English-Czech Output Bias in LLMs: A Geometry-Based Case Study

Michaela Tichá, Jiří Přibyl and Magdalena Krátká
University of J. E. Purkyně in Ústí nad Labem, Czech Republic

michaela.ticha@ujep.cz
jiri.pribyl@ujep.cz
magdalena.kratka@ujep.cz

Abstract: The rapid integration of large language models (LLMs) into educational, professional, and public discourse has prompted increasing scrutiny of their multilingual capabilities. While English dominates as a testing and training language, understanding LLM performance in less-resourced languages—such as Czech—is critical for equitable AI deployment. This study investigates a subtle but systematic bias in LLM behaviour: the relative verbosity of their responses in Czech versus English within the domain of elementary geometry. We compiled a dataset of 48 paired mathematical prompts, posed in both Czech and English to six prominent LLMs (ChatGPT, Claude, Gemini, Mistral Large, Copilot Quick-Nuance, and Copilot Deep-Thinker), yielding 576 total responses. Each model was accessed in a controlled language-specific context to ensure fair comparison. Using surface-level metrics—word count and character count—we observed a consistent pattern: English responses were significantly longer than Czech ones across all models. Statistical analysis confirmed the robustness of these differences, with medium to large effect sizes (Cohen's *d*) in both metrics. Notably, even morphologically richer Czech did not yield longer outputs in character count, contradicting initial assumptions. Beyond confirming a consistent verbosity gap, our analysis employed rigorous statistical testing, including paired t-tests and Wilcoxon signed-rank tests, as well as effect size estimation to quantify the magnitude of the disparity. We interpret these findings in the context of known architectural and training imbalances in LLM development—particularly differences in how text is segmented and processed, alongside the relative abundance of English-language data. While stylistic conventions and user context may also influence response length, our results consistently indicate that LLMs, even those marketed as multilingual, tend to produce more verbose output in English. This raises concerns about potential discrepancies in explanation quality across languages, which may have implications for fairness and pedagogical effectiveness in multilingual educational settings. The study lays the groundwork for follow-up research that will move beyond surface metrics toward semantic content analysis of mathematical reasoning across languages. Future work will assess whether English verbosity corresponds to greater mathematical depth, or if Czech responses deliver equivalent content more concisely. This line of inquiry is vital for ensuring fairness, clarity, and effectiveness in multilingual AI deployment—especially in contexts such as mathematics education, where explanation quality directly impacts learning outcomes.

Keywords: Language model evaluation, Multilingual text generation, Response length analysis, Educational AI prompts, Czech-English comparison

1. Introduction and Theoretical Background

The integration of large language models (LLMs) into multilingual education has revealed significant disparities in how these models' reason, generate, and elaborate responses across languages.

1.1 Motivation

This study was motivated by an unexpected finding from our previous work Krátká, Přibyl and Tichá (2025), where Czech-language prompts in lower-secondary geometry tasks led most language models to give more detailed responses than identical prompts in English. While that study focused on geometric features rather than response length, the consistent pattern of greater detail in Czech contradicted expectations, given the dominance of English in model training data. This raised a broader question: do language models systematically generate longer or richer outputs in Czech than in English, even with semantically equivalent prompts? To investigate, we conducted a follow-up study comparing response length in Czech and English across broader mathematical contexts.

1.2 Language Bias and Encoding Inequity

Despite their apparent multilingual capabilities, most large language models (LLMs) are primarily trained on English data. Quantitative studies confirm that performance in lower-resourced languages—such as Czech—is consistently lower than in English, especially in tasks requiring detailed explanations or structured reasoning (Ahuja et al., 2023; Gupta et al., 2025; Sallam et al., 2024, Lai, et al., 2023). A major factor contributing to this disparity lies in how text is internally represented across languages. Petrov et al. (2023) demonstrated that the length of internal sequences for semantically equivalent content can vary by up to 15× between languages, which affects output verbosity, response time, and even usage costs.

Beyond technical performance metrics, sociolinguistic and ethical perspectives highlight that language variation carries identity and power implications (cf. Bender & Friedman, 2018; Blodgett et al., 2020). These frameworks underscore that multilingual fairness in AI is not only a question of accuracy, but of equitable representation and respect for linguistic diversity—a dimension increasingly emphasized in AI ethics discourse.

1.3 Mathematical Reasoning Across Languages

Mathematics, with its structured logic and abstract concepts, is a valuable domain for testing multilingual model robustness. Several recent multilingual benchmarks have been introduced to evaluate these capabilities. The *MMATH* dataset (Luo et al., 2025) spans 10 languages (unfortunately not including Czech) and includes 374 high-quality math problems. Findings revealed not only performance disparities but also issues with language consistency—where models often "think in English" even when prompted in another language.

Similarly, the *PolyMath* benchmark (Wang et al., 2025), which covers 18 languages (not including Czech) and multiple difficulty levels, showed that the "thinking length" (stepwise reasoning verbosity) and semantic precision varied significantly across languages. In both cases, English outputs were consistently more elaborated, reinforcing the idea that verbosity and completeness are influenced by language choice.

1.4 Multilingual Benchmarks and Performance Stratification

The necessity of robust multilingual evaluation frameworks is clear. Benchmarks such as MEGA (Ahuja et al., 2023), MMATH (Luo et al., 2025), and PolyMath (Wang et al., 2025) allow for structured comparison of LLM reasoning across languages and domains. These studies show that model performance correlates strongly with the presence of each language in pretraining data (Li et al., 2024) and that even large models exhibit persistent accuracy and fluency gaps in low-resource languages.

In long-context tasks, Kim et al. (2025) showed that English is not always the best-performing language; however, performance tends to degrade more rapidly with increasing input length in underrepresented languages. Differences in how textual data is broken down and processed across languages further compound this effect (Petrov et al., 2023), reinforcing the need for multilingually fair model design.

2. Methodology

This study investigated whether large language models (LLMs) systematically produce longer or more detailed responses in English than in Czech, using a controlled set of prompts in the domain of elementary geometry. Our methodological design focused on creating paired, semantically equivalent prompts in both languages, eliciting responses from a diverse set of models, and analysing their surface-level output in a transparent and replicable manner.

2.1 Prompt Design and Translation

We constructed a set of 48 geometry-related prompts originally in Czech, each focusing on properties, comparisons, or definitions of basic geometric shapes and relations. Each prompt was carefully translated into English using a combination of ChatGPT and DeepL, and all translations were subsequently reviewed for consistency and semantic equivalence.

The prompt set was designed to simulate natural, school-level mathematical questions that students might reasonably pose when exploring geometry independently or with AI support. All prompts focused on fundamental geometric reasoning rather than technical terminology. This ensured broad comprehensibility for both human readers and AI models, and encouraged the generation of moderately elaborated answers. Prompts were phrased neutrally and consistently to avoid instructional bias or regional assumptions.

Table 1 shows a sample of five representative prompts in both languages:

Table 1: Sample prompt pairs in Czech and English

Czech Prompt	English Prompt
Co mají společného čtverec a obdélník? A v čem se liší?	What do a square and a rectangle have in common? And how are they different?
Jaké vlastnosti má kosočtverec?	What are the properties of a rhombus?

Czech Prompt	English Prompt
Vysvětli mi, čím se liší pravoúhlý a rovnoramenný trojúhelník a co mají společného.	Can you explain the difference between a right-angled triangle and an isosceles triangle, and what they have in common?
Proč musí být součet vnitřních úhlů v trojúhelníku vždycky 180 stupňů?	Why do the angles inside a triangle always add up to 180 degrees?
Co znamená, že je těleso pravidelný mnohostěn? Jak vypadá?	What does it mean when a solid shape is called a regular polyhedron? What does it look like?

Each prompt was phrased clearly, without model-specific instructions, to minimize stylistic bias. Full transcripts of the AI model responses (all 576 interactions) as well as the complete list of 48 prompts in both Czech and English are available from the authors upon request.

2.2 Model Selection and Testing Protocol

We tested six large language model configurations accessible through public-facing web platforms as of June 12, 2025. The models included: ChatGPT (free tier), Claude 3.0 Sonnet, Gemini 1.5 Pro, Mistral Large (via the Le Chat interface), and two conversational variants of Microsoft Copilot: Quick-Nuance and Deep-Thinker. All interactions were conducted via each model's standard web interface without API access, and no system prompts or internal settings were modified beyond the input language and IP geolocation.

The two Copilot variants represent contrasting model personas within the same Microsoft-hosted environment: Quick-Nuance prioritizes fluency and brevity, while Deep-Thinker produces slower but more elaborate responses. Mistral Large, accessed via Le Chat, is an open-weight European model noted for its multilingual capacity. Claude 3.0 Sonnet and Gemini 1.5 Pro are proprietary models with multilingual support and high performance across tasks. ChatGPT remains one of the most widely used freely accessible models, though its behaviour in less-resourced languages like Czech may vary.

All 48 prompts were submitted to each model in both Czech and English, resulting in a total of 576 responses. Interactions were conducted manually between 8:30 and 11:30 CEST on June 12, 2025. Each prompt was entered in a fresh session or chat window to avoid context carryover. Czech prompts were submitted from a Czech IP address (Ústí nad Labem), and English prompts via a VPN located in Florida, USA. The system language was not manually changed, but the use of region-specific IP addresses ensured that models defaulted to the appropriate language environment for each set of prompts.

In two isolated cases, the Czech version of Gemini responded with a refusal message unrelated to the prompt content. Specifically, for the prompts *"Co mají společného kvádr a hranol?"* and *"Jaký je rozdíl mezi čtyřúhelníkem a deltoidem?"*, the model returned generic avoidance responses (e.g., "This conversation isn't my cup of tea"). In both instances, the prompt was immediately resubmitted in a new session without any modification, and the second response was appropriate and included in the dataset. No other model or language combination exhibited such refusals.

2.3 Response Length Measurement

We evaluated each response using two surface-level metrics: (1) the number of words (tokens) and (2) the raw character count. Word counts were computed using a custom tokenization heuristic in Python, which treated each contiguous alphanumeric sequence, standalone numbers, and common mathematical symbols (e.g., π, ±, √) as individual tokens. LaTeX-style expressions enclosed in dollar signs (e.g., *$S = \pi r^2$*) were counted as a single unit. Emojis, emphasis markers (e.g., asterisks, bullets, arrows), and similar stylistic elements were excluded from the token count. This approach ensured consistent and content-focused word counts across languages and models.

Raw character counts were computed using Python's built-in *len()* function and included all characters, such as spaces, punctuation, LaTeX syntax, markdown symbols, emojis, and whitespace. Cleaned character counts excluded stylistic or non-semantic elements (e.g., emojis, visual formatting markers), using a combination of regular expressions and emoji filtering. This dual metric design allowed for both intra-model comparison between languages.

We chose these surface-level metrics—word count and character count—because they provide language-agnostic indicators of verbosity without requiring deeper semantic annotation. While they do not directly

measure information content, they serve as a robust proxy for response elaboration, especially in controlled, paired-prompt settings. In follow-up work, we aim to investigate whether the observed differences in verbosity also correspond to differences in mathematical content (e.g., properties, definitions, relations) versus peripheral narrative or explanatory elements. This will help determine whether Czech responses are merely shorter or also semantically less complete.

2.4 Statistical Analysis

To evaluate whether response length differed systematically between English and Czech outputs, we compared each model's responses across 48 matched prompts. For every model, we computed the difference in response length per prompt (English minus Czech) using two surface metrics: word count and character count. This resulted in 12 paired distributions (6 models × 2 metrics).

Based on general linguistic patterns, we expected English responses to contain more words but potentially fewer characters than their Czech counterparts, as English tends to rely on shorter lexical units and more auxiliary constructions. As such, a significant difference in word count alone would not necessarily indicate truly longer responses. However, if statistically significant differences also emerged in character count, this would more reliably signal that English outputs are indeed more verbose in absolute terms.

We began with exploratory box plot visualizations, which showed that most distributions were approximately symmetric and centred above zero. However, some models—notably Copilot Deep-Thinker and Gemini—exhibited wider spread and visible outliers, particularly in word count differences.

To formally assess the assumption of normality, we applied the Shapiro–Wilk test to each of the 12 paired distributions. In 10 out of 12 cases, the null hypothesis of normality was not rejected at the $\alpha = 0.05$ level. This included both metrics for Gemini ($p = 0.09$ for word count; $p = 0.12$ for character count), where distributions showed minor deviations but no statistically significant departure from normality. Based on these results and the roughly symmetric shapes observed in the plots, we proceeded with a two-tailed paired t-test for these 10 cases.

For Copilot Deep-Thinker, normality was rejected for both metrics ($p < 0.05$). Consequently, we used the Wilcoxon signed-rank test, a non-parametric alternative suitable for paired data with non-normal distributions.

In addition to significance testing, we computed Cohen's d for each difference distribution as a standardized measure of effect size. This statistic quantifies the magnitude of the observed difference in units of pooled standard deviation, facilitating interpretation across models and metrics beyond mere statistical significance.

All statistical computations were performed in Python using the scipy.stats and statsmodels libraries. We did not apply corrections for multiple comparisons, as each test corresponded to a distinct model–metric pair and served to confirm a consistent directional trend. Full test statistics, descriptive summaries, and p-values are reported in Section 3.

3. Results

This section presents the results of the quantitative analysis comparing Czech and English outputs across six large language models.

3.1 Response Length Differences Across Languages

We began by visualizing the distribution of length differences between English and Czech responses across all six models. Figures 1 and 2 display box plots of per-prompt differences (English – Czech), measured in word count and character count, respectively.

Figure 1 shows that all models produced longer responses in English than in Czech, on average. Each box plot summarizes the distribution of word count differences for a given model. Positive values indicate that the English version of the response contained more words than the Czech version. While the direction of the difference was consistent across all models, the magnitude and variability differed. In particular, Copilot Deep-Thinker and Gemini exhibited wider interquartile ranges and a greater number of outliers, suggesting higher variability in verbosity across prompts.

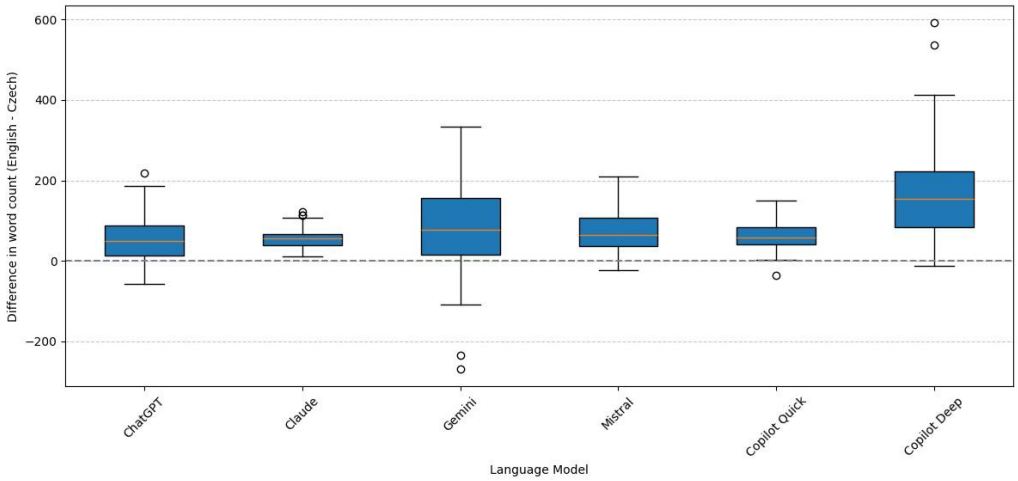

Figure 1: Distribution of word count differences (English – Czech) across six language models

Figure 2 presents the character count differences between English and Czech responses for each model. Once again, all distributions are centred above zero, indicating that English outputs were consistently longer. Copilot Deep-Thinker produced the most pronounced character-level differences, both in median length and in the spread of values. Several outliers, especially for Gemini and Copilot Deep-Thinker, suggest occasional extreme discrepancies between language versions.

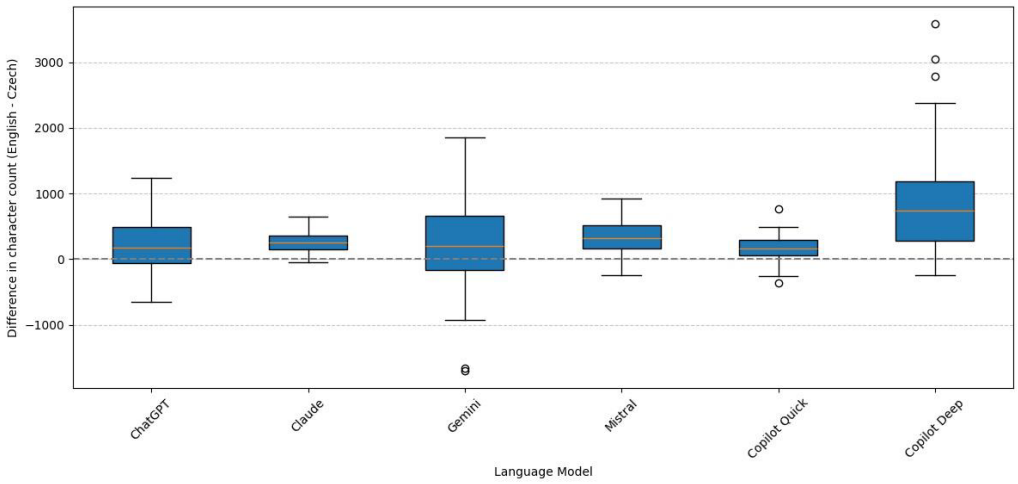

Figure 2: Distribution of character count differences (English – Czech) across six language models

These consistent positive differences suggest a systematic tendency among large language models to produce more verbose responses in English, regardless of prompt content or model type.

3.2 Tests of Normality

We conducted Shapiro–Wilk tests on the distributions of word count and character count differences for each model to assess the assumption of normality. The results are summarized in Table 2.

Table 2: Shapiro–Wilk test for normality of English–Czech response length differences across six large language models, evaluated separately for word count and character count

Model	Word Count (p)	Character Count (p)
ChatGPT	0.17085	0.98701
Claude	0.13835	0.45978
Gemini	0.09457	0.12328
Mistral	0.31635	0.63737

Model	Word Count (p)	Character Count (p)
Copilot Quick	0.89925	0.67320
Copilot Deep	0.00077	0.00041

For Copilot Deep-Thinker, the null hypothesis of normality was clearly rejected for both word and character count differences ($p < 0.001$), justifying the use of non-parametric methods in subsequent analyses. For Gemini, the p-values were relatively close to the 0.05 threshold (0.095 for word count; 0.123 for character count), indicating mild deviations from normality. However, visual inspection of the corresponding box plots revealed approximately symmetric distributions, and thus we retained the use of parametric tests for this model.

3.3 Inferential Results: Statistical Significance and Effect Sizes

We evaluated the statistical significance and magnitude of response length differences (English – Czech) for each model using both word count and character count as dependent measures.

3.3.1 Statistical test results

Across all six models, English responses were significantly longer than Czech responses. Paired t-tests (used for all models except Copilot Deep-Thinker) yielded highly significant results for both word and character counts (see Table 2). For Copilot Deep-Thinker, where the assumption of normality was violated (see Section 3.2), we applied the Wilcoxon signed-rank test (see Table 3); the results remained strongly significant.

Table 2: Paired t-test p-values (all models except Copilot Deep-Thinker)

Model	Word Count (p)	Character Count (p)
ChatGPT	3.28e-07	8.57e-04
Claude	8.76e-19	2.50e-14
Gemini	3.96e-05	*3.27e-02*
Mistral	6.13e-14	5.28e-12
Copilot Quick	2.25e-15	2.23e-07

Table 3: Wilcoxon test p-values for Copilot Deep-Thinker

Model	Word Count (p)	Character Count (p)
Copilot Deep	4.97e-14	2.65e-11

3.3.2 Effect size and model comparisons

To complement the significance tests, we computed Cohen's d as a standardized measure of effect size. This quantifies the magnitude of the English–Czech length difference, helping to distinguish between statistically significant but trivial effects and those that are substantial and consistent across prompts.

Table 4 summarizes the mean differences, standard deviations, and effect sizes for both word and character counts. Figures 3 and 4 visualize these effect sizes using bar charts, with dashed lines indicating thresholds for small (0.2), medium (0.5), and large (0.8) effects.

Table 4: Effect Sizes (Cohen's d) for Word and Character Count Differences

Model	Cohen's d (Words)	Mean Diff (Words)	SD (Words)	Cohen's d (Chars)	Mean Diff (Chars)	SD (Chars)
ChatGPT	0.8579	53.71	62.60	0.5141	211.69	411.78
Claude	2.0697	56.85	27.47	1.5593	261.58	167.76
Gemini	0.6547	74.60	113.94	0.3176	218.77	688.83
Mistral	1.5181	73.85	48.65	1.3200	335.48	254.16
Copilot Quick	1.6723	59.92	35.83	0.8738	173.56	198.63
Copilot Deep	1.3333	167.08	125.32	1.0601	874.81	825.21

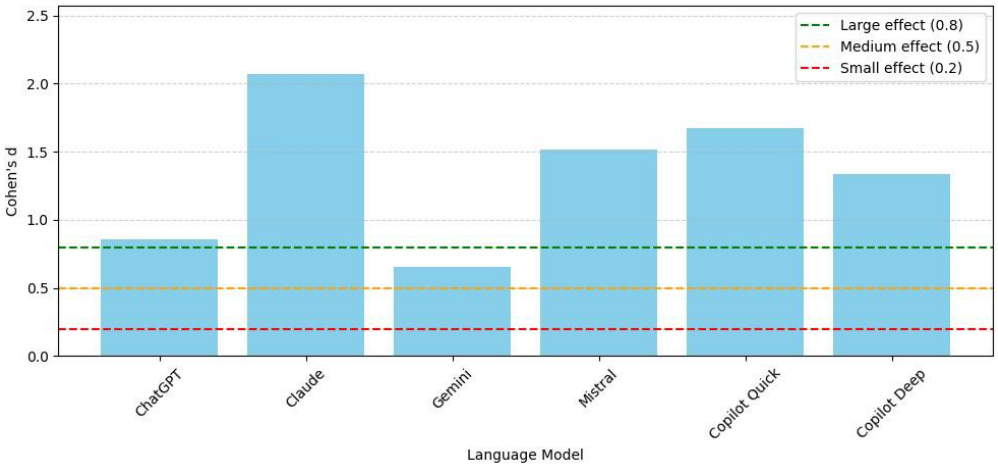

Figure 3: Cohen's d effect sizes for word count differences (English – Czech) across six language models

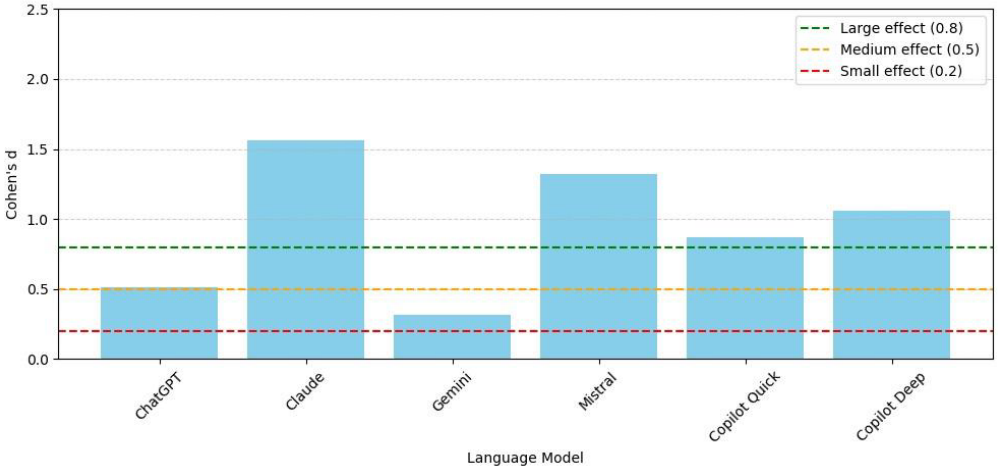

Figure 4: Cohen's d effect sizes for character count differences (English – Czech) across six language models

Across all models, English responses were longer and the effect sizes exceeded the medium threshold (*d > 0.5*) in nearly every case. The most pronounced differences were observed in Claude (*d = 2.07 words; d = 1.56 characters*), followed by Copilot Quick and Mistral, both of which also showed large effects. This indicates not only a reliable pattern, but one of substantial magnitude.

In contrast, Gemini exhibited the weakest effects (*d = 0.65 words; d = 0.32 characters*), suggesting either more balanced generation across languages or a general tendency toward conciseness. Interestingly, Copilot Deep-Thinker showed the largest absolute differences in both metrics (+167 words, +875 characters), but with slightly lower standardization due to higher variance—consistent with its design to produce elaborate, but variable, outputs.

Notably, character-level effects were generally smaller than word-level ones, particularly for ChatGPT and Gemini. This suggests that additional words in English responses may include redundant phrasing or syntactic padding rather than denser informational content, depending on the model.

3.3.3 Interpretation

Taken together, these results suggest that the observed verbosity in English is not merely an incidental artifact of translation choices or prompt formulation. Crucially, the fact that English responses were significantly longer than Czech ones not only in word count but also in character count reinforces the interpretation that the difference reflects a true increase in textual length, rather than just a structural property of the language. While English often uses more—but shorter—words than Czech, a consistent increase in both metrics indicates that responses in English were genuinely more expansive. This pattern was observed across all tested LLMs,

regardless of model architecture or interface type, underscoring a systematic tendency to generate longer and often more elaborated outputs in English, even when provided with semantically equivalent prompts.

3.4 Summary of Statistical Findings

In summary, all six models generated significantly longer responses in English than in Czech across both word and character counts. The strongest effects were observed in Claude and Copilot Quick, while Copilot Deep-Thinker exhibited the greatest variability. The consistency of this pattern across both parametric and non-parametric tests supports the conclusion that LLM verbosity is systematically influenced by language, even under tightly controlled prompt conditions. These findings provide a robust foundation for further semantic and qualitative analyses.

4. Discussion

The results of this study reveal a consistent and statistically robust tendency across large language models (LLMs) to produce longer responses in English than in Czech, even when presented with semantically equivalent prompts in a controlled setting. This pattern held across all six tested models and both surface-level verbosity metrics—word count and character count.

4.1 Alignment with Prior Research

Our findings are largely consistent with existing literature that has documented performance asymmetries in multilingual outputs of LLMs. In particular, several large-scale benchmarking efforts (Ahuja et al., 2023; Wang et al., 2025; Luo et al., 2025) have shown that model responses in English are typically more complete, coherent, and accurate than in lower-resource languages such as Czech. These patterns have also been observed in domain-specific evaluations, including educational (Gupta et al., 2025), medical (Sallam et al., 2024), and long-context reasoning tasks (Kim et al., 2025). Moreover, underlying architectural factors such as tokenization granularity and training data imbalance are recognized contributors to these disparities (Petrov et al., 2023; Li et al., 2024).

Interestingly, our results contrast with the findings of Krátká, Přibyl and Tichá (2025), who reported higher feature richness in Czech-language geometry responses. However, their study focused on qualitative geometric detail rather than holistic completeness or answer length, and applied a different evaluation scheme targeting concept-level granularity. This divergence highlights the importance of metric selection in multilingual LLM research, and suggests that certain linguistic or contextual factors—such as overcompensation or literalism in Czech—may modulate output richness in ways not directly comparable across studies.

Overall, our results reaffirm the prevailing view that English remains the most optimized language for LLM output, particularly in tasks requiring structured, multi-step reasoning.

4.2 Possible Explanations

Several factors may contribute to the observed differences in verbosity. One possibility is that LLMs have been trained on disproportionately larger and more diverse English-language corpora, resulting in more confident, expansive output in English. Additionally, differences in language structure could play a role. While Czech is morphologically richer and can express meaning more compactly, English often requires more lexical units and auxiliary constructions to convey the same ideas. However, this linguistic asymmetry alone cannot account for the consistent English-dominant verbosity, particularly since we also observed longer character counts in English responses—contrary to our initial intuition that Czech might require more characters due to its inflectional nature.

Interface conventions may also contribute. In English-language environments—where many LLMs are trained and optimized—responses may reflect culturally embedded norms of elaboration, instructional tone, and over explanation. This stylistic bias could lead to longer outputs in English, independent of the underlying content. In contrast, responses in Czech may reflect a more concise style, either due to linguistic economy or because such elaborative conventions are less prevalent in the Czech training data or user context.

4.3 Implications

The implications of these findings are twofold. First, they raise questions about fairness and content equivalence in multilingual LLM applications, especially in educational contexts where completeness and clarity of explanations are critical. Shorter responses in Czech may lead to reduced informativeness or perceived support, even if the semantic content is technically sufficient.

Second, the results highlight the importance of cross-lingual evaluation in LLM development. Models that appear robust in English may perform inconsistently in other languages not only in terms of accuracy but also in surface fluency, elaboration, and tone. These qualitative differences can have downstream effects on user trust, comprehension, and accessibility.

4.4 Pedagogical Consequences

From an educational perspective, the observed brevity of Czech responses may have tangible consequences for learners. Shorter explanations can reduce perceived guidance, limit opportunities for reflection, and weaken user trust—especially in self-paced learning settings where AI feedback replaces human scaffolding. Conversely, English users may benefit from richer elaboration that reinforces conceptual understanding. This asymmetry suggests that LLM-mediated learning experiences could inadvertently reproduce linguistic inequities, privileging students who interact in English. Addressing this issue is therefore not only a matter of linguistic fairness, but also of educational quality and inclusion.

4.5 Limitations

4.5.1 Surface metrics vs. semantic content

This study evaluated only surface-level verbosity metrics—specifically, word count and character count—which do not directly reflect the semantic richness or mathematical depth of responses. While these metrics are useful proxies for response elaboration, they cannot distinguish between substantive mathematical reasoning and stylistic or redundant padding.

4.5.2 Domain and prompt scope

The prompt set was limited to 48 questions in the domain of elementary geometry. Although the questions were designed to be pedagogically relevant and broadly representative of school-level reasoning, this domain specificity may limit generalizability to other subject areas. Similarly, other question types might elicit different language behaviours.

4.5.3 Model behaviour and language detection

Although prompt translations were reviewed for accuracy and the prompts were submitted under tightly controlled conditions (including geolocation and browser language settings), we cannot fully exclude the possibility that internal language detection mechanisms or user-context heuristics affected model behaviour. Some platforms may infer user language preferences based on prior usage or hidden system prompts, potentially influencing output verbosity in ways we could not observe or control.

4.5.4 Temporal validity

All data were collected on June 12, 2025. As LLMs are continuously updated and retrained—often without public versioning or changelogs—model behaviour may have changed since data collection. Results should therefore be interpreted as a snapshot of model performance at that time.

4.5.5 Context and interaction effects

Our design focused on isolated, single-turn prompts submitted in fresh sessions to minimize prior context effects. While this setup increases experimental control, it does not fully reflect real-world usage, where models often operate in multi-turn conversations. Prompt position, interaction history, or cumulative dialogue context may modulate verbosity differently in Czech and English, especially in dynamic, tutoring-style exchanges.

4.6 Future Firections

Building on the current findings, future work will move beyond surface-level metrics to examine the semantic content of LLM-generated responses across languages. Specifically, we plan to conduct a phenomenon-level content analysis in which each response is annotated for the presence of mathematical reasoning elements, such as definitions, properties, examples, comparisons, and deductive steps. This will allow us to determine whether the observed verbosity differences correspond to actual content disparities, or whether Czech responses merely express the same ideas more concisely.

4.7 Design Recommendations and Mitigation Strategies

Future research should also explore strategies to mitigate these cross-lingual disparities. Potential approaches include balanced fine-tuning with pedagogically relevant Czech data, adaptive verbosity controls that equalize

elaboration across languages, and transparent reporting of multilingual performance. Collaborations between AI developers, linguists, and educators could inform guidelines ensuring that LLMs deliver equally supportive and trustworthy responses regardless of language.

5. Conclusion

This study investigated whether large language models produce longer responses in English than in Czech when given semantically equivalent mathematical prompts. Analysing responses from six widely used LLMs, we found a consistent and statistically significant pattern: English responses were longer in both word and character count across all models, with medium to large effect sizes.

These findings suggest that LLM output is influenced not only by prompt content but also by language-dependent behaviours likely rooted in training data distributions, tokenization practices, and optimization objectives. While verbosity alone does not equate to quality, the observed differences raise important questions about fairness, completeness, and content equity in multilingual LLM deployment.

Addressing these issues requires a deeper understanding of how semantic content is distributed across languages in AI outputs. Our future work will focus on content-based and rhetorical analysis to uncover whether verbosity differences entail differences in mathematical depth. Ultimately, ensuring linguistic parity in AI-generated content is essential for building models that serve diverse users equitably—regardless of the language they speak.

Ethics declaration: This study did not involve human participants or personal data and therefore did not require ethical approval.

AI declaration: This study examines the output behaviour of large language models (LLMs), which also served as the primary data sources for the experimental analysis. Responses were collected from six publicly accessible AI systems—ChatGPT, Claude, Gemini, Mistral, and two configurations of Microsoft Copilot—and subsequently analysed by the authors.

During the manuscript preparation, the authors utilized ChatGPT (OpenAI, GPT-4) to support select aspects of English-language phrasing, editorial refinement, and structural organization. All analytical decisions, result interpretation, and formulation of research conclusions were conducted exclusively by the human authors. Any AI-assisted text was carefully reviewed, edited, and validated prior to inclusion.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K. and Sitaram, S. (2023) *MEGA: Multilingual Evaluation of Generative AI*. arXiv:2303.12528. https://doi.org/10.48550/arXiv.2303.12528

Bender, E.M. and Friedman, B. (2018) *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics*, Vol 6, pp 587–604. https://doi.org/10.1162/tacl_a_00041

Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020) *Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Gupta, V., Chowdhury, S.P., Zouhar, V, Rooein, D. and Sachan, M. (2025) *Multilingual Performance Biases of Large Language Models in Education*. arXiv:2504.17720. https://doi.org/10.48550/arXiv.2504.17720

Kim, Y., Russell, J., Karpinska, M. and Iyyer, M. (2025) *One Ruler to Measure Them All: Benchmarking Multilingual Long-Context Language Models*. arXiv:2503.01996. https://doi.org/10.48550/arXiv.2503.01996

Krátká, M., Přibyl, J. & Tichá, M. (2025). "AI in the Classroom: Didactical Misalignments in Geometry Between Czech and Anglo-Saxon Contexts". *To appear at the ECEL 2025 Conference*.

Lai, V.D., Ngo, N.T., Pouran ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T. and Nguyen, T.H. (2023) *ChatGPT Beyond English: Comprehensive Evaluation Across Languages*. arXiv: 2304.05613. https://doi.org/10.48550/arXiv.2304.05613

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N. and Du, M. (2024) *Quantifying Multilingual Performance of Large Language Models Across Languages*. arXiv:2404.11553. https://doi.org/10.48550/arXiv.2404.11553

Luo, W., Zhao, W.X., Sha, J., Wang, S. and Wen, J.-R. (2025) *MMATH: A Multilingual Benchmark for Mathematical Reasoning*. arXiv:2505.19126. https://doi.org/10.48550/arXiv.2505.19126

Petrov, A., La Malfa, E., Torr, P.H.S. and Bibi,A. (2023) "Language Model Tokenizers Introduce Unfairness Between Languages", *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp 36963–36990. Available at: https://dl.acm.org/doi/10.5555/3666122.3667730

Sallam, M., Al-Mahzoum, K., Alshuaib, O., Alhajri, H., Alotaibi, F., Alkhurainej, D., Al-Balwah, M.Y., Barakat, M. and Egger, J. (2024) "Language Discrepancies in the Performance of Generative AI Models: An Examination of Infectious Disease

Queries in English and Arabic", *BMC Infectious Diseases*, Vol 24, 799, pp 1–13. https://doi.org/10.1186/s12879-024-09725-y

Wang, Y., Zhang, P., Tang, J., Wei, H., Yang, B., Wang, R., Sun, C., Sun, F., Zhang, J., Wu, J., Cang, Q., Zhang, Y., Huang, F., Lin, J., Huang, F. and Zhou, J. (2025) *PolyMath: Evaluating Mathematical Reasoning in Multilingual Contexts*. arXiv:2504.18428. https://doi.org/10.48550/arXiv.2504.18428