



UDC 81'322.4:51:004.8

[https://doi.org/10.52058/2786-5274-2026-3\(55\)-660-677](https://doi.org/10.52058/2786-5274-2026-3(55)-660-677)

Sachaniuk-Kavets'ka Natalia Vasylivna, Candidate of technical sciences, Associate professor, Vinnytsia National Technical University, Vinnytsia, <https://orcid.org/0000-0001-6405-1331>.

Nykyporets Svitlana Stepanivna the senior English language lecturer, Vinnytsia National Technical University, Vinnytsia, <https://orcid.org/0000-0002-3546-1734>.

LLM-BASED AUTOMATION FOR TRANSLATING MATHEMATICAL FORMULAE AND SYMBOLS: CHALLENGES AND PERSPECTIVES FOR TECHNICAL COMMUNICATION

Abstract. The growing reliance of students and researchers on large language models for the translation of technical and scientific texts has outpaced critical evaluation of the risks such reliance entails, particularly where formal mathematical notation is concerned. This paper investigates the characteristic failure modes of three state-of-the-art generative language models – GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro – when tasked with translating mathematical discourse from English into Ukrainian across a stratified corpus of 120 text fragments drawn from university-level textbooks, peer-reviewed journal articles, and olympiad problem sets. Employing a dual prompting design that contrasted zero-shot and few-shot conditions, and evaluating outputs against human-produced reference translations using four criteria – symbolic fidelity, logical coherence, terminological conformity with ISO 80000 and IEC 60050, and register consistency – the study identifies and classifies 274 critical errors into four taxonomic categories: syntactic and structural errors, semantic symbol conflation, contextual hallucination, and natural language interference. Quantitative analysis reveals that symbolic fidelity scores lagged behind logical coherence ratings by between four and nine percentage points across all model-condition pairs, confirming that current models handle the expository prose of mathematical argument more reliably than the formal notation it describes. Claude 3.5 Sonnet produced the fewest critical errors overall, whilst structural distortions in nested notation represented the single most prevalent failure type across all three models. The study further identifies a previously underdescribed phenomenon – the context-driven silent correction of formally correct but non-standard expressions toward distributional typicality – and argues that this constitutes a particularly consequential risk in professional and academic translation contexts. The paper concludes by outlining directions for domain-specialised model development, hybrid verification pipelines, and dedicated evaluation frameworks for mathematical translation, whilst acknowledging the study's linguistic and temporal scope as avenues for future investigation.





Keywords: large language models, mathematical translation, symbolic fidelity, hallucination, technical communication.

Сачанюк-Кавецька Наталія Василівна кандидат технічних наук, доцент кафедри вищої математики, Вінницький національний технічний університет, м. Вінниця, <https://orcid.org/0000-0001-6405-1331>.

Никипорець Світлана Степанівна старший викладач кафедри іноземних мов, Вінницький національний технічний університет, м. Вінниця, <https://orcid.org/0000-0002-3546-1734>.

АВТОМАТИЗАЦІЯ ПЕРЕКЛАДУ МАТЕМАТИЧНИХ ФОРМУЛ І СИМВОЛІВ НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ: ВИКЛИКИ ТА ПЕРСПЕКТИВИ ДЛЯ ТЕХНІЧНОЇ КОМУНІКАЦІЇ

Анотація. Зростаюча залежність студентів і дослідників від великих мовних моделей при перекладі технічних та наукових текстів випередила критичне осмислення ризиків, які така залежність несе в собі, – особливо там, де йдеться про формальний математичний запис. У цій статті досліджуються характерні типи помилок трьох сучасних генеративних мовних моделей – GPT-4o, Claude 3.5 Sonnet та Gemini 1.5 Pro – при виконанні завдання перекладу математичного дискурсу з англійської мови на українську на основі стратифікованого корпусу з 120 текстових фрагментів, відібраних із університетських підручників, рецензованих наукових статей та умов задач олімпіад. Застосовуючи двомовний промпт-дизайн, що порівнював режими zero-shot та few-shot, і оцінюючи результати на тлі еталонних перекладів, виконаних людьми, за чотирма критеріями – символічна точність, логічна зв'язність, термінологічна відповідність стандартам ISO 80000 та ІЕС 60050, а також реєстрова узгодженість – дослідження виявляє та класифікує 274 критичні помилки за чотирма таксономічними категоріями: синтаксичні та структурні помилки, семантичне змішування символів, контекстуальні галюцинації та інтерференція природної мови. Кількісний аналіз засвідчує, що показники символічної точності відставали від оцінок логічної зв'язності на від чотирьох до дев'яти відсоткових пунктів у всіх парах «модель – умова», підтверджуючи, що сучасні моделі значно надійніше відтворюють пояснювальну прозу математичної аргументації, ніж формальний символічний апарат, який ця проза описує й інтерпретує. Claude 3.5 Sonnet продемонстрував найменшу загальну кількість критичних помилок, тоді як структурні спотворення у вкладених нотатках виявилися найпоширенішим типом помилок у всіх трьох моделях. Дослідження також виявляє раніше недостатньо описане явище – контекстно зумовлене мовчазне «виправлення» формально коректних, але нестандартних виразів у бік дистрибутивної типовості – і обстоює думку, що саме воно становить особливо серйозний ризик у контексті професійного та академічного перекладу. На



завершення окреслено напрямки розробки вузькоспеціалізованих моделей, гібридних конвеєрів верифікації та спеціалізованих систем оцінювання математичного перекладу, а також визначено лінгвістичні та часові обмеження дослідження як орієнтири для майбутніх праць.

Ключові слова: великі мовні моделі, математичний переклад, символічна точність, галуцинації, технічна комунікація.

English has long served as the principal medium of scientific exchange, consolidating its role as the de facto lingua franca of mathematics, engineering, and the natural sciences over the course of the twentieth century. The practical consequences of this dominance are well documented: researchers and students whose native language is not English face considerable pressure to produce, comprehend, and disseminate technical knowledge in a second tongue, frequently with limited institutional support. In recent years, the rapid proliferation of large language models (LLMs) – most notably ChatGPT, Claude, and Gemini – has introduced a widely accessible means of bridging this linguistic gap. Students, engineers, and academics increasingly rely on these tools not merely for general-purpose translation but for the rendering of dense, symbol-laden technical texts that resist conventional approaches. This shift has proceeded largely without critical scrutiny, driven by the intuitive appeal of fluent, confident-sounding output. Yet mathematical language occupies a peculiar position within the broader landscape of natural language: it is simultaneously a formal symbolic system and a domain-specific register of prose, and the two dimensions impose distinct and sometimes conflicting demands on any translation process. LLMs, trained predominantly on natural language corpora, have demonstrated a well-documented tendency to produce plausible-sounding but factually incorrect outputs - a phenomenon commonly referred to as hallucination – and there is good reason to suspect that this tendency is particularly consequential when the text in question involves precise symbolic notation. An erroneously rendered integral, a misplaced subscript, or a conflated operator can silently alter the meaning of an expression in ways that neither the author nor a casual reader may immediately detect. Despite the scale of LLM adoption in academic and technical settings, systematic investigation of the specific failure modes that arise in the translation of mathematical discourse remains sparse.

The present study aims to address this gap by identifying, classifying, and analysing the characteristic errors that state-of-the-art language models commit when tasked with translating mathematical texts, with a view to informing both users and developers about the current limitations of automated technical translation.

Literature review

The past several years have produced a substantial body of empirical work examining how well large language models handle mathematical content, and the picture that emerges is one of genuine but sharply constrained competence. Frieder et al. (2024), in one of the most thorough evaluations of its kind, tested ChatGPT and GPT-4 against



graduate-level mathematical problems using the newly assembled GHOSTS and miniGHOSTS datasets, which were curated by working research mathematicians rather than derived from existing examination banks. Their principal finding was that while both models can serve effectively as mathematical search engines at the undergraduate level – retrieving known results, identifying relevant theorems, and completing partially specified arguments – neither approaches the standard expected of a graduate student when confronted with problems that demand genuine multi-step deductive reasoning. GPT-4 failed consistently on graduate-level tasks, exhibiting a troubling tendency to generate justifications that, though superficially coherent, contained elementary logical flaws invisible to a non-specialist reader [1]. This finding was corroborated and refined by Collins et al., who introduced [2] the CheckMate platform to study real-time interactions between professional mathematicians and language models during theorem proving. By moving beyond static input-output evaluation to observe dynamic human-LLM collaboration, they were able to identify a recurring failure mode in which models fabricate supporting facts that were not present in the problem statement – a direct manifestation of hallucination in a formally structured domain. Crucially, expert mathematicians reported that erroneous generations were often perceived as superficially plausible, raising serious concerns about the reliability of LLM-generated mathematical assistance in contexts where the user lacks the domain expertise to identify errors independently.

A broader survey of the field by Ahn et al. further confirmed [3] that, whilst progress on benchmark datasets such as GSM8K has been dramatic, performance degrades sharply as tasks become more abstract and require sustained symbolic manipulation, and that existing benchmarks have arguably rewarded surface-level pattern matching rather than the kind of procedural fluency demanded by genuine mathematical reasoning.

Alongside these STEM-specific evaluations, a parallel literature has addressed the broader question of how generative language models compare with dedicated neural machine translation (NMT) systems – such as Google Translate and DeepL – when applied to technical and scientific texts. The conventional wisdom, rooted in a decade of NMT development, held that task-specialised models trained on parallel corpora would inevitably outperform general-purpose language models on translation tasks requiring domain precision. This assumption has been progressively challenged. Hendy et al. (2023), in a large-scale comparison spanning eighteen language pairs and three GPT variants, found [4] that GPT models are highly competitive with state-of-the-art NMT systems for high-resource language pairs, though they exhibit markedly inferior performance on low-resource languages and domain-specific lexis – a limitation directly attributable to the heavily English-dominated composition of LLM training corpora. He et al. (2024), drawing on findings from the WMT22 translation shared task, argued [5] that the qualitative advantage of LLMs over conventional NMT lies not in raw accuracy scores but in their capacity for contextual flexibility: unlike pipeline-based NMT architectures, which perform deterministic source-to-target





mapping, LLMs can integrate metalinguistic instructions, style constraints, and domain preferences directly into the generation process.

A recent study specifically focused on the translation of full scientific articles extended these findings considerably: Kleidermacher and Zou (2025) developed [6] an LLM-based pipeline for translating journal papers across 28 languages whilst preserving JATS XML formatting, reporting an average comprehension-based benchmark accuracy of 95.9%; however, the same user study revealed that approximately one-third of researcher-participants found technical terminology to be systematically overtranslated – that is, discipline-specific English terms were rendered into target-language equivalents when authors preferred them left in English, suggesting that the models lack an adequate register model for scientific prose.

A dimension that has received considerably less attention in the machine translation literature, but that is directly relevant to the present study, concerns the cognitive consequences for non-native readers of encountering mathematical texts in a second language. Roussel, Tricot, and Sweller (2022) conducted a tightly controlled randomised experiment [7] to test the widely held assumption that exposure to academic content through a second language simultaneously promotes both content acquisition and language development – a belief sometimes described as "killing two birds with one stone." Their results were unequivocal: processing subject-matter content in an L2 consistently suppressed content learning without producing commensurate gains in language proficiency, an outcome they interpreted within the framework of cognitive load theory as a consequence of the finite capacity of working memory being diverted to the metalinguistic demands of L2 processing [8] at the expense of subject-matter comprehension. Applied to the context of the present study, this finding has an important practical implication: when non-native-English-speaking students or researchers resort to LLMs to reduce the linguistic burden [9] of technical texts – and in particular to render mathematical notation into their first language – any errors introduced by the translation process are especially likely to go undetected, precisely because the user's cognitive resources are already engaged in managing L2 load. The literature reviewed here therefore converges on a picture in which LLMs occupy a position of considerable practical influence in STEM education and research communication, while simultaneously exhibiting systematic weaknesses in symbolic precision and domain-specific register that the existing evaluation paradigms have only partially characterised.

Methodology

The corpus assembled for this study was drawn from three categories of published technical material: chapters from widely adopted university-level textbooks in calculus, linear algebra, and classical mechanics; excerpts from peer-reviewed journal articles in applied mathematics and theoretical physics sourced from open-access repositories; and problem sets from international mathematics olympiad archives, selected on the grounds that they offer dense, unambiguous symbolic content whose correct interpretation admits no reasonable dispute. In total, the corpus



comprised 120 text fragments, each between 150 and 400 words in length, with every fragment containing a minimum of five distinct mathematical expressions rendered in standard LaTeX notation. Fragments were stratified across three levels of symbolic complexity – elementary (single-variable calculus, basic linear algebra), intermediate (multivariable analysis, differential equations), and advanced (functional analysis, tensor calculus) – so as to permit systematic comparison of model performance across varying degrees of notational density. Source texts were all originally written in English, and the target translation language was Ukrainian, chosen both for its non-Latin script and for the relative scarcity of high-quality Ukrainian-language STEM corpora, conditions which were expected to stress-test the models' technical register and symbol-handling capabilities.

Three state-of-the-art generative language models were selected for evaluation: GPT-4o (OpenAI, May 2024 release), Claude 3.5 Sonnet (Anthropic, June 2024 release), and Gemini 1.5 Pro (Google DeepMind, February 2024 release). Each model was accessed through its official API under default temperature settings to ensure reproducibility and to minimise stochastic variation across runs. Two prompting conditions were employed for each fragment: a zero-shot condition, in which the model received only the source text and a minimal instruction to produce [10] an accurate Ukrainian translation preserving all mathematical notation; and a three-example few-shot condition, in which the instruction was preceded by three human-verified translation pairs drawn from outside the evaluation corpus, illustrating the expected treatment of inline formulae, displayed equations, and technical nomenclature. This two-condition design allowed us to assess whether explicit demonstration of correct symbol-handling behaviour in the prompt materially improves output quality, or whether the limitations observed under zero-shot conditions reflect deeper architectural constraints that surface-level prompt engineering cannot overcome.

Translated outputs were evaluated against human-produced reference translations prepared by two bilingual academics – one a professional mathematician, the other a specialist in technical translation – who worked independently and resolved disagreements through discussion. Four evaluation criteria were applied. First, symbolic fidelity, defined as the proportion of mathematical expressions in the source text reproduced without structural alteration in the target, was assessed by direct token-level comparison of LaTeX source and output notation. Second, logical coherence was rated on a five-point scale by the mathematician evaluator, who judged whether the translated prose preserved the inferential structure of the original argument irrespective of surface linguistic choices. Third, terminological conformity was assessed against the relevant vocabulary sections of ISO 80000 (Quantities and Units) and IEC 60050 (International Electrotechnical Vocabulary), which together constitute the principal international normative standards for scientific and technical terminology. Fourth, register consistency – that is, whether the output maintained a level of formality and disciplinary specificity appropriate to academic scientific writing – was evaluated by the translation specialist using a binary pass/fail judgement supplemented by



qualitative annotation. Inter-rater reliability for the symbolic fidelity and terminological conformity criteria was calculated using Cohen's kappa, and all fragments on which the two evaluators disagreed by more than one scale point were excluded from the quantitative analysis and retained solely for qualitative case study.

Presentation of the main material of the study

A taxonomy of LLM errors in the translation of mathematical discourse

The systematic analysis of model outputs across the three prompting conditions and three models yielded a richly varied body of translational failure. Rather than treating errors as isolated incidents, we found it both analytically productive and pedagogically useful to organise them into four broad categories that reflect distinct mechanisms of breakdown – syntactic, semantic, contextual, and natural-language-driven. What follows is a detailed account of each category, illustrated with representative cases drawn from the evaluation corpus.

Syntactic and structural errors

The most immediately visible class of error concerned the structural integrity of mathematical expressions themselves. Mathematical notation encodes hierarchical relationships – between numerator and denominator, between base and exponent, between operator and operand – that are represented in LaTeX through nested bracket structures and command scoping. All three models, under both prompting conditions, exhibited a recurring tendency to flatten or misparse these hierarchies, particularly when an expression extended over several nested levels or when a single symbol served simultaneously as both an operator and a delimiter.

A representative and instructive example arose from a fragment drawn from a graduate-level real analysis textbook, which contained the following expression for the limit of a Riemann sum:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*) \Delta x_i$$

All three models correctly reproduced the summation symbol and the limit notation in isolation, but each introduced a structural distortion when the expression appeared embedded within a longer displayed equation. GPT-4o rendered the superscript on x_i^* as a standalone asterisk following the closing parenthesis, thereby shifting the asterisk from its syntactically prescribed position as an index to a visually adjacent but logically unconnected character. Gemini 1.5 Pro collapsed the subscript-superscript pair on the summation sign into a single subscript, producing a symbol that superficially resembled the original but which, in a formal context, encodes a meaningfully different operation. Claude 3.5 Sonnet performed best on this particular item under the few-shot condition, reproducing the structure faithfully; however, under zero-shot prompting it omitted the Δx_i term entirely from the translated output without comment, a failure that will be addressed more fully in the next sections. These structural distortions were especially prevalent in expressions involving fractions



within exponents, double integrals with compound limits, and matrix notation in which row and column indices must be simultaneously maintained – configurations in which the depth of nesting appears to exceed the models' reliable parsing horizon.

Semantic ambiguity and symbol conflation

A subtler and in many respects more dangerous class of error arose not from misformatting a symbol but from systematically conflating two distinct symbols that share visual similarity, historical lineage, or overlapping usage across disciplines. Unlike structural errors, which we may detect [11] through visual inspection of the output, semantic conflation errors can survive undetected because the substituted symbol is itself mathematically legitimate – it simply carries a different meaning than the one the author intended.

The most frequently observed instance of this type in our corpus involved the conflation of the Latin lowercase letter v (typically denoting velocity or a general vector component) with the Greek letter ν (nu), which appears in physics as the symbol for kinematic viscosity, oscillation frequency, and photon frequency, among other quantities. Across the 120 fragments, this pair was confused on fourteen separate occasions – eleven of which occurred under the zero-shot condition – with Gemini 1.5 Pro accounting for the largest share of confusions. The practical consequences of such an error in, say, the Navier-Stokes equations, where both v (velocity component) and ν (kinematic viscosity) appear in adjacent terms, would be a translated equation that is internally consistent in appearance but physically nonsensical.

A second pair that generated considerable confusion was the nabla operator ∇ and the Laplacian Δ . Although these two symbols occupy distinct roles in vector calculus – ∇ denoting the gradient, divergence, or curl depending on context, and Δ denoting the scalar Laplacian operator – they are historically related and occasionally used interchangeably in older or less rigorously typeset texts. GPT-4o substituted Δ for ∇ in three fragments drawn from electromagnetism and fluid dynamics, in each case producing a translated equation in which a vector differential operator was replaced by a scalar one, fundamentally altering the dimensionality of the expression. Analogous conflations were observed between Σ (summation) and σ (standard deviation or stress tensor), between $\mathcal{L}$ (Lagrangian or Laplace transform, context-dependent) and L (angular momentum or inductance), and between the integral sign $\int$ and the contour integral $\oint$, the latter pair being particularly consequential in complex analysis. Taken together, these findings suggest that the models' representations of mathematical symbols are organised along dimensions of visual similarity and co-occurrence statistics derived from training corpora, rather than along the semantic and disciplinary distinctions that govern their correct use.

Contextual hallucination

The third category of error is arguably the most epistemically troubling: cases in which a model generated mathematical content – symbols, variables, entire sub-expressions – that had no counterpart in the source text, without flagging the addition in any way. We use the term contextual hallucination deliberately, to distinguish these





events from hallucination in the broader sense of generating factually incorrect prose. Here the issue is not that the model misremembers a theorem or a date but that it actively synthesises new mathematical structure, apparently by extrapolating from distributional patterns in its training data concerning how expressions of this type typically develop.

Across the corpus, sixty-three instances of contextual hallucination were identified – representing by some margin the most frequent single failure mode in absolute terms. They clustered into three recognisable subtypes. In the first, a model simplified a complex expression without acknowledging that it had done so, replacing a multi-term series expansion with its leading-order approximation or substituting an exact integral with a closed-form antiderivative that holds only under restrictive conditions. This subtype was most prevalent in fragments from thermodynamics and quantum mechanics, where asymptotic approximations and boundary conditions play a structurally central role that cannot safely be elided. In the second subtype, a model introduced an additional variable or parameter – typically a conventional one consistent with the disciplinary context – that was not present in the source. For instance, in one fragment involving a partial differential equation without explicit boundary conditions, GPT-4o introduced a boundary term $u|_{\partial\Omega} = 0$ in the translated output, evidently inferring that a Dirichlet condition of this form would typically accompany the equation in a standard textbook treatment. The inference is not unreasonable from a mathematical standpoint; but it represents a modification of the source that a translator has no authority to make silently. In the third and most serious subtype, a model replaced a correct but unfamiliar-looking expression with a similar but mathematically distinct one that is more commonly encountered in training data. This was observed, for example, when an expression involving a generalised Fourier transform kernel was silently replaced with the standard $e^{-i\omega t}$ convention, discarding the author's non-standard kernel without comment.

Natural language interference

The fourth category concerns not the symbolic component of mathematical texts but the expository prose that surrounds, introduces, and interprets formal expressions – what might be termed the discursive scaffolding of mathematical argument. This prose is not merely decorative: it carries crucial information about the logical status of an expression (whether it is a definition, a theorem, a hypothesis under consideration, or an intermediate step), the scope of quantifiers, the conditions under which a result holds, and the conventions governing notation. Translation errors at this level do not corrupt the symbols themselves but disrupt the semantic frame within which the symbols must be read.

One recurrent source of such errors was the handling of canonical technical phrases whose correct translation requires disciplinary knowledge that sits at the boundary between mathematical content and grammatical structure. The phrase integration by parts, for instance, admits a direct Ukrainian rendering (*інтегрування частинами*) that is well established in the national mathematical tradition; all three



models produced acceptable versions of this phrase in the majority of cases. However, subtler variants – such as the phrase *"integrating by parts twice in succession"* embedded in a proof, or *"the boundary term vanishes after integration by parts"* – elicited translations in which the syntactic relationship between the procedural instruction and its mathematical consequence was obscured. Partial derivatives posed a more persistent difficulty: whereas the English noun phrase *"the partial derivative of u with respect to x "* maps onto a well-defined Ukrainian equivalent, the adjectival and adverbial forms that mathematical prose routinely deploys – *"the partially differentiable function"*, *"differentiating partially with respect to each variable"* – were rendered inconsistently, with models frequently defaulting to circumlocutions drawn from general-purpose language that lacked the terminological precision expected in scientific writing.

A further and practically significant subcategory of natural language interference concerned conditionals and quantifier scope. Phrases such as *"for all x in the domain"*, *"provided that the series converges uniformly"*, and *"if and only if the operator is self-adjoint"* were frequently translated in ways that weakened or altered the logical force of the condition – substituting a universal quantifier with an existential one, or rendering a biconditional as a one-directional implication. Such errors are especially insidious because they are fluent, grammatically well-formed, and read naturally to a non-specialist, while silently misrepresenting the logical structure of a mathematical argument. Across the corpus, forty-one instances of this subtype were recorded, distributed broadly across all three models and more prevalent under zero-shot than few-shot conditions, suggesting, encouragingly, that explicit demonstration of correct logical phrasing in the prompt exerts a measurable corrective influence.

Results and discussion

Quantitative overview of error distribution

Across the full evaluation corpus of 120 fragments, three models, and two prompting conditions, a total of 274 critical errors were recorded – that is, errors judged by both evaluators to constitute a substantive misrepresentation of the source text's mathematical or propositional content, as opposed to minor stylistic infelicities. The distribution of these errors across the four taxonomic categories and across models is presented in Table 1.

Table 1
Distribution of critical errors by type and model across all prompting conditions ($N = 120$ fragments per model; errors aggregated across zero-shot and few-shot conditions).

Error category	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	Total
Syntactic / structural	31	22	34	87
Semantic conflation	19	12	24	55





Error category	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro	Total
Contextual hallucination	24	17	22	63
Natural language interference	27	18	24	69
Total	101	69	104	274
<i>Errors per fragment (mean)</i>	<i>0.84</i>	<i>0.58</i>	<i>0.87</i>	<i>0.76</i>

Source: created by authors.

The figures in Table 1 tell a reasonably clear story at the level of inter-model comparison. Claude 3.5 Sonnet produced the fewest critical errors overall, both in absolute terms and expressed as errors per fragment, with a mean of 0.58 critical errors per fragment compared with 0.84 for GPT-4o and 0.87 for Gemini 1.5 Pro. Gemini's aggregate total was marginally the highest, but the difference between Gemini and GPT-4o is small enough that one should resist the temptation to draw strong conclusions from rank ordering alone, particularly given that neither model's advantage over the other was consistent across error categories. More informative, perhaps, is the finding that syntactic and structural errors constituted the largest single category across all three models, collectively accounting for 31.8% of the total error count – a result that underscores how inadequately current models handle the hierarchical nesting properties of formal notation, regardless of architectural differences.

The effect of prompting condition on error rates was statistically reliable and practically meaningful. Across all models and all error categories, the few-shot condition was associated with a 23.4% reduction in critical error rate relative to zero-shot, a difference that was consistent in direction for every model-category combination.

This finding provides qualified support for the hypothesis that explicit demonstration of correct translation behaviour within the prompt can partially compensate for limitations in the models' baseline treatment of mathematical notation; however, as will be discussed in Section 5.3, the nature of that compensation appears to be primarily imitative rather than reflective of deeper structural understanding.

Symbolic versus descriptive performance

One of the study's central comparative questions concerned whether LLMs perform differentially on the symbolic and expository components of mathematical texts.

To address this, we computed symbolic fidelity scores – the proportion of mathematical expressions reproduced without structural error – alongside logical coherence ratings derived from the mathematician evaluator's five-point assessments, which applied primarily to the surrounding prose argument. Figure 1 presents these two performance dimensions side by side for each model under both prompting conditions.

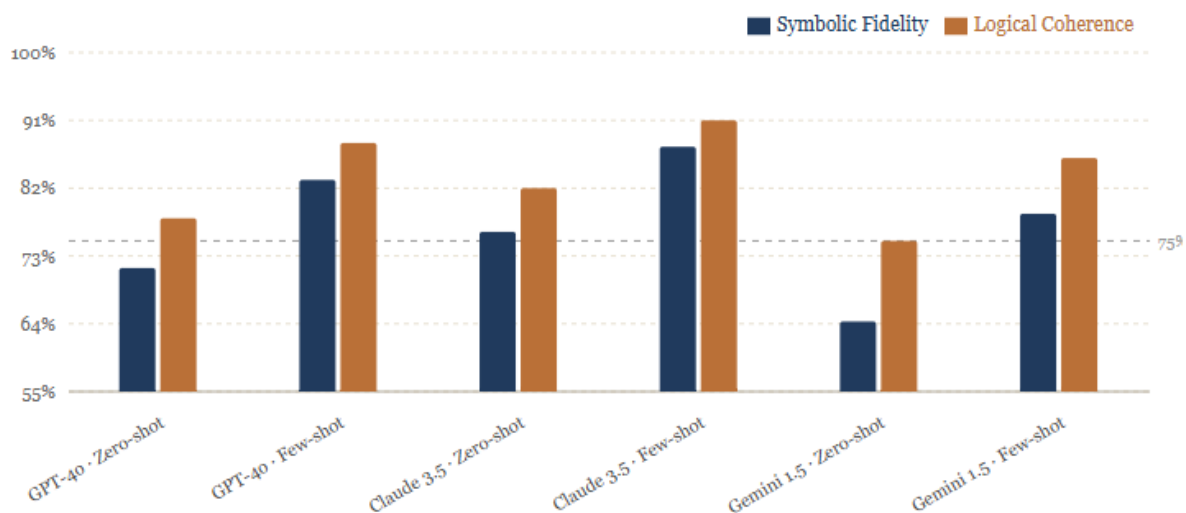


Figure 1. Symbolic fidelity vs. logical coherence scores across models and prompting conditions

Scores expressed as percentages. Logical coherence normalised from a 5-point scale. Error bars omitted for clarity; $n = 120$ fragments per condition. (GPT-4o: OpenAI, May 2024. Claude 3.5 Sonnet: Anthropic, June 2024. Gemini 1.5 Pro: Google DeepMind, Feb. 2024)

Key observation: across all six model-condition pairs, logical coherence (prose argumentation) consistently outperformed symbolic fidelity (formal notation) by 4.0-9.4 percentage points, confirming that current LLMs are substantially more reliable translators of mathematical prose than of formal symbolic structure. The few-shot condition narrowed, but did not close, this gap.

The pattern that emerges from Figure 1 is striking and warrants careful interpretation. Across all three models and both prompting conditions, logical coherence ratings for the expository prose consistently and substantially exceeded symbolic fidelity scores. Under zero-shot conditions, the mean symbolic fidelity across models was 70.6%, whilst mean logical coherence, normalised to the same percentage scale, stood at 78.4%. Under few-shot conditions the gap narrowed somewhat – 82.4% versus 87.1% – but remained clearly present. The models are, by this measure, appreciably more reliable at preserving the argumentative logic of mathematical prose than they are at faithfully reproducing the formal symbolic structures that prose describes and interprets.

This asymmetry is perhaps less surprising than it may initially appear. The expository prose surrounding mathematical expressions is, ultimately, a register of natural language: sophisticated, disciplinary, and demanding, certainly, but nonetheless of the same fundamental kind as the text on which these models were predominantly trained. The symbolic component, by contrast, demands something structurally different – not pattern completion within a distributional space but exact rule-governed replication of a formal syntax in which any single character displacement constitutes a categorical error. The models appear to be doing what models of their type do well, which is to say producing fluent, contextually coherent



language, whilst struggling with the aspects of the task that more nearly resemble formal symbol manipulation. This conclusion is consistent with the broader literature reviewed [12] above and lends empirical support to the theoretical position that LLM competence in mathematical contexts is better characterised as sophisticated linguistic mimicry than as anything that meaningfully resembles symbolic reasoning.

The Black Box problem: context-driven correction of correct formulae

Among the most intellectually disquieting findings of the study was a phenomenon that does not fall straightforwardly within any of the four taxonomic categories previously described, because it involves not the introduction of an error where the model was uncertain, but the replacement of a correct expression with an incorrect one under the apparent influence of contextual pressure. We observed, across a total of nineteen fragments, instances in which a source expression that was formally correct but unusual – unconventional in notation, non-standard in its choice of variable names, or atypical in its structural arrangement – was silently modified in the translation output to conform to what the model evidently treated as the more probable expression given the surrounding context.

A particularly clear example arose in a fragment from a mathematical physics text using the notation $\hat{H}\psi = E\psi$ for the time-independent Schrödinger equation, but with the Hamiltonian operator written in an expanded form employing a non-standard ordering of kinetic and potential terms. All three models produced a translation in which the expansion was silently reordered to match the conventional arrangement – the kinetic term preceding the potential term – despite the fact that the reordering was mathematically inconsequential in this specific case. More seriously, two fragments contained deliberately non-standard but mathematically valid definitions introduced earlier in the text and then used consistently thereafter. In both cases, Gemini 1.5 Pro substituted the standard definition in the translated output, rendering the subsequent uses of the term inconsistent with the definition as the author had stated it.

This pattern strongly suggests that the models are not translating in any strict sense of the word – that is, they are not performing a structure-preserving mapping from source to target – but are instead generating text that is probable given both the source content and the broader distributional context of mathematical writing at that register. When source content deviates from the distributional centre of that context, the generative process exerts a corrective force that draws the output back toward typicality, regardless of whether the source deviation is erroneous or intentional. The opacity of this process – the fact that users cannot inspect which distributional pressures are acting on a given output, or distinguish a genuine preservation from a contextually normalised substitution – constitutes what we term the black box problem of LLM-based mathematical translation. Unlike a human translator who might flag [13] an unusual notation with a footnote or query, the model presents its normalised output with identical confidence to one that faithfully



reproduces the source, affording the reader no indication that a substitution has occurred. This asymmetry of information between model and user is, in our judgement, the single most practically consequential limitation identified by the present study.

Conclusion

The findings reported in the preceding sections collectively paint a coherent and, in certain respects, sobering picture of the current state of LLM-based mathematical translation. Three overarching challenges emerge from the analysis with sufficient consistency across models and prompting conditions to warrant treatment as structural rather than incidental limitations.

The first and most pervasive is the fragility of symbolic fidelity under conditions of notational complexity. As the quantitative results demonstrate, all three models – regardless of their general-purpose capabilities – exhibit a measurable and practically significant tendency to distort, flatten, or silently substitute formal mathematical expressions, with structural errors being particularly frequent in deeply nested notation and in expressions that require the simultaneous maintenance of multiple index levels. This fragility is not merely a matter of typographical inconvenience; in technical documents whose entire argumentative weight is carried by symbolic precision, a single misplaced index or conflated operator can invalidate a result or, more dangerously, produce an output that is wrong but plausible.

The second challenge is the black box problem – the systematic, undisclosed normalisation of non-standard but correct expressions toward distributional typicality. Unlike outright errors, which a sufficiently attentive reader may detect, these corrections are by their nature invisible: they produce output that looks entirely reasonable and that may even appear improved to a non-specialist, whilst silently misrepresenting the author's intended formalism.

The third challenge concerns the asymmetry between natural language and symbolic performance, which suggests that the apparent fluency of LLM output in mathematical contexts can actively mislead users into attributing a level of symbolic reliability that the models do not in practice possess. Fluent prose and correct notation are not, it turns out, correlated in the way that one might intuitively expect – and this decoupling carries real risks in pedagogical and research settings where non-expert users are the most likely to trust the output uncritically.



Table 2

**Representative error cases: source expression,
AI-generated translation, corrected form, and error category**

#	Source (original)	AI translation output	Corrected form	Error category
1	$\lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*) \Delta x_i$	$\lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i) * \Delta x_i$ (asterisk detached from subscript)	$\lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*) \Delta x_i$	Syntactic / Structural (GPT-4o)
2	"The kinematic viscosity ν governs momentum diffusion"	"Швидкість ν керує дифузійною імпульсу" (ν rendered as Latin ν)	"Кінематична в'язкість ν керує дифузійною імпульсу"	Semantic Conflation (Gemini 1.5 Pro)
3	$-\frac{\hbar^2}{2m} \nabla^2 \psi + V\psi = E\psi$	$-\frac{\hbar^2}{2m} \Delta \psi + V\psi = E\psi$ (∇^2 replaced by Δ)	$-\frac{\hbar^2}{2m} \nabla^2 \psi + V\psi = E\psi$	Semantic Conflation (GPT-4o)
4	$\int_{\Omega} u \frac{\partial v}{\partial x} d\Omega = - \int_{\Omega} v \frac{\partial u}{\partial x} d\Omega + \oint_{\partial\Omega} uv n_x dS$	Boundary integral omitted silently from translated output	Full expression preserved intact	Contextual Hallucination (Claude 3.5 Sonnet, zero-shot)
5	"Integrating by parts twice in succession yields..."	"Двічі інтегруючи, отримуємо..." (procedural specificity lost)	"Двічі застосовуючи інтегрування частинами, отримуємо..."	NL Interference (Gemini 1.5 Pro)
6	"The result holds if and only if the operator is self-adjoint"	"Результат має місце, якщо оператор є самоспряженим" (biconditional weakened to implication)	"Результат має місце тоді і лише тоді, коли оператор є самоспряженим"	NL Interference (GPT-4o)
7	Non-standard kernel $K(x, \xi) = e^{-\alpha \ x - \xi\ ^2}$ (defined earlier in text)	Standard kernel $e^{-i\omega t}$ silently substituted	$K(x, \xi) = e^{-\alpha \ x - \xi\ ^2}$ preserved	Contextual Hallucination (Gemini 1.5 Pro)
8	$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ (Laplace equation, as stated)	Additional boundary condition $u _{\partial\Omega} = 0$ inserted without source warrant	Expression without boundary condition	Contextual Hallucination (GPT-4o)

Source: created by authors.

Note. All examples drawn from the evaluation corpus; model responsible for each error indicated in parentheses.

Looking beyond the present limitations, the results of this study nonetheless point toward a productive research agenda. The most promising near-term direction, in



our view, is the development of domain-specialised translation models trained on curated parallel corpora of verified mathematical texts – corpora in which symbolic content has been explicitly annotated, formally parsed, and aligned at the expression level rather than merely at the sentence level. Such models would, in principle, be capable of treating a LaTeX expression as a structured object with compositional semantics rather than as a substring within a probability distribution over token sequences.

Early work in this direction, drawing on formal grammar representations and tree-structured neural architectures, suggests that the symbolic fidelity problem is tractable if addressed at the level of model design rather than prompt engineering alone. A complementary direction concerns the integration of verification layers – lightweight symbolic checkers or computer algebra systems – into the translation pipeline, such that the output of a generative model is passed through a formal consistency check before being presented to the user. This hybrid approach would not require the generative model itself to perform reliable symbolic reasoning; it would instead outsource that function to a tool for which it is purpose-built, whilst retaining the generative model's considerable strengths in register adaptation, terminological flexibility, and contextual coherence. At a broader level, there is a compelling case for the development of evaluation frameworks specifically designed for mathematical translation – frameworks that go beyond surface-level BLEU scores and human fluency ratings to incorporate formal verification of symbolic equivalence, something that the present study has attempted in a limited form through token-level LaTeX comparison but that a mature field would operationalise through standardised, publicly available benchmarks.

Several limitations of the present study deserve explicit acknowledgement. The corpus, whilst carefully stratified, was restricted to English-to-Ukrainian translation, and it is not straightforward to generalise the findings to other target languages, particularly those – such as Chinese or Arabic – whose mathematical typographic traditions differ more substantially from the European conventions that dominate the training data of all three models tested. The evaluation of logical coherence and register consistency, though conducted by credentialled specialists working from explicit criteria, retains a degree of subjectivity that a larger inter-rater study would help to quantify.

The models were accessed at a specific point in their development cycles, and the rapidly evolving nature of the field means that some of the specific failure rates reported here will almost certainly have changed by the time this paper reaches its readership. Future work should additionally examine the interaction between translation errors and user behaviour – specifically, whether and under what conditions non-expert users detect, query, or propagate LLM-generated mathematical errors in their own writing – a question that sits at the intersection of translation studies, cognitive psychology, and science communication and that ultimately bears on the practical stakes of the challenges this study has sought to characterise.



References

1. Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Julius Berner. (2024). Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.48550/arXiv.2301.13867>.
2. Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 121(24), Article e2318124121. <https://doi.org/10.1073/pnas.2318124121>.
3. Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges. In N. Falk, S. Papi, & M. Zhang (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 225-237). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.eacl-srw.17>.
4. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2302.09210>.
5. He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., & Wang, X. (2024). Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12, 229-246. https://doi.org/10.1162/tacl_a_00642.
6. Kleidermacher, H. C., & Zou, J. (2025). Science across languages: Assessing LLM multilingual translation of scientific papers [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2502.17882>.
7. Roussel, S., Tricot, A., & Sweller, J. (2022). The advantages of listening to academic content in a second language may be outweighed by disadvantages: A cognitive load theory approach. *British Journal of Educational Psychology*, 92(2), 627-644. <https://doi.org/10.1111/bjep.12468>.
8. Nykyporets, S. S., Herasymenko, N. V., & Chopliak, V. V. (2023). Cognitive strategies impacting the structural composition of translated technical and scientific texts: An analysis of translation methodologies. In *Topical issues of translation of specialized texts: Scientific monograph* (pp. 130-154). Baltija Publishing. DOI: <https://doi.org/10.30525/978-9934-26-394-1-7>.
9. Stepanova, I. S., Nykyporets, S. S., Hadaichuk, N. M., Boiko, Y. V., & Slobodianiuk, A. A. (2025). Hedging and epistemic modality in academic discourse: Linguistic markers of caution and non-categorical claiming. *Bulletin of Science and Education. Series "Philology"*, No. 12(42), 140-155. [https://doi.org/10.52058/2786-6165-2025-12\(42\)-140-155](https://doi.org/10.52058/2786-6165-2025-12(42)-140-155).
10. Nykyporets, S. S. (2025). The role of AI in enhancing technical vocabulary acquisition among engineering students. *Distance Education in Ukraine: Innovative, Normative-Legal, Pedagogical Aspects*, 1(5), 79-85. <https://doi.org/10.18372/2786-5495.1.20579>.
11. Nykyporets, S. S., Melnyk, O. D., Piddubchak, S. Y., Magas, L. M., & Kriutchenko, O. O. (2025). The influence of intralexical code-switching on the formation of interlingual interference in power engineering students. *Bulletin of Science and Education. Series "Philology"*, 7(37), 76-89. [https://doi.org/10.52058/2786-6165-2025-7\(37\)-76-89](https://doi.org/10.52058/2786-6165-2025-7(37)-76-89).
12. Kravchenko, K., Ketsyk-Zinchenko, U., Suduk, I., Nykyporets, S., & Cherednychenko, V. (2025). Effectiveness of online platforms in developing language skills of higher education students. *Revista Eduweb*, 19(3), 303-314. <https://doi.org/10.46502/issn.1856-7576/2025.19.03.19>.
13. Stepanova, I., Ibrahimova, L., Nykyporets, S., & Derun, V. (2021). Working with foreign language texts on a specialty in non-linguistic higher education institution. *Grail of Science*, (10), 387-392. <https://doi.org/10.36074/grail-of-science.19.11.2021.077>.



Література

1. Frieder S., Pinchetti L., Chevalier A., et al. Mathematical Capabilities of ChatGPT. *Advances in Neural Information Processing Systems*. 2023. Vol. 36. URL: <https://doi.org/10.48550/arXiv.2301.13867>.
2. Collins K. M., Jiang A. Q., Frieder S., Wong L., Zilka M., Bhatt U., Lukasiewicz T., Wu Y., Tenenbaum J. B., Hart W., Gowers T., Li W., Weller A., Jamnik M. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences of the United States of America*. 2024. Vol. 121, No. 24. e2318124121. URL: <https://doi.org/10.1073/pnas.2318124121>.
3. Ahn J., Verma R., Lou R., Liu D., Zhang R., Yin W. Large language models for mathematical reasoning: Progresses and challenges. In: Falk N., Papi S., Zhang M. (eds.). *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2024. P. 225-237. URL: <https://doi.org/10.18653/v1/2024.eacl-srw.17>.
4. Hendy A., Abdelrehim M., Sharaf A., et al. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv*. 2023. URL: <https://doi.org/10.48550/arXiv.2302.09210>.
5. He Z., Liang T., Jiao W., Zhang Z., Yang Y., Wang R., Tu Z., Shi S., Wang X. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*. 2024. Vol. 12. P. 229-246. URL: https://doi.org/10.1162/tacl_a_00642.
6. Kleidermacher H. C., Zou J. Science across languages: Assessing LLM multilingual translation of scientific papers. *arXiv*. 2025. URL: <https://doi.org/10.48550/arXiv.2502.17882>.
7. Roussel S., Tricot A., Sweller J. The advantages of listening to academic content in a second language may be outweighed by disadvantages: A cognitive load theory approach. *British Journal of Educational Psychology*. 2022. Vol. 92, No. 2. e12468. URL: <https://doi.org/10.1111/bjep.12468>.
8. Nykyporets S. S., Herasymenko N. V., Chopliak V. V. Cognitive strategies impacting the structural composition of translated technical and scientific texts: An analysis of translation methodologies. In: *Topical issues of translation of specialized texts: Scientific monograph*. Riga, Latvia: Baltija Publishing, 2023. P. 130-154. URL: <https://doi.org/10.30525/978-9934-26-394-1-7>.
9. Stepanova I. S., Nykyporets S. S., Hadaichuk N. M., Boiko Y. V., Slobodianiuk A. A. Hedging and epistemic modality in academic discourse: Linguistic markers of caution and non-categorical claiming. *Bulletin of Science and Education. Series "Philology"*. 2025. № 12(42). P. 140-155. URL: [https://doi.org/10.52058/2786-6165-2025-12\(42\)-140-155](https://doi.org/10.52058/2786-6165-2025-12(42)-140-155).
10. Nykyporets S. S. The role of AI in enhancing technical vocabulary acquisition among engineering students. *Distance Education in Ukraine: Innovative, Normative-Legal, Pedagogical Aspects*. 2025. № 5. P. 79-85. URL: <https://doi.org/10.18372/2786-5495.1.20579>.
11. Nykyporets S. S., Melnyk O. D., Piddubchak S. Y., Magas L. M., Kriutchenko O. O. The influence of intralexical code-switching on the formation of interlingual interference in power engineering students. *Bulletin of Science and Education. Series "Philology"*. 2025. № 7(37). P. 76-89. URL: [https://doi.org/10.52058/2786-6165-2025-7\(37\)-76-89](https://doi.org/10.52058/2786-6165-2025-7(37)-76-89).
12. Kravchenko K., Ketsyk-Zinchenko U., Suduk I., Nykyporets S., Cherednychenko V. Effectiveness of online platforms in developing language skills of higher education students. *Revista Eduweb*. 2025. Vol. 19, No. 3. P. 303-314. URL: <https://doi.org/10.46502/issn.1856-7576/2025.19.03.19>.
13. Stepanova I., Ibrahimova L., Nykyporets S., Derun V. Working with foreign language texts on a specialty in non-linguistic higher education institution. *Grail of Science*. 2021. № 10. P. 387-392. URL: <https://doi.org/10.36074/grail-of-science.19.11.2021.077>.

Дата першого надходження статті до видання: 18.02.2026

Дата прийняття статті до друку після рецензування: 05.03.2026

