities assessment. While headlines often emphasize high performance on mathematical benchmarks, our difficulty-stratified analysis reveals that models' capabilities are far from uniform. Problems requiring advanced techniques, non-obvious insights, or creative approaches, precisely the problems that distinguish strong mathematical reasoners, remain challenging for current systems.

6.4. Domain-Specific Performance Patterns

The substantial variation across mathematical domains, algebra at 93.42%, number theory at 92.11%, geometry at 82.14%, and combinatorics at 80.06%, likely reflects both training data distributions and inherent problem characteristics. Algebra problems often follow recognizable patterns and solution templates that may be well-represented in training data. In contrast, combinatorics problems typically require creative problem decomposition and less formulaic approaches, potentially explaining their lower accuracy.

The particularly strong performance on algebra across all models (Claude Sonnet 4.5: 89.37%, Gemini 2.5 Pro: 95.19%, GPT-5: 95.70%) suggests that algebraic manipulation and equation-solving capabilities are robust even in Icelandic. However, the widening gaps for geometry and combinatorics, where Claude Sonnet 4.5 drops to 70.29% and 67.09% respectively, indicate that these domains present compounding challenges, possibly combining visual reasoning difficulties (for geometry) with limited training examples (for both domains).

6.5. Implications for Low-Resource Language AI

Our results contribute to understanding how LLMs handle mathematical reasoning in low-resource languages. With approximately 350,000 speakers,

Icelandic represents a language with limited training data compared to English. The fact that models achieve 81-93% overall accuracy demonstrates that mathematical reasoning capabilities transfer across languages to a substantial degree, likely because mathematical concepts transcend linguistic barriers.

However, comparing our results to English mathematical benchmarks reveals performance gaps. Recent evaluations on the American Invitational Mathematics Examination (AIME), a prestigious high-school mathematics competition, show frontier models achieving over 90% accuracy on these challenging problems with GPT-5 scoring 93.4%. While direct comparison is complicated by differences in problem distributions and difficulty calibration, our finding that even the best model in our evaluations (GPT-5 at 92.60%) shows declining performance on higher-difficulty problems (dropping to 68.8% at Level 8 and 43.3% at Level 9) suggests that advanced mathematical reasoning remains challenging across languages. Recent work has introduced increasingly sophisticated benchmarks beyond AIME, including FrontierMath (Glazer et al., 2024), where current models solve less than 2% of research-level problems, OlymMATH (Sun et al., 2025) with rigorous olympiad-level problems, MATH-Vision (Wang et al., 2024a) for multimodal mathematical reasoning with 3,040 problems across 16 disciplines, and MathOdyssey (Fang et al., 2025) spanning high school to olympiad levels. These benchmarks demonstrate that substantial challenges remain in mathematical reasoning, particularly for multilingual and multimodal contexts.

These findings emphasize the importance of developing evaluation resources for diverse languages. Performance on English benchmarks, while informative, may not fully reflect capabilities across linguistic contexts. The challenges we observe, particularly around visual reasoning and complex problem-solving, may manifest differently or more acutely in languages with limited training data.

6.6. Future Directions

Several promising directions for future work emerge from our findings. First, investigating tool use approaches such as code generation for computational verification (Wang et al., 2024b) could improve performance on difficult problems. Recent work has demonstrated that seamless code integration in LLMs significantly enhances mathematical reasoning capabilities (Wang et al., 2024b), achieving substantial improvements on challenging benchmarks. Examining how these techniques transfer to low-resource languages like Icelandic would provide valuable insights.

Second, expanding the benchmark to include applied mathematics, word problems grounded in Icelandic cultural contexts, and problems requiring extended reasoning would provide additional insights into mathematical reasoning capabilities beyond competitive mathematics. This would better reflect the diverse mathematical challenges encountered in educational and real-world contexts.

Third, systematic error analysis (Li et al., 2024b) examining solution quality, reasoning coherence, and specific failure modes would deepen our understanding of model capabilities. Recent work on error identification and correction in mathematical reasoning has revealed that models struggle with different error types, from computational mistakes to logical flaws, and that understanding these patterns is crucial for developing more robust systems.

Finally, cross-lingual transfer experiments (Ko et al., 2025) comparing model performance on translated versions of the same problems could illuminate whether performance differences stem from language-specific challenges or from the mathematical content itself. Recent approaches using explicit cross-lingual Chain-of-Thought reasoning have shown promise in bridging multilingual mathematical reasoning gaps, suggesting methodologies that could be adapted for Icelandic.

7. Conclusion

We introduced Icelandic Math Eval, the first comprehensive benchmark for evaluating large language models on competitive mathematics problems in Icelandic. Our dataset comprises 1,027 problems from Icelandic mathematics competitions spanning 1984-2025, providing a unique resource for assessing mathematical reasoning capabilities in a low-resource language context.

Through extensive evaluation of three state-of-the-art models, Claude Sonnet 4.5, Gemini 2.5 Pro, and GPT-5, we identified several key findings. First, while models demonstrate substantial mathematical reasoning capabilities in Icelandic (81-93% overall accuracy), significant challenges remain, particularly for problems involving visual reasoning (17.48 percentage point performance drop for image-containing problems) and high difficulty levels (dropping to 43-60% accuracy on the most challenging problems).

Second, our dual evaluation mode methodology reveals that evaluation format substantially affects measured performance, with a 6.65 percentage point gap between with-choices and without-choices modes. This finding has important implications for benchmark design and suggests that multiple-choice evaluation may overestimate genuine mathematical reasoning capabilities.

Third, we observe substantial variation across

mathematical domains, with algebra achieving 93.42% accuracy compared to 80.06% for combinatorics. This domain-specific variation, combined with multimodal reasoning challenges, highlights specific areas where current models require improvement.

Our work contributes to the growing body of multilingual evaluation resources and demonstrates the importance of developing benchmarks for diverse languages. The performance patterns we observe, particularly around visual reasoning, difficulty scaling, and evaluation methodology, provide insights relevant beyond Icelandic to multilingual AI development more broadly.

We release our complete dataset, evaluation framework, and detailed results to facilitate future research in multilingual mathematical reasoning. As LLMs continue to advance and expand their language coverage, resources like Icelandic Math Eval help ensure that progress is measured across diverse linguistic communities, promoting more equitable and comprehensive AI capabilities assessment.

8. Limitations

Several limitations of our work warrant consideration and suggest caution in generalizing our findings.

Evaluation methodology: Our LLM-as-judge evaluation approach (Li et al., 2024a), while flexible and increasingly standard in the field, introduces potential sources of error. In particular, using GPT-5 as the primary judge while also evaluating GPT-5 as a test-taker creates a self-assessment scenario that could introduce bias. To quantify this risk, we validated judgments with two independent models (Gemini 3 Flash and Claude Sonnet 4.6), obtaining Fleiss' $\kappa = 0.944$ across all three judges (see Section 4.3). This near-perfect agreement indicates that the primary judge's verdicts are not inflated by self-preferencing. Nonetheless, all three judges are LLMs and may share systematic blind spots; validation against human expert evaluation, particularly for problems requiring nuanced mathematical reasoning, would further strengthen confidence in these results.

Dataset scope: Our dataset, while comprehensive for Icelandic competitive mathematics, represents a specific distribution of problem types and difficulties. The problems are drawn exclusively from mathematics competitions, which may not fully represent the broader space of mathematical reasoning required in educational or real-world contexts. Additionally, the concentration of problems in mid-range difficulty levels (Levels 2-4) means our insights about extreme difficulty levels are based on smaller sample sizes.

Model evaluation settings: We evaluated models using their default configurations and standard prompting approaches. We did not explore extensive prompt engineering, few-shot learning, tool use (e.g., code execution for computational verification), or chain-of-thought variations that might improve performance. Our results therefore reflect baseline capabilities rather than optimized performance, which may underestimate the potential of these models with careful tuning.

Temporal considerations: The problems in our dataset span four decades (1984-2025), during which mathematical education and competition problem design may have evolved. We do not control for potential temporal trends in problem characteristics. Additionally, more recent problems may have appeared in model training data, though the use of Icelandic reduces this concern compared to widely-distributed English benchmarks. Furthermore, commercial LLMs accessed via APIs are not guaranteed to remain stable over time; model updates or deprecation may affect the reproducibility of our specific numerical results, though the benchmark itself and evaluation methodology remain fully reusable.

Language-specific factors: While our work contributes to multilingual evaluation, we examine only one language (Icelandic). The generalizability of our findings to other low-resource languages, particularly those with different linguistic typology or cultural contexts, remains an open question.

9. Ethics Statement

This dataset is collected from publicly available competition problems with appropriate permissions. We ensure that no personally identifiable information is included in the dataset. The dataset is released under a permissive license to facilitate research while respecting the intellectual property of the original problem authors.

10. Data and Code Availability

All data and code are available at [redacted for the sake of anonymity].

11. Acknowledgements

We would like to thank Hallgrímur Haraldsson for useful discussions and valuable input preparing the dataset in the preparation of this work.

12. Bibliographical References

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2025. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *Scientific Data*, 12(1):1392.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. 2024. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *CoRR*, abs/2501.02448.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024b. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11316–11360, Bangkok, Thailand. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025. Mmath: A multilingual benchmark for mathematical reasoning. *arXiv preprint arXiv:2505.19126*.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for Icelandic 2019-2023. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France. European Language Resources Association.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Vésteinn Snæbjarnarson. 2021. Automated methods for question-answering in Icelandic. Master's thesis, University of Iceland.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Natural questions in Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4488–4496, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. MM-MATH: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375, Miami, Florida, USA. Association for Computational Linguistics.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pages 95095–95169. Curran Associates, Inc.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024b. Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pages 169–186, Cham. Springer Nature Switzerland.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust llm evaluators. *CoRR*, abs/2405.11966.

13. Language Resource References

A. Prompt Templates

We provide the complete prompt templates used in our evaluation to ensure reproducibility and facilitate future research. This work was done as part of work supported by the European Commission under grant agreement no. 101135671.

A.1. Model Generation Prompts

A.1.1. With Choices Mode - Multiple Choice Problems

System Prompt (in Icelandic):

```
Þú ert sérfræðingur í stærðfræði. Þér
↪  verður gefið
stærðfræðiverkefni með nokkrum
↪  svarmöguleikum.
```

```
Lestu verkefnið vandlega og veldu rétta
↪  svarið
(A, B, C eða D).
Svaraðu eingöngu með stafnum sem táknar
↪  rétta
svarið (A, B, C, eða D).
```

User Prompt Template:

```
{problem_text}

Svarmöguleikar:
A) {choice_a}
B) {choice_b}
C) {choice_c}
D) {choice_d}

Svaraðu eingöngu með stafnum (A, B, C
↪  eða D):
```

A.1.2. Without Choices Mode

System Prompt (in Icelandic):

```
Þú ert sérfræðingur í stærðfræði. Þér
↪  verður gefið
stærðfræðiverkefni.
Lestu verkefnið vandlega og skilaðu
↪  rétta svarinu.
```

User Prompt Template (Multiple Choice):

```
{problem_text}

Hvert er svarið?
```

User Prompt Template (Numeric):

```
{problem_text}

Svaraðu með tölu:
```

A.2. Judge LLM Prompt

We use GPT-5 as our judge model with structured output to evaluate answer correctness.

System Prompt (in English):

```
You are an expert mathematics evaluator.
↪  Your task is to
determine whether a given answer to a
↪  mathematical problem
is correct.

You will be provided with:
1. A problem statement (in Icelandic)
2. The correct answer
3. The answer provided by an LLM

Your job is to evaluate whether the
↪  LLM's answer matches
the correct answer. Consider the
↪  following:
- For multiple choice questions, the
↪  LLM's answer is
  correct if it either:
  * Matches the correct letter (A, B, C,
  ↪  or D), OR
  * Matches the actual value/content of
  ↪  the correct choice
  * Be flexible with formatting (e.g.,
  ↪  '2000' matches
    '2000 kr.', '$2000$', or similar
    ↪  variations)
- For numeric answers, extract the final
↪  numerical answer
  from the LLM's response and compare it
  ↪  to the correct
  answer
- The LLM may provide reasoning or
↪  explanations - focus
  on the final answer
- Minor formatting differences are
↪  acceptable if the
  mathematical content is correct
- Be objective and fair in your
↪  evaluation

Note: The problem statements and answers
↪  are in Icelandic,
but you should evaluate them objectively
↪  based on
mathematical correctness.
```

User Prompt Template (Multiple Choice):

```
**Problem Statement:**
{problem_text}

**Multiple Choice Options:**
A) {choice_a}
B) {choice_b}
C) {choice_c}
D) {choice_d}

**Correct Answer Letter:**
↪  {correct_answer}

**LLM's Response:**
{llm_response}

**Extracted Answer:** {extracted_answer}

**Answer Type:** {answer_type}

Please evaluate whether the LLM's answer
↪  is correct.
The answer is correct if it matches
↪  either:
1. The correct letter ({correct_answer}),
↪  OR
2. The actual value of the correct
↪  choice (accept
   formatting variations)
```

User Prompt Template (Numeric):

```
**Problem Statement:**
{problem_text}

**Correct Answer:** {correct_answer}
```

```
**LLM's Response:**
{llm_response}

**Extracted Answer:** {extracted_answer}

**Answer Type:** {answer_type}

Please evaluate whether the LLM's answer
↪  is correct.
Consider both the extracted answer and
↪  the full response
context.
```

The judge model uses structured output (via Pydantic schema) to return a JSON object containing:

```
{
  "is_correct": boolean,
  "explanation": string
}
```

B. Dataset Examples

We provide three representative examples from our dataset to illustrate the range of problems and difficulty levels. All problems are presented in their original Icelandic with English translations.

B.1. Example 1: Algebra (Level 3 - Easy)

Problem (Icelandic): Rögnvaldur og vinir hans fjórir eiga að að meðaltali 220 krónur, en Rögnvaldur sjálfur á 380 kr. Hve mikið eiga vinirnir fjórir að meðaltali?

English Translation: Rögnvaldur and his four friends have on average 220 krónur, but Rögnvaldur himself has 380 kr. How much do the four friends have on average?

Multiple Choice Options:

- A) 160 krónur
- B) 180 krónur
- C) 220 krónur
- D) 300 krónur

Correct Answer: B
Problem Type: Algebra
Source: Stærðfræðikeppni framhaldsskólanema 2000-2001 - neðra stig (National High School Mathematics Competition 2000-2001 - Lower Level)

Solution Approach: This is a straightforward average problem. If 5 people have an average of 220 kr, their total is $5 \times 220 = 1100$ kr. Rögnvaldur has 380 kr, so the four friends together have $1100 - 380 = 720$ kr. Therefore, their average is $720/4 = 180$ kr.

B.2. Example 2: Geometry (Level 4 - Medium)

Problem (Icelandic): Á myndinni hér að neðan er $ABCD$ ferningur og P er punktur á hringnum, CB er miðstrengur, $CP = 7$ og $PB = 11$. Hvert er flatarmál ferningsins?

English Translation: In the figure below, $ABCD$ is a square and P is a point on the circle, CB is a diameter, $CP = 7$ and $PB = 11$. What is the area of the square?

Multiple Choice Options:

- A) 144
- B) 169
- C) 170
- D) 180

Correct Answer: C
Problem Type: Geometry
Has Image: Yes (geometric diagram showing square with inscribed circle)
Source: Stærðfræðikeppni framhaldsskólanema 2000-2001 - neðra stig

Solution Approach: This problem combines circle geometry with the Pythagorean theorem. Since CB is a diameter and P is on the circle, angle CPB is a right angle (Thales' theorem). Using the Pythagorean theorem: $CB^2 = CP^2 + PB^2 = 7^2 + 11^2 = 49 + 121 = 170$. Since CB is both a diameter and a side of the square, the area of the square is $CB^2 = 170$.

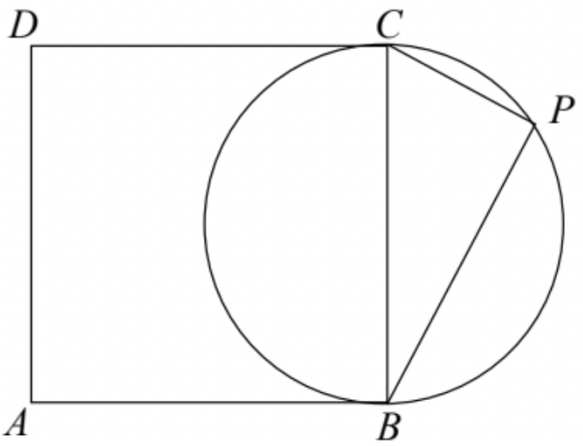

Figure 5: The geometry problem described in Problem B.2

B.3. Example 3: Number Theory (Level 8 - Hard)

Problem (Icelandic): Bjarni skrifaði niður allar jákvæðar heiltölur sem innihalda allt að sjö tölustafi

og þá eingöngu 0 og 1. Hversu oft skrifaði Bjarni töluna 1?

English Translation: Bjarni wrote down all positive integers that contain up to seven digits and only use the digits 0 and 1. How many times did Bjarni write the digit 1?

Multiple Choice Options:

- A) 127
- B) 254
- C) 381
- D) 508

Correct Answer: B

Problem Type: Number Theory (Talnafræði)

Source: Stærðfræðikeppni framhaldsskólanema 2006-2007 - efra stig (National High School Mathematics Competition 2006-2007 - Upper Level)

Solution Approach: This combinatorial number theory problem requires systematic counting. For n-digit numbers using only 0 and 1 (with the first digit being 1), there are 2^{n-1} such numbers. Each position (except the first) has equal probability of being 0 or 1. For 1-digit: 1 number (1), contributing 1 occurrence of digit 1. For 2-digit: 2 numbers (10, 11), contributing 1 (first position) + 1 (second position) = 2 occurrences. Continuing this pattern through 7-digit numbers and summing: the total count follows the formula $\sum_{k=1}^{7} k \cdot 2^{k-1} = 254$.