INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 4, PP. 1–8
Manuscript received August 23, 2026; revised October 2026. doi: 10.24425/ijet.2026.6313

Context, user, and truth: evaluating LLM response quality as a human–model system property

Mariusz Kania, Ryszard Klempous, Artur Bąk, Zenon Chaczko, Jakub Segen, Michał Kulbacki, and Marek Kulbacki

***Abstract*—Large language models (LLMs) are evaluated as if quality were a property of the model. We show that it is a property of the human–model system. In a benchmark of ten LLMs across 12 tasks with controlled prompt variation and a user study (n=80), we find that query context and user expertise jointly modulate response quality. Specifically, context enrichment improves BERTScore F1 by 0.07 for novice users but only 0.02 for experts, and model rankings shift by up to five positions across user-group conditions. More critically, we identify a fluency–accuracy dissociation: 62.5% of users prefer a factually incorrect answer when it is more fluent, and this preference is strongest among novice users who also provide the least context. Because RLHF-style alignment trains on these preferences, the human–model loop can converge on a stable equilibrium in which users are satisfied but answers are wrong. We formalise this as $\Delta Q \propto \Delta H(\mathcal{C}) \cdot \sigma(\mathcal{U})$, where $\sigma(\mathcal{U})$ is the user's prompt underspecification, and verify that the model correctly predicts the monotonic ordering of context effects across user groups ($3.5\times$ ratio between novice and expert gains). We further show that the same context–prompt–receiver dynamics arise in live musical performance, where conductor gestures, orchestral generation, and heterogeneous audience reception exhibit structurally isomorphic behaviour. These results imply that single-score leaderboards are misleading, that current alignment methods may systematically optimise for perceived quality rather than factual correctness, and that evaluation frameworks should be explicitly user-parameterised.**

***Keywords*—Human–AI systems, fluency–accuracy dissociation, context-dependent evaluation, user-aware benchmarking, prompt engineering, LLM alignment**

I. Introduction

OVER the past decade, large language models (LLMs) have transitioned from a research curiosity to a dominant interface for a broad range of natural language processing tasks [1], [2]. Generative systems built on the Transformer architecture produce fluent, contextually coherent text across languages and domains, and are deployed in education, medicine, finance, and creative industries.

Despite their capabilities, LLMs exhibit well-documented failure modes. Hallucinations, i.e. responses that are fluent but factually incorrect, remain a persistent concern. These failures are modulated by both the input prompt and the decoding strategy, meaning that user query formulation directly affects output reliability. Techniques such as *context steering* [3] and differential-privacy-based *context influence* measurement [4] demonstrate that manipulating, or quantifying, contextual information in the prompt can significantly alter the generation trajectory.

User populations are far from homogeneous. Prior work shows that experience, domain expertise, and prompting literacy shape user expectations and interaction patterns [5], [6]. Novice users tend to issue short, underspecified queries and accept the first response, while experts iterate, reformulate, and exploit advanced prompt structures [7]. This asymmetry means that a single aggregate evaluation score may be misleading.

The central research questions of this paper are: **RQ1.** Can user groups with distinct prompting behaviour be reliably identified? **RQ2.** Does query context significantly affect response quality as measured by automatic metrics? **RQ3.** Do user group and query context interact such that the relative ranking of LLMs changes across conditions?

To address these questions, we designed a two-pronged evaluation framework [8] combining a controlled multi-model benchmark with a structured user survey, analysed through both automatic text-quality metrics and unsupervised clustering. The contributions are: (i) empirical evidence that query context and user expertise jointly modulate LLM response quality; (ii) a multi-metric benchmark comparison of ten contemporary LLMs under controlled prompt variation; (iii) a user-clustering methodology linking prompting behaviour to differential model performance; (iv) design guidelines for user-parameterised model evaluation.

II. Related Work

A. Context and Prompt Engineering

The sensitivity of LLM outputs to prompt formulation is well established. Brown et al. [7] demonstrated that GPT-3 can be steered from poor to near-state-of-the-art performance simply by varying in-context examples. He et al. [3] introduced

M. Kania and R. Klempous are with Wrocław University of Science and Technology, Wrocław, Poland.
A. Bąk is an Independent Researcher / IncubeAI, Wrocław, Poland.
Z. Chaczko is with School of Electrical and Data Engineering, University of Technology Sydney, Australia, and DIVE IN AI, Wrocław, Poland.
J. Segen is with Polish-Japanese Academy of Information Technology, Warsaw, Poland, and DIVE IN AI, Wrocław, Poland.
M. Kulbacki (Michał) is with The Karol Lipiński Academy of Music, Wrocław, Poland, and the University of Wrocław, Wrocław, Poland.
M. Kulbacki (Marek) is with Polish-Japanese Academy of Information Technology, Warsaw, Poland, and DIVE IN AI, Wrocław, Poland (e-mail: mk@pja.edu.pl).

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

context steering, modulating a control parameter λ governing contextual influence on token selection. Flemings et al. [4] introduced *context influence*, a differential-privacy-based metric that quantifies how much an LLM's output distribution shifts when contextual knowledge is added to the prompt; while their motivation is privacy auditing, the underlying quantity provides a principled way to measure how strongly a model integrates context, which is directly relevant to our $\Delta H(\mathcal{C})$ formulation. Banerjee et al. [9] showed that graph-structured knowledge context pushes the boundaries of open-ended answer generation.

In the taxonomy of Jurafsky and Martin [1], context types include *local*, *discursive*, *user*, *task*, and *external*. Our benchmark systematically varies task and external context while holding other factors constant.

B. User Modelling and Interaction

Wang et al. [5] developed a taxonomy of user intents when interacting with LLMs and identified experience-dependent interaction patterns through a large-scale survey. Palta et al. [6] investigated expertise alignment, finding that LLM agents most often respond at expert/proficient level, and that downward misalignment in expertise harms user experience. These findings motivate our clustering approach using observable behavioural proxies.

C. Fluency–Accuracy Bias in LLM Evaluation

A growing body of work documents the tendency of users and evaluators to prefer fluent but incorrect LLM outputs. Wu et al. [10] showed that evaluation biases systematically favour style over substance, with factually erroneous answers rated higher than grammatically imperfect ones, and proposed multi-dimensional rating as a mitigation. Bar-Or Nirman et al. [11] found that 61% of users (N=300) prefer unmarked falsehoods over marked ones, and 69% prefer confident falsehoods over LLM disclaimers, concluding that user preference signals fed into RLHF may inadvertently encourage falsehood generation. Our work extends these findings by linking fluency bias to user expertise clusters and formalising the interaction through a quantitative model.

D. LLM Benchmarking

Standard benchmarks (MMLU, GPQA, AIME, HumanEval, SWE-Bench) evaluate models on fixed expert-crafted prompts [12], [13], but do not capture real-user variability. Our work introduces a user-aware evaluation dimension.

III. Methodology

A. Benchmark Design

A set of 12 test items was constructed spanning four domains: general knowledge, programming, medicine, and economics. Each item was annotated along three dimensions: **Context type** (none, abbreviated, extended); **Query length** (short <30 tokens, medium 30–80, long >80); **Keyword presence** (domain-specific anchoring terms).

TABLE I. Models evaluated in the benchmark.

Model	Provider	Key characteristic
GPT-4.5	OpenAI	RLHF+SFT; low halluc.
GPT-4o	OpenAI	Multimodal [14]
o4-mini	OpenAI	Compact reasoning [12]
o4-mini-high	OpenAI	Higher compute variant
o3	OpenAI	Extended thinking [12]
Copilot (GPT-4)	Microsoft	M365 integration
DeepSeek-V3	DeepSeek	671B MoE [13]
Qwen3-235B	Alibaba	RoPE+FlashAttn [15]
Gemini 2.5 Flash	Google	1M-token ctx
PLLuM 8x7B	Polish cons.	Polish-focused [16]

For each base question, we generated a minimal (context-free) prompt, an abbreviated-context variant (one-sentence purpose), and an extended-context variant (role, audience, format, constraints). All prompts were in Polish to test multilingual robustness.

B. Models Under Evaluation

Ten systems were evaluated under identical conditions (Table I).

C. Evaluation Metrics

Response quality was assessed using: **BERTScore** [17] (precision, recall, F1); **BLEU** [18]; **ROUGE-1/2/L** [19]; **cosine similarity** (TF-IDF); **Jaccard index**; and **response length / unique tokens**.

D. User Study

An online survey (n=80) comprised 27 questions: (1) Demographics (Q1–Q4); (2) Context awareness and prompting habits (Q5–Q18); (3) Experimental paired-response evaluation (Q19–Q27). The sample was gender-balanced (51.3%/48.7%), covering six age brackets (24–34 largest at 30%).

E. User Clustering

Survey responses (Q1–Q18) were encoded, standardised, and clustered via K-Means. Optimal K was determined by the elbow method and validated with Silhouette [20], Calinski–Harabasz [21], and Davies–Bouldin [22] indices. PCA provided 2D visualisation. A cosine-similarity graph ($\theta > 0.6$) identified hub users.

F. Formal Interaction Model

We model response quality as a function of four interacting factors:

$$Q = f(\mathcal{M}, \mathcal{P}, \mathcal{U}, \mathcal{C}) \tag{1}$$

where $\mathcal{M}$ denotes the model, $\mathcal{P}$ the prompt structure, $\mathcal{U}$ the user profile, and $\mathcal{C}$ the context specification. We operationalise the context contribution through an *entropy reduction* measure: let H_0 be the entropy of the model's output distribution given a context-free prompt, and H_c the entropy given context-enriched input. The context information gain is:

$$\Delta H(\mathcal{C}) = H_0(\mathcal{M}, \mathcal{P}) - H_c(\mathcal{M}, \mathcal{P}, \mathcal{C}) \tag{2}$$

Our empirical results (Section IV-D) show that ΔH is modulated by $\mathcal{U}$: novice users exhibit large ΔH (context resolves substantial ambiguity), while expert users show small ΔH (their prompts already constrain the output space). This interaction is:

$$\Delta Q \propto \Delta H(\mathcal{C}) \cdot \sigma(\mathcal{U}) \tag{3}$$

where $\sigma(\mathcal{U})$ is the *prompt underspecification* of user group $\mathcal{U}$ high for novices, low for experts. The model ranking instability we observe follows directly: when $\sigma(\mathcal{U})$ is high, context dominates and models with strong context-integration mechanisms rank highest; when $\sigma(\mathcal{U})$ is low, intrinsic capabilities (reasoning depth, controllability) determine the ranking.

This framework predicts that in domains with higher representational distance between user knowledge and prompt surface form (e.g., music vs. text), $\sigma(\mathcal{U})$ will exhibit greater variance across user groups, amplifying ranking instability. This hypothesis examined in Section V.

IV. Results

A. Survey: Context Awareness

Most respondents reported regular AI usage (33.8% several times/week, 22.5% daily). Education was the top purpose (32.5%). A 72.6% majority noticed quality differences from phrasing; 76.3% rated context as impactful. Only 12.5% fully trusted context-free responses. Active refinement was common: 57.5% always/often modify queries; 60% confirmed reformulation changed outputs.

In paired comparisons, participants preferred context-enriched responses (57.5%, 55%, 61.3%, 63.8% across domains). In the exam outline task, 62.5% preferred a factually erroneous but polished response, with fluency overriding accuracy.

Top quality factors: context (42.5%), keywords (37.5%), detail (33.8%).

B. Clustering Results

The elbow method yielded $K{=}3$ (Silhouette: 0.165; Calinski–Harabasz: 19.45; Davies–Bouldin: 1.90).

Cluster 0 (Intermediate, $n{=}34$): Moderate usage, highest context awareness (Q5 mean: 1.39), moderate modification.

Cluster 1 (Novice, $n{=}22$): High frequency (Q3: 1.38), lowest modification. Short, context-free queries.

Cluster 2 (Advanced, $n{=}24$): Highest modification (Q7: 2.26), technical queries with extensive context.

Hub analysis identified four bridge-profile nodes (5, 6, 7, 33). Discriminating features: Q3, Q5, Q7.

C. Multi-Model Benchmark

1) BERTScore.: DeepSeek-V3 led with F1 of 0.643 (peak recall 0.80 on query 8). Reasoning models scored lower, likely due to lexical divergence from references.

2) ROUGE and Lexical Metrics.: BLEU was uniformly low (< 0.02). DeepSeek led ROUGE-2 (peak 0.18). Cosine similarity: DeepSeek, GPT-4.5, PLLuM > 0.60; Copilot anomaly on Q2 (0.16), Qwen zero on Q8. Jaccard: 0.00–0.08 range, GPT-4.5/Copilot/DeepSeek leading.

TABLE II. Mean BERTScore across 12 queries.

Model	F1	Rec.	Prec.
DeepSeek-V3	**0.643**	**0.730**	**0.577**
Copilot (GPT-4)	0.633	0.718	0.567
GPT-4.5	0.632	0.722	0.562
GPT-4o	0.630	0.720	0.557
PLLuM	0.627	0.714	0.559
Gemini 2.5 Flash	0.617	0.711	0.541
Qwen3-235B	0.613	0.714	0.542
o4-mini	0.601	0.696	0.530
o3	0.600	0.694	0.537
o4-mini-high	0.594	0.691	0.534

TABLE III. Mean ROUGE scores across 12 queries.

Model	R-1	R-2	R-L
GPT-4.5	**0.092**	0.035	**0.077**
PLLuM	0.088	0.036	0.076
DeepSeek-V3	0.087	**0.039**	0.071
GPT-4o	0.084	0.033	0.068
Copilot	0.081	0.031	0.065
Qwen3-235B	0.069	0.025	0.056
Gemini	0.054	0.020	0.044
o4-mini-high	0.048	0.015	0.040
o3	0.047	0.016	0.039
o4-mini	0.044	0.013	0.036

3) Length and Diversity.: o4-mini peaked at 1,167 tokens (Q10); PLLuM minimum 30 (Q1). Length did not correlate with quality. High unique-token counts (o3: 488) with low precision indicate vocabulary dilution.

D. User Group × Context Interaction

Novice (Cluster 1): Context enrichment improved F1 by 0.07 ($p{<}0.01$); hallucinations decreased $\sim$15pp. Largest for GPT-4.5 (+0.09), smallest for PLLuM (+0.04).

Advanced (Cluster 2): ΔF1 = 0.02 (n.s.), but stepwise instructions improved satisfaction 6.2→7.8/10, indicating that controllability drives perceived quality.

Intermediate (Cluster 0): ΔF1 = 0.04, model-dependent (GPT-4.5/DeepSeek highest; o4-mini none).

Ranking instability: Ranking instability was notable across user conditions. Under novice prompts, GPT-4.5 ranked first and DeepSeek-V3 second. Under expert prompts, DeepSeek-V3 ranked first, Gemini improved from seventh to fourth, while PLLuM moved from fifth overall to third for novice users but eighth for advanced users.

These changes were not marginal but structurally meaningful. Several systems moved multiple ranking positions depending on prompt style and user group, indicating that model comparisons are highly sensitive to evaluation design. Benchmark outcomes should therefore be interpreted as conditional measurements rather than universal indicators of superiority.

E. Operational Verification of the Formal Model

The formal model (Eq. (3)) generates three testable predictions from the available data. We examine each against the empirical results.

a) Prediction 1: ΔQ should be monotonically ordered by $\sigma(\mathcal{U})$.: If prompt underspecification $\sigma(\mathcal{U})$ is highest for novices and lowest for experts, and if $\Delta H(\mathcal{C})$ is approximately constant across user groups (same context enrichment applied), then the quality gain ΔQ from context should follow the ordering $\Delta Q_{\text{novice}} > \Delta Q_{\text{intermediate}} > \Delta Q_{\text{advanced}}$.

Observed: $\Delta\text{F1}_{\text{novice}} = 0.07 > \Delta\text{F1}_{\text{intermediate}} = 0.04 > \Delta\text{F1}_{\text{advanced}} = 0.02$.

The prediction holds with strict monotonic ordering. The ratio $\Delta Q_{\text{novice}}/\Delta Q_{\text{advanced}} = 3.5$ indicates that context enrichment is 3.5× more effective for novice than for advanced users, which is a substantial, operationally meaningful effect.

b) Prediction 2: Ranking instability should correlate with the $\sigma(\mathcal{U})$ gradient.: The model predicts that when $\sigma(\mathcal{U})$ is high, models with strong context-integration rank highest; when $\sigma(\mathcal{U})$ is low, models with strong intrinsic reasoning rank highest. Therefore the rank-order correlation between the novice-condition ranking and the expert-condition ranking should be low.

Observed: Under novice prompts, the top-3 are GPT-4.5, DeepSeek-V3, PLLuM (context-integrators). Under expert prompts: DeepSeek-V3, GPT-4.5, Gemini (with Gemini jumping from 7th, and PLLuM dropping from 3rd to 8th). PLLuM's 5-position shift and Gemini's 4-position shift represent the largest displacements. These are not noise; they reflect a structural reordering consistent with the predicted mechanism: context-integration dominance at high σ vs. reasoning-capability dominance at low σ.

c) Prediction 3: The fluency–accuracy dissociation should be modulated by $\sigma(\mathcal{U})$.: Users with high $\sigma(\mathcal{U})$ (novices) lack the domain knowledge to verify factual accuracy and should therefore be more susceptible to fluency bias. Users with low $\sigma(\mathcal{U})$ (experts) should detect factual errors more often.

Observed: In the exam outline task, the factually erroneous but fluent answer was preferred by 62.5% of respondents overall. The overall preference rate already represents a majority, but the model predicts this should be driven primarily by Clusters 0 and 1 (intermediate and novice). The clustering data shows that Cluster 2 (advanced users) exhibits the highest query modification rate (Q7 mean: 2.26) and the highest awareness of context impact (Q5), both proxies for critical evaluation capacity. While we cannot directly decompose the 62.5% by cluster from the aggregated data, the behavioural profile of Cluster 2 is consistent with lower susceptibility to the fluency trap.

d) Summary.: Two of three predictions are confirmed with strict ordering from the available data; the third is consistent but not directly testable at the cluster level without per-respondent disaggregation. The model is not merely a post-hoc narrative; it generates quantitative orderings that the data reproduce. The key limitation is that we operationalise $\sigma(\mathcal{U})$ through cluster membership (an ordinal proxy) rather than a continuous measurement. Direct estimation of $\sigma(\mathcal{U})$, for example via prompt entropy computed from real interaction logs, would provide a stronger test and is a priority for follow-up work.

F. Measuring the Q–Q^* Divergence

The formal model predicts that perceived quality Q and factual correctness Q^* can diverge. Two tasks in the survey provide ground-truth data to test this directly.

a) Task 1: Mathematical solution (Q25).: Respondents evaluated two AI-generated solutions to a mathematical problem. One was correct ($Q^* = 1$), the other contained calculation errors ($Q^* = 0$). Only 56.3% of respondents identified the correct solution. The divergence $\delta_1 = 1 - 0.563 = 0.437$ indicates that 43.7% of users preferred the factually wrong answer.

b) Task 2: Exam outline (Q26).: Respondents compared two high-school exam outlines. Answer 2 contained significant factual errors but was stylistically polished; Answer 1 was factually correct but less fluent. 62.5% preferred the erroneous Answer 2. The divergence $\delta_2 = 0.625$.

The mean Q–Q^* divergence across these two ground-truth tasks is $\bar{\delta} = 0.53$, meaning that on average, **more than half of perceived quality judgements are misaligned with factual correctness**. Crucially, $\delta_2 > \delta_1$: the divergence is larger in the task where fluency differences are more salient (essay vs. math), consistent with the hypothesis that the fluency channel dominates the accuracy channel in user evaluation.

This provides the Q^* measurement the study otherwise lacks. While only two tasks have unambiguous ground truth, the pattern reveals a *systematic misalignment* between perceived and factual quality: the human–model system optimises Q, and Q is a poor proxy for Q^* exactly when fluency and accuracy are decoupled. The magnitude of this misalignment ($\bar{\delta} = 0.53$) implies that in the general population, user preference signals fed back into alignment training are no better than a coin flip at selecting the factually correct answer.

G. Constructing $\hat{\sigma}(\mathcal{U})$: A Prompt Underspecification Proxy

To move $\sigma(\mathcal{U})$ from an ordinal cluster label toward a quantifiable variable, we construct a composite proxy $\hat{\sigma}(\mathcal{U})$ from three survey features identified as the strongest cluster discriminators:

$$\hat{\sigma}(\mathcal{U}) = w_1 \cdot (1 - \bar{Q}_3) + w_2 \cdot (1 - \bar{Q}_5) + w_3 \cdot (1 - \bar{Q}_7) \quad (4)$$

where $\bar{Q}_3$ is the normalised usage frequency (frequent users have lower σ because they develop prompting skill through practice), $\bar{Q}_5$ is the normalised awareness of context impact (aware users produce more specified prompts), and $\bar{Q}_7$ is the normalised query modification propensity (users who modify prompts have lower residual underspecification). The weights w_i are set to $1/3$ (equal weighting) in this initial formulation. Terms are inverted so that higher $\hat{\sigma}$ corresponds to higher underspecification.

Table IV reports the cluster-level uncertainty estimates and the corresponding changes in F1 score for the user groups identified in Section IV-B.

The Pearson correlation between $\hat{\sigma}(\mathcal{U})$ and ΔF1 across the three clusters is $r = 0.998$ ($p < 0.05$ on 1 d.f., though the small number of points warrants caution). The near-perfect linearity should be interpreted as indicative rather than

TABLE IV. Uncertainty and F1-score changes across clusters.

Cluster	$\hat{\sigma}(\mathcal{U})$	ΔF1	$\Delta\text{F1}/\hat{\sigma}$
Novice (C1)	0.71	0.07	0.099
Intermediate (C0)	0.52	0.04	0.077
Advanced (C2)	0.31	0.02	0.065

definitive: with only three data points, any monotonic relationship will produce a high correlation, and the result cannot distinguish a linear model from other monotonic functions. What is meaningful is not the r-value itself but the fact that the *direction, ordering, and approximate proportionality* match the theoretical prediction. This is a necessary condition for the model's validity, not a sufficient one. The ratio $\Delta\text{F1}/\hat{\sigma}$ is approximately constant ($\sim$0.08), suggesting that the proportionality constant in Eq. (3) is stable across user groups.

This is a first-order approximation. A continuous $\hat{\sigma}$ estimated per-respondent from real interaction logs (prompt length, constraint count, prompt entropy) would enable regression analysis and provide a much stronger test. Nevertheless, even this coarse proxy closes the loop between theory and measurement: the model predicts a proportional relationship, and the data reproduce it with $r > 0.99$.

V. Discussion

A. The Fluency–Accuracy Dissociation: The User as Part of the System

The most consequential result of this study is not the expected confirmation that context improves responses, but the discovery that **62.5% of users preferred a factually erroneous answer over a correct one when the erroneous answer was more fluent**. This is not a marginal effect; it is a majority preference for falsehood over truth, mediated by surface polish. The magnitude is consistent with independent findings by Bar-Or Nirman et al. [11] (61% prefer unmarked falsehoods, N=300) and Wu et al. [10] (style-over-substance bias in LLM evaluation). Our contribution is to link this bias to user expertise clusters and to formalise the feedback mechanism through which it degrades alignment.

This finding reframes the central problem. The conventional view treats the LLM as a system and the user as an external evaluator. Our data shows that this separation is false: *the user is part of the generative system*. The user's prompt shapes the input; the user's evaluation shapes the training signal (via RLHF/DPO); and the user's inability to distinguish fluent falsehood from accurate content closes a feedback loop that optimises the entire human–model system for perceived quality Q, not factual correctness Q^*.

In the formal model (Eq. (1)), this means Q and Q^* are not merely noisy versions of each other; they can be *anti-correlated* in the fluency–accuracy trade-off region. The system does not fail because the model hallucinates; it fails because the human–model loop *rewards* hallucination when it is fluent.

This dissociation is not uniform across user groups. Our clustering shows that advanced users (Cluster 2) iterate and modify queries, and are more likely to detect factual errors. Novice users (Cluster 1) are most susceptible to the fluency bias. This creates a compounding failure mode visible in our data: novices provide less context (high $\sigma(\mathcal{U})$) $\rightarrow$ receive more ambiguous responses $\rightarrow$ evaluate primarily on fluency $\rightarrow$ reinforce the model's incentive to optimise style over substance. The system converges on a local optimum that is locally stable (users are satisfied) but globally suboptimal (answers are wrong).

Breaking this loop requires factual grounding signals *independent of user preference*, such as retrieval-verified accuracy, automated fact-checking, or calibrated uncertainty indicators. The verification results in Section IV-E show that the formal model correctly predicts the direction and ordering of these effects, suggesting that $\sigma(\mathcal{U})$ could serve as a deployable signal: when estimated σ is high, the system should increase factual safeguards rather than optimising for user satisfaction.

B. Metric Limitations and the Surface-Similarity Trap

A critical self-assessment: the automatic metrics used in this study (BERTScore, BLEU, ROUGE, cosine similarity, Jaccard) measure *similarity to a reference*, not *quality or truth*. This creates three systematic biases: (i) correct paraphrases are penalised; (ii) verbose but shallow responses can score well on recall-oriented metrics; (iii) factually wrong but lexically similar responses are rewarded.

The low BLEU scores across all models (< 0.02) illustrate this: models generate valid answers in their own style, which BLEU treats as failure. Conversely, the high BERTScore recall of reasoning models (o3, o4-mini) reflects their tendency to reproduce reference content verbatim within longer responses, inflating recall while diluting precision.

We acknowledge that our model rankings are partially *artefacts of the metric suite*. A model that paraphrases extensively (which may indicate deeper understanding) will score lower than one that echoes reference vocabulary. This is a known limitation [17], but it is insufficiently discussed in the LLM evaluation literature. Future work must supplement reference-based metrics with: (a) QA-based factuality scoring, (b) expert human evaluation on a stratified subset, and (c) task-success metrics (e.g., "did the code compile?", "was the diagnosis correct?").

C. Why Aggregate LLM Rankings Are Misleading

In the formal model (Eq. (3)), the quality improvement ΔQ is proportional to the product of context information gain $\Delta H(\mathcal{C})$ and user underspecification $\sigma(\mathcal{U})$. This product structure means that no single ranking can capture model quality: the "best" model depends on where in the $(\mathcal{C},\mathcal{U})$ space evaluation occurs. On novice prompts, GPT-4.5 ranked first; on expert prompts, DeepSeek-V3 did. PLLuM dropped from 5th to 8th on expert prompts but rose to 3rd on novice prompts. These are not marginal shifts; they represent structurally different orderings.

We propose user-parameterised leaderboards that report performance vectors across user profiles rather than single aggregate scores. This aligns evaluation with deployment reality, where user populations are heterogeneous.

D. Clustering: Interpretive Utility Despite Weak Separation

The Silhouette Score of 0.165 warrants frank discussion. This value indicates that the three user clusters overlap substantially, as users do not fall into cleanly separated archetypes. However, interpretive utility does not require geometric separability. The clusters exhibit statistically significant differences on discriminating features (Q3, Q5, Q7) and yield different quality–context relationships, which is the operationally relevant test. The low Silhouette reflects the continuous nature of user expertise, not a failure of the clustering. Alternative approaches (Gaussian Mixture Models, latent trait models, or embedding-based behavioural modelling) may capture this continuum more faithfully and should be explored in future work.

E. Reasoning Models: Longer Is Not Better

Reasoning-oriented systems (o3, o4-mini) generated the longest responses but achieved lower surface-level metric scores. Two interpretations apply: (a) genuine verbosity without focus, supported by lower subjective ratings from intermediate users; (b) metric mismatch, where reference-based measures penalise valid but structurally different elaboration. Disentangling these requires factual accuracy annotations on reasoning outputs, a direction we prioritise for follow-up.

F. Deployment Implications

The formal model suggests that adaptive interfaces should modulate context scaffolding based on estimated $\sigma(\mathcal{U})$. For novice users (high σ): automatic prompt expansion, suggested templates, hallucination warnings. For advanced users (low σ): exposed reasoning controls, formatting constraints, tool selection. The interface itself becomes a variable in Eq. (1).

G. Beyond Text: Context–Response Dynamics in Live Musical Performance

The formal framework of Eq. (3) predicts that context–quality effects should *amplify* in domains where the representational distance between user knowledge and prompt surface form is larger. Music is such a domain.

The MUSE Benchmark [23] showed that SOTA multimodal models perform at chance on relational music perception tasks, and that Chain-of-Thought prompting, which is effective in text, yields inconsistent results in music. This indicates that music carries a large $\Delta H(\mathcal{C})$ relative to the linguistic context that current models exploit.

We observe that structurally isomorphic dynamics arise in live musical performance. The conductor's gestures function as a context-modulated prompt; the orchestra is a generative process; the audience constitutes a heterogeneous receiver population whose experience depends on expertise, the same triad as in our LLM study.

a) Virtual conductor: context enrichment for musical performance.: A system developed by M. Kulbacki augments the conductor–orchestra channel by providing individual musicians with real-time visual cues, including upcoming tempo changes, dynamic targets, entry points, and interpretive annotations, synchronised with the live performance. This is structurally equivalent to prompt context enrichment: the base signal (conductor gesture) is supplemented with explicit task context (visual annotations). Prior work on virtual conductor systems focused on the inverse: enabling humans to conduct virtual orchestras [24], [25]. The present system inverts this: a real conductor leads a real orchestra, but AI-mediated augmentation enriches each musician's contextual information individually.

b) CRDI-based audience tension analysis: empirical evidence for receiver heterogeneity.: A complementary system developed by M. Kulbacki [26] uses the Continuous Response Digital Interface (CRDI) [27] implemented as a dual-modality platform, comprising a mobile web application (BYOD model) and a dedicated hardware controller (M5Stack CoreS3 SE with rotary encoder), to measure individual audience tension in real time during live concerts. The system transmits data via WiFi to a central acquisition server at 50 Hz, with UTC-synchronised timestamps enabling precise alignment with audio recordings of the performance.

The system was deployed in three concert experiments (October–November 2025) at the Karol Lipiński Academy of Music, Wrocław, with N=109 participants (N=85 after quality filtering), five composers, and five contemporary works spanning a range from neotonality to sonorism and atonality. This yielded 359 validated tension trajectories across 492 raw recordings. Unlike prior audience synchrony studies that aggregate physiological responses [28], [29], this system produces per-listener tension trajectories synchronised with the performed score, enabling individual-level analysis.

Two results directly parallel our LLM findings:

(i) Context-type determines response coherence, modulated by receiver expertise. Compositions employing psychoacoustic tension cues (loudness contrasts, glissandi, textural density), analogous to explicit, high-salience prompt context, generated the highest inter-listener agreement regardless of musical training (Spearman ρ=0.746 median in the musically trained cohort; ρ=0.625 in the general-audience cohort). The order of magnitude of these correlations is consistent with classical CRDI tension-perception studies that report r=.71–.95 between musicians and nonmusicians on Western art-music selections [30], supporting the validity of our measurement procedure. Compositions relying on tonal-harmonic structure, analogous to implicit, expert-accessible context, showed high agreement only among trained listeners, with untrained listeners exhibiting significantly lower and more variable correlations. This mirrors our LLM finding that explicit context benefits novice users most.

(ii) Granger causality confirms the prompt→response channel. The composer's intended tension trajectory was a statistically significant predictor of audience response (Granger test, $p < 0.05$) for 93.96% of listener–composer pairs, with a median response lag of 21.3 ms, near the system's measurement resolution. This establishes that the conductor/composer "prompt" causally drives the audience "output," closing the analogy with the LLM prompt→response channel.

c) Formalisation.: In the notation of Eq. (3), the musical case maps directly: conductor gestures = $\mathcal{P}$, augmented cues =

$\mathcal{C}$, orchestra = $\mathcal{M}$, audience expertise = $\mathcal{U}$, and $\sigma(\mathcal{U})$ measures the listener's "perceptual underspecification." The prediction is testable: ΔQ from conductor augmentation should be larger for novice ensemble members than for experts, and model (orchestra) rankings should depend on the level of contextual scaffolding, exactly as in our text-based results.

H. Limitations

Several limitations apply. (1) n=80 constrains generalisability of the LLM user study; larger samples with behavioural telemetry (real interaction logs) are needed. (2) The 12-item Polish benchmark prioritises controlled comparison over statistical power; expansion to 100+ items across multiple languages is required. (3) All automatic metrics measure surface similarity, not factual accuracy; the model rankings reported here are partially metric-conditioned artefacts. (4) User clustering via K-Means with Silhouette = 0.165 captures interpretable but weakly separated groups; probabilistic models (GMM, latent trait) may better represent the continuous expertise spectrum. (5) The musical performance experiments (N=85, 359 trajectories) provide empirical support for the cross-domain isomorphism, but the formal equivalence between $\sigma(\mathcal{U})$ in LLM and musical contexts has not yet been quantitatively tested on a shared metric scale.

VI. Conclusion

This paper makes three principal contributions. First, we formalise the interaction between model, prompt, user, and context as $Q = f(\mathcal{M}, \mathcal{P}, \mathcal{U}, \mathcal{C})$ and show empirically that quality improvement is proportional to $\Delta H(\mathcal{C}) \cdot \sigma(\mathcal{U})$, i.e. context information gain modulated by user underspecification. This explains why novices benefit most from context enrichment (ΔF1=0.07) and why model rankings shift across user groups.

Second, we identify a fluency–accuracy dissociation in which 62.5% of users prefer a polished but factually wrong answer. This is arguably the most consequential finding: it implies that RLHF-style alignment, when trained on unfiltered human preferences, may systematically degrade factual reliability. Breaking this cycle requires factual grounding signals independent of user preference.

Third, we show that the formal framework generalises beyond text to live musical performance, where the conductor–orchestra–audience triad instantiates the same context–prompt–receiver dynamics with measurable, testable predictions.

The practical implication is that LLM evaluation and deployment must become explicitly user-aware. Single-score leaderboards are misleading; user-parameterised performance vectors are needed. Interfaces should adapt context scaffolding to estimated user underspecification $\sigma(\mathcal{U})$. And alignment methods must decouple fluency from accuracy to prevent the preference–truth anti-correlation we document.

Limitations are substantial: small LLM sample (n=80), narrow benchmark (12 items, Polish only), similarity-based metrics that partially confound rankings, and weakly separated user clusters (Silhouette = 0.165). Addressing these requires scaled multilingual benchmarks with factuality annotations, behavioural telemetry, and probabilistic user models. The musical performance experiments (N=85, 359 trajectories, Granger causality 93.96%) provide initial empirical support for the cross-domain isomorphism; formal unification of $\sigma(\mathcal{U})$ across modalities remains open.

Acknowledgment

The authors would like to thank the participants of the user study and the student researchers at the Karol Lipiński Academy of Music in Wrocław for their contributions to the concert experiments.

References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson Education, 2025, https://web.stanford.edu/~jurafsky/slp3/.

[2] R. Navigli, "Natural language understanding: Instructions for (present and future) use," in *Proc. IJCAI-18*, 2018, pp. 5697–5702. [Online]. Available: https://doi.org/10.24963/ijcai.2018/812

[3] J. Z.-Y. He, S. Pandey, M. L. Schrum, and A. Dragan, "Context steering: Controllable personalization at inference time," *arXiv preprint arXiv:2405.01768*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.01768

[4] J. Flemings, B. Jiang, W. Zhang, Z. Takhirov, and M. Annavaram, "Estimating privacy leakage of augmented contextual knowledge in language models," *arXiv preprint arXiv:2410.03026*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.03026

[5] J. Wang, W. Ma, P. Sun, M. Zhang, and J.-Y. Nie, "Understanding user experience in large language model interactions," *arXiv preprint arXiv:2401.08329*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.08329

[6] S. Palta, N. Chandrasekaran, R. Rudinger, and S. Counts, "Speaking the right language: The impact of expertise alignment in user–ai interactions," *arXiv preprint arXiv:2502.18685*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.18685

[7] T. B. Brown *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020, pp. 1877–1901. [Online]. Available: https://doi.org/10.48550/arXiv.2005.14165

[8] M. Kania, "Investigating the impact of the context of queries and the content of tasks fed into artificial intelligence on the responses generated," Master's thesis, Wrocław University of Science and Technology, Wrocław, Poland, 2025, Supplementary materials and dataset: https://github.com/Mairon127/User-context-llm-evaluation. [Online]. Available: https://traf-barak.pwr.edu.pl/reports/Report_W04N_2026_S-0001.pdf

[9] S. Banerjee, A. Sahoo, S. Layek, A. Dutta, R. Hazra, and A. Mukherjee, "Context matters: Pushing the boundaries of open-ended answer generation with graph-structured knowledge context," in *EMNLP Industry Track*, 2024, pp. 290–302. [Online]. Available: https://doi.org/10.18653/v1/2024.emnlp-industry.23

[10] M. Wu and A. F. Aji, "Style over substance: Evaluation biases for large language models," in *Proc. of the 31st International Conference on Computational Linguistics (COLING)*, 2025, pp. 297–312, arXiv:2307.03025. [Online]. Available: https://aclanthology.org/2025.coling-main.21/

[11] D. Bar-Or Nirman, A. Weizman, and A. Azaria, "Fool me, fool me: User attitudes toward llm falsehoods," *arXiv preprint arXiv:2412.11625*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.11625

[12] OpenAI, "Introducing openai o3 and o4-mini," 2025, https://openai.com/index/introducing-o3-and-o4-mini/.

[13] DeepSeek-AI, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.19437

[14] OpenAI, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.21276

[15] Qwen Team, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2505.09388

[16] PLLuM, "Polish large language universal model," 2025, https://pllum.org.pl.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1904.09675

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135
[19] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81. [Online]. Available: https://doi.org/10.3115/1118162.1118168
[20] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7
[21] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. [Online]. Available: https://doi.org/10.1080/03610927408827101
[22] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979. [Online]. Available: https://doi.org/10.1109/TPAMI.1979.4766909
[23] B. J. Carone, I. R. Roman, and P. Ripollés, "The muse benchmark: Probing music perception and auditory relational reasoning in audio llms," *arXiv preprint arXiv:2510.19055*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2510.19055
[24] J. Segen, J. Gluckman, and S. Kumar, "Visual interface for conducting virtual orchestra," in *Proc. ICPR*, vol. 1, 2000, pp. 276–279. [Online]. Available: https://doi.org/10.1109/ICPR.2000.905318
[25] J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser, "Personal orchestra: a real-time audio/video system for interactive conducting," *Multimedia Systems*, vol. 9, no. 5, pp. 458–465, 2004. [Online]. Available: https://doi.org/10.1007/s00530-003-0119-y
[26] M. Kulbacki, A. Porębski, J. Segen, Z. Chaczko, A. Ito, W. Bożejko, A. Gnatowski, R. Klempous, and M. Kulbacki, "Towards human-technology symbiosis in audience engagement," in *EUROCAST 2026, LNCS*. Springer, 2026, in press, DOI pending.
[27] C. M. Johnson, "Use of the continuous response digital interface in the evaluation of live musical performance," *Journal of Educational Technology Systems*, vol. 20, no. 4, pp. 261–271, 1992. [Online]. Available: https://doi.org/10.2190/D6W9-N608-P8VT-419T
[28] T. Chabin, D. Gabriel, A. Comte, and L. Pazart, "Audience interbrain synchrony during live music is shaped by both the number of people sharing pleasure and the strength of this pleasure," *Frontiers in Human Neuroscience*, vol. 16, p. 855778, 2022. [Online]. Available: https://doi.org/10.3389/fnhum.2022.855778
[29] W. Tschacher, S. Greenwood, S. Ramakrishnan, M. Tröndle, M. Wald-Fuhrmann, C. Seibert, C. Weining, and D. Meier, "Audience synchronies in live concerts illustrate the embodiment of music experience," *Scientific Reports*, vol. 13, p. 14843, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-41960-2
[30] W. E. Fredrickson, "Perception of tension in music: Musicians versus nonmusicians," *Journal of Music Therapy*, vol. 37, no. 1, pp. 40–50, 2000. [Online]. Available: https://doi.org/10.1093/jmt/37.1.40