



Prompt Clarity, Hallucination Risk, and Trustworthy LLM Performance: Mediation Analysis, Cross-Lingual/Domain Robustness, and Routing Gains

Hathairat Ketmaneechairat
Faculty of Information Technology, King Mongkut's University
of Technology North Bangkok, Bangkok, Thailand
hathairat.k@cit.kmutnb.ac.th

ABSTRACT

Despite the rapid advancement of Large Language Models (LLMs), hallucinations remain a critical barrier to their trustworthy deployment in high stakes domains. This study examines how prompt clarity influences LLM response quality and trustworthiness, focusing on the mediating role of hallucination risk. Utilizing a synthetic benchmark of 15,000 prompt response pairs spanning five languages, multiple domains, four difficulty levels, and five model families, we employed causal style mediation analysis, fairness audits, and routing policy simulations. Results indicate that hallucination risk is the dominant mediator, accounting for approximately 66% to 71% of the total effect of prompt clarity on response quality. While LLM performance demonstrated strong cross-lingual and cross-domain robustness, trustworthiness measured via a novel Trustworthy Response Index (TRI) declined significantly as task difficulty increased. Furthermore, model family significantly moderated the mediation pathway; Reasoning-LLMs showed greater resilience to prompt ambiguity than other families. System level simulations revealed that intelligent model routing, particularly strategies prioritizing Reasoning LLMs, yields substantial gains in overall quality and TRI. These findings underscore that prompt engineering and model aware routing are interconnected determinants of reliable AI, necessitating integrated strategies for trustworthy LLM deployment.

Keywords: Large Language Models (LLMs), Hallucination Risk, Prompt Clarity, Mediation Analysis, Trustworthy Response Index (TRI), Model Routing, AI Reliability, Cross-Lingual Robustness

Received: 6 April 2026, Revised: 11 July 2026, Accepted: 28 July 2026

Copyright: DLINE

1. Introduction and Background

Large Language Models (LLMs) have recently emerged as foundational technologies in modern natural lan

guage processing (NLP), supporting an expanding range of applications from conversational agents and scientific writing assistants to automated code generation. Their growing adoption is closely tied to their extraordinary capabilities in zero shot and few shot learning tasks [1]. These capabilities have substantially broadened LLMs' potential applications across diverse domains. However, despite their remarkable performance, a fundamental challenge remains: hallucinations, in which an LLM produces content that is fluent and coherent but factually incorrect, fabricated, or logically inconsistent [2, 3, 4].

Hallucinations pose a substantial challenge to the reliability and efficiency of AI-based systems. The problem becomes especially consequential when LLMs are deployed in high impact domains where inaccurate or fabricated information can have significant downstream consequences. Existing research has documented concerns associated with LLM use in medicine [5], law [6], journalism [7], and scientific communication [8, 9]. Beyond the immediate problem of incorrect outputs, hallucinations may contribute to misinformation, diminish user trust, and create accountability gaps in AI-mediated decision-making and communication [10, 11]. Consequently, understanding the mechanisms, manifestations, and determinants of hallucination remains a key research priority.

1.1 Hallucination and the Internal Dynamics of LLMs

Researchers have traditionally investigated hallucination by analysing model outputs, focusing on factual inaccuracies or inconsistencies in the final generated response. However, emerging perspectives suggest that hallucination may also stem from information deficiencies within LLMs' internal computational processes. Kim revealed that hallucination manifests as usable information deficiencies in inter layer transmissions. Rather than focusing exclusively on the final layer output, this perspective emphasizes the importance of examining information dynamics across the model's layers. Tracking cross layer information dynamics ($\mathcal{LI}$) therefore provides potentially robust indicators of model reliability [12].

This internal perspective broadens the conceptualization of hallucination from an output level error to a phenomenon potentially associated with the flow and transformation of information throughout the model. Such an approach is particularly relevant for trustworthy LLM evaluation because final outputs alone may not fully capture the processes through which unreliable information emerges. Monitoring cross layer information dynamics may consequently provide additional evidence for assessing the reliability of model behavior.

The broader body of research further indicates that hallucination cannot be understood as a simple or isolated software defect. A comprehensive account of its causes encompasses factors related to data quality, model architecture, training processes, and inference mechanisms. Together, these factors suggest that hallucination is better understood as an emergent property of the current LLM design paradigm rather than a single, independently identifiable bug [13, 14, 15]. This interpretation highlights the need for multidimensional approaches to hallucination assessment that consider both model characteristics and the conditions under which the model is prompted, trained, and deployed.

1.2 Trustworthiness as a Multidimensional Property of LLMs

Because hallucinations directly affect the reliability of generated information, their investigation is closely connected to the broader question of LLM trustworthiness. Liu presented a comprehensive survey identifying key dimensions for assessing LLM trustworthiness. The survey organizes LLM trustworthiness into seven major categories: reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness [9]. This multidimensional framework demonstrates that trustworthy LLM performance cannot be reduced to factual accuracy alone. Instead, trustworthy behavior encompasses the capacity of a model to produce reliable outputs while remaining safe, fair, robust, explainable, resistant to misuse, and consistent with relevant social expectations.

Within this broader framework, alignment has been proposed as a crucial step toward developing trustworthy LLMs. Alignment seeks to ensure that models can effectively benefit and serve human users constructively [16, 17]. This perspective connects model performance with the intended objectives, expectations, and interests

of human users. Consequently, improving trustworthiness requires attention not only to what models can generate but also to whether their behavior is appropriately aligned with human needs and expectations.

Explainability and transparency constitute another important dimension of trustworthy LLM performance. A trustworthy LLM should be able to explain its reasoning and be transparent about how it generates content. Nevertheless, the black box nature of most machine learning models makes it difficult for users to understand the reasoning underlying model decisions. This limitation raises particular concerns when LLMs are used in critical scenarios and across domains where users may need to evaluate, verify, or challenge model-generated information [18-23]. Thus, transparency and explainability are central considerations in evaluating whether LLM outputs can be appropriately trusted and incorporated into consequential workflows.

1.3 Prompt Quality and the Reliability of LLM Outputs

Beyond intrinsic model characteristics, prompt quality is an important factor in determining LLM performance. Systematic reviews remain labor intensive, particularly when methodological details must be extracted from full text scholarly publications. Using mediation analysis as a case study, Zhang, [24] evaluated whether LLMs can achieve human expert level performance in full text methodological review of key causal assumptions, including the absence of unmeasured confounding and temporal ordering. The evaluation also considered methodological best practices such as sensitivity analyses, interaction assessments, and covariate adjustment. This work illustrates LLMs' potential to support complex methodological evaluation while also highlighting the importance of carefully assessing their performance on research intensive tasks.

Prompt quality has consequently become an important component of LLM evaluation and application. Reinhard systematically examined evaluation criteria and optimization methods that can improve prompt quality. Drawing on a systematic literature review, the study identified clarity, accuracy, and precision as important criteria for evaluating prompts, together with initial measurement techniques for assessing prompt quality. The review also synthesized commonly employed optimization approaches, including iterative refinement and shot based prompting. [25] These findings indicate that prompt construction is not merely a superficial interactional consideration but an important component of effective LLM utilization.

Jacobsen further supports the link between prompt quality and model selection, showing that both LLM choice and prompt quality predict performance. The findings suggest that targeted prompt engineering principles can generate high quality feedback, and that domain specific language is particularly effective. Consequently, the selection of appropriate LLMs and the development of effective prompting skills should be fostered as relevant characteristics of AI literacy [26].

1.4 Research Context and Rationale

Taken together, the literature indicates that multiple interacting dimensions shape reliable LLM performance. Hallucination poses a major threat to reliability because fluent, coherent outputs can still contain factual errors, fabricated information, or logical inconsistencies. At the same time, hallucination appears to arise from multiple sources, including data quality, model architecture, training processes, and inference mechanisms, suggesting it should be treated as an emergent property rather than a simple, isolated defect. [13, 14, 15].

The literature further demonstrates that trustworthiness encompasses a broader set of properties, including reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness [9]. Alignment and transparency are therefore essential considerations when evaluating the extent to which LLMs can be responsibly integrated into human centered applications [16, 17]. At the same time, prompt quality provides an important operational dimension because clarity, accuracy, precision, iterative refinement, shot-based prompting, and domain specific language can influence model performance [25, 26]. These strands of research collectively motivate a more integrated examination of LLM performance that considers prompt characteristics, hallucination risk, model reliability, methodological performance, and trustworthiness together. Such an approach is particularly important because improving model capability alone does not necessarily guarantee trustworthy behavior. Instead, reliable LLM deployment requires

considering the interaction among model characteristics, prompting strategies, internal information dynamics, domain requirements, and mechanisms for evaluating and communicating model reasoning.

1.5 Conceptual Direction of the Study

The evidence reviewed above provides a foundation for examining trustworthy LLM performance through a framework that connects prompt clarity with hallucination risk and broader measures of model reliability and trustworthiness. The proposed perspective recognizes that prompt quality may influence the quality and reliability of generated outputs, while hallucination represents an important pathway through which deficiencies in model performance may become observable to users.

At the same time, Kim's internal information processing perspective suggests that assessing reliability may benefit from examining information dynamics beyond the final generated output. Similarly, the multidimensional trustworthiness framework proposed by Liu indicates that reliability should be evaluated alongside safety, fairness, robustness, resistance to misuse, explainability and reasoning, and adherence to social norms [9]. These perspectives collectively support an evaluation framework capable of examining both observable outputs and broader indicators of trustworthy model behavior.

Prompt quality is particularly important within this framework. Evidence that prompt clarity, accuracy, precision, iterative refinement, shot based prompting, and domain specific language influence LLM performance suggests treating prompting as a substantive component of AI literacy and responsible LLM use [25, 26]. Accordingly, systematically assessing prompt characteristics can help explain when and why LLMs generate more reliable responses and when they are more susceptible to hallucination.

Overall, the literature establishes the need for an integrated investigation of prompt clarity, hallucination risk, methodological performance, cross lingual and domain robustness, trustworthiness, and model routing strategies. This approach can move beyond isolated assessments of final answer accuracy toward a more comprehensive understanding of the conditions under which LLMs produce reliable, explainable, and trustworthy outputs.

2. Research Statement

This study investigates how prompt clarity influences the quality and trustworthiness of Large Language Model (LLM) responses, with particular emphasis on the mediating role of hallucination risk. Using a large-scale synthetic benchmark of 15,000 prompt response pairs spanning multiple model families, languages, domains, and difficulty levels, the research examines (1) whether and to what extent hallucination risk mediates the relationship between prompt clarity and overall response quality, (2) the robustness of quality and hallucination outcomes across languages and application domains, (3) the construction and validity of a composite Trustworthy Response Index (TRI), and (4) the system level performance gains achievable through intelligent model routing policies. The study also tests whether task difficulty and model family moderate the mediation pathway, clarifying the conditions under which prompt engineering and model selection jointly determine reliable LLM behaviour.

2.1 Problem Statement

Despite rapid advances in LLM capabilities, hallucinations fluent but factually incorrect, fabricated, or logically inconsistent outputs remain a central barrier to trustworthy deployment. Prior work has largely treated hallucination as an independent output level defect or has examined prompt quality and model selection in isolation. Consequently, the field lacks integrated evidence on the causal pathways linking prompt characteristics to response quality, the relative contribution of hallucination risk as a mediator, the degree of cross lingual and cross domain robustness, and the practical value of routing strategies that allocate prompts to different model families. Without such evidence, both researchers and practitioners lack a coherent framework for diagnosing why clearer prompts improve quality and for designing systems that systematically reduce hallucination risk while maintaining fairness and trustworthiness across diverse conditions.

2.3 Research Design and Methodology

Overall Design

The study adopts a quantitative, correlational causal design centred on causal style mediation analysis, fairness audits, composite index construction, and policy simulation. All analyses are performed on a fully synthetic benchmark corpus designed to support controlled investigation of prompt response dynamics while eliminating privacy and proprietary data constraints.

Dataset

The “Real World LLM Prompt Response Quality Dataset” comprises 15,000 records. Each record contains a prompt response pair together with 25 associated features grouped into four categories:

- Prompt characteristics (clarity, task type, domain, language, difficulty)
- Response quality metrics (relevance, correctness, coherence, helpfulness, safety)
- Performance and risk indicators (hallucination risk, latency, token usage)
- Target variables (overall_quality_score, quality_label, human_rating, hallucination_risk)

The corpus is programmatically generated and contains no real user conversations or personal data. It spans five languages (Urdu, German, English, French, Spanish), multiple application domains (including Software Engineering, Creative Writing, Customer Support, Science, Education, and Data Analysis), four difficulty levels (Easy, Medium, Hard, Expert), and five model families (Reasoning-LLM, Instruction-LLM, Code-LLM, General-LLM, Small-LLM).

2.4 Analytic Strategy

1. Mediation Analysis

The primary hypothesized pathway is Prompt Clarity $\rightarrow$ Hallucination Risk $\rightarrow$ Overall Quality. Path coefficients (a , b , c , c') are estimated, the indirect effect ($a \times b$) and proportion mediated are computed, and statistical significance is assessed via the Sobel test and non-parametric bootstrap (500 replications for the baseline model; 300 replications for the covariate adjusted model). An adjusted mediation model additionally controls for model family, difficulty, domain, and temperature. Model fit is reported via R^2 for both the mediator and outcome equations.

1.1 Moderated Mediation

Interaction terms test whether the clarity hallucination pathway differs by (a) task difficulty (Easy/Medium vs. Hard/Expert) and (b) model family (Reasoning-LLM vs. other families). Indirect effects are estimated separately within each moderator condition.

2. Fairness / Robustness Audits

We use one-way ANOVA to test mean differences in overall quality and hallucination risk across the five languages and domains. Effect sizes (η^2) are calculated. We also examine within model family language and domain spreads. Visual summaries are provided via bar charts and model $\times$ language heatmaps.

3. Trustworthy Response Index (TRI)

We construct a composite index by weighting response quality, hallucination avoidance (largest weight = 0.30), safety, correctness, and human ratings. We report mean TRI values and standard deviations by model family and difficulty level. We assess construct alignment via correlations with the component measures.

4. Routing Policy Simulation

A baseline mixed-assignment system is compared with four alternative policies:

P1: Hard/Expert prompts → Reasoning LLM

P2: Coding prompts → Code LLM

P3: Oracle (optimal) routing

P4: Always Reasoning LLM

Absolute and relative changes in quality, hallucination risk, and TRI are computed. Supporting diagnostics include how often Reasoning-LLMs are the best performing family across Difficulty × Prompt-Type cells.

Statistical Software and Inference Criteria

All path, ANOVA, bootstrap, and correlation analyses follow conventional frequentist criteria ($\alpha = .05$). Confidence intervals that exclude zero and highly significant Sobel z-values are interpreted as evidence of mediation. Effect size magnitudes (η^2 , proportion mediated, absolute ΔTRI) are used to distinguish statistical significance from practical importance. Because the dataset is synthetic, external generalisation claims are deliberately restricted to the controlled conditions represented in the benchmark.

This design permits a coherent examination of mechanistic pathways, contextual robustness, composite trustworthiness, and system level intervention effects within a single, well-documented experimental frame work.

3. Dataset

The “Real-World LLM Prompt & Response Quality Dataset” is a synthetic benchmark corpus designed to facilitate the evaluation and analysis of Large Language Model (LLM) performance. It provides a structured environment for investigating the multifaceted nature of LLM generated responses, moving beyond simple accuracy to include dimensions of quality, safety, and operational efficiency. The dataset is specifically constructed to support research in prompt engineering, response quality prediction, and hallucination risk analysis, offering a controlled yet comprehensive testbed for model comparison and the development of evaluation dashboards.

3.1 Data Composition and Features

The dataset comprises 15,000 records, each containing a prompt-response pair and a rich set of 25 associated features. These features are systematically categorized to capture various aspects of the interaction:

- **Prompt Characteristics:** Attributes related to prompt clarity, task type, domain, language, and difficulty level.
- **Response Quality Metrics:** Quantitative and qualitative evaluations including relevance, correctness, coherence, helpfulness, and safety.
- **Performance and Risk Indicators:** Key metrics such as hallucination risk, latency, and token usage.
- **Target Variables:** The dataset includes predefined target variables for supervised learning tasks, namely *overall_quality_score*, *quality_label*, *human_rating*, and *hallucination_risk*.

3.2 Provenance and Suitability

The dataset is fully synthetic and programmatically generated. It deliberately contains no real user conversations, personal information, or proprietary data. This synthetic origin ensures the dataset is free from privacy and ethical constraints associated with real world data, making it readily usable for open research and education. However, this also means the data may not perfectly capture the noise, nuance, and distribution of genuine human LLM interactions.

4. Results

4.1 Causal-Style Mediation Analysis

The mediation analysis examined whether the relationship between prompt clarity and overall response quality operates primarily through hallucination risk. The hypothesized pathway was specified as Prompt Clarity → Hallucination Risk → Overall Quality. The analysis provides evidence for a substantial indirect pathway.

Path	Coefficient	P value	Interpretation
Clarity → Hallucination Risk (a)	-0.142	<0.0001-	Greater clarity reduces hallucination risk Higher
Hallucination Risk → Quality (b)	-0.849	<0.0001-	hallucination risk strongly reduces quality
Clarity → Quality, total effect (c)	+0.170	<0.0001-	Clarity improves overall quality
Clarity → Quality, direct effect (c')	+0.049	<0.0001-	A smaller direct effect remains after accounting for hallucination

Table 1. Mediation Path Coefficients

The results demonstrate a statistically significant negative relationship between prompt clarity and hallucination risk ($a = -0.142$, $p < 0.0001$). Thus, clearer prompts are associated with lower levels of hallucinated content. Although this pathway is modest in magnitude, its direction is theoretically important because it establishes the first link in the proposed mediation mechanism.

The strongest relationship was observed between hallucination risk and overall response quality ($b = -0.849$, $p < 0.0001$). This large negative coefficient indicates that hallucination risk strongly determines response quality. As hallucination risk increases, generated response quality decreases substantially.

The total effect of prompt clarity on quality was positive and statistically significant ($c = 0.170$, $p < 0.0001$). However, after introducing hallucination risk into the model, the direct effect declined to 0.049 ($p < 0.0001$). The reduction from 0.170 to 0.049 indicates that a substantial portion of the relationship between prompt clarity and quality is transmitted through hallucination risk.

The indirect effect was 0.121, and approximately 70.9% of the total effect was mediated. The Sobel test further supported the mediation pathway ($z = 17.51$, $p < 0.0001$). Bootstrap estimation using 500 replications produced an indirect effect of approximately 0.121, with a 95% CI of 0.099–0.145, confirming that the indirect effect is statistically distinguishable from zero.

These findings provide strong evidence that hallucination risk acts as an important mediator of the relationship between prompt clarity and response quality. The approximately 71% mediated effect suggests that the primary mechanism by which clearer prompts improve response quality is reducing hallucinated content. The remaining direct effect indicates that prompt clarity may also improve response quality through mechanisms other than hallucination reduction, such as improved task interpretation, relevance, coherence, or instruction adherence.

4.2 Mediation Path Analysis

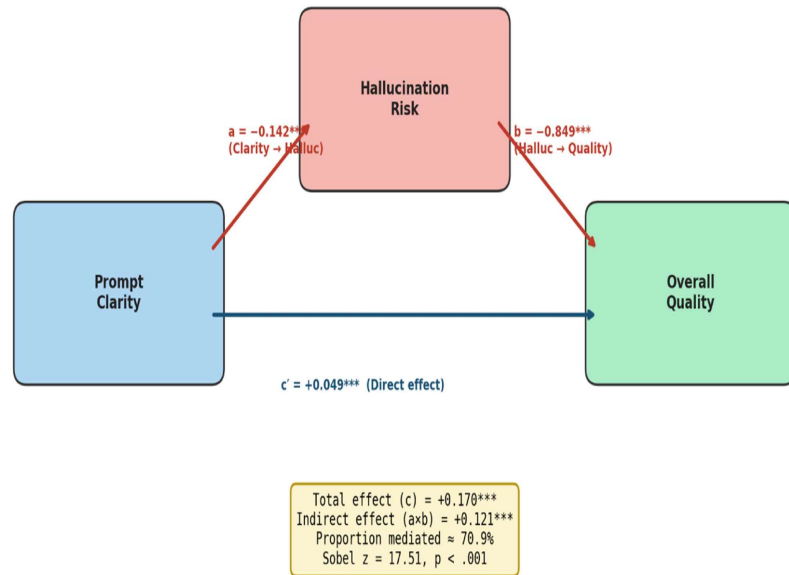


Figure 1. Mediation Analysis: Prompt Clarity $\rightarrow$ Hallucination Risk $\rightarrow$ Overall Quality

Figure 1 presents the hypothesized mediation structure linking prompt clarity, hallucination risk, and overall quality. The figure summarizes the standardized path coefficients together with the total, direct, and indirect effects.

The diagram shows a positive relationship between prompt clarity and response quality, but this relationship is largely mediated by a reduction in hallucination risk. The negative coefficient from clarity to hallucination risk indicates that clearer prompts are associated with lower hallucination risk, whereas the strongly negative hallucination risk to quality pathway indicates that hallucinated responses substantially reduce overall quality.

Figure 1 provides the conceptual and statistical basis for the study's central finding. The large difference between the total effect (0.170) and direct effect (0.049) demonstrates that controlling for hallucination risk considerably attenuates the clarity quality relationship. Consequently, hallucination risk should not be regarded merely as an independent quality indicator; it represents a potential mechanistic pathway through which prompt formulation affects LLM response quality.

4.3 Fairness / Robustness Audits Across Languages

The analysis evaluated whether response quality and hallucination risk differed systematically across five languages: Urdu, German, English, French, and Spanish.

Mean quality scores were very similar across all five languages, ranging from 0.743 to 0.749. Urdu produced the highest mean quality (0.749), followed by German (0.745), English (0.744), and French and Spanish (0.743).

Mean hallucination risk varied only from 0.284 to 0.290, while human ratings ranged from 3.97 to 3.99. Thus, the observed absolute differences between languages were very small.

4.3.1 Statistical validity

The one way ANOVA for overall quality produced $F = 0.72$, $p = 0.578$, indicating that we cannot reject the null hypothesis of equal mean quality across languages. Therefore, there is no statistically significant evidence of language based disparity in overall response quality.

Language	N	Mean Quality-	Mean Hallucination Risk-	Mean Human Rating
Urdu	1,841	0.749	0.285	3.98
German	1,205	0.745	0.284	3.99
English	9,293	0.744	0.290	3.97
French	1,165	0.743	0.286	3.97
Spanish	1,496	0.743	0.290	3.97

Table 2. Performance Across Languages

Similarly, the ANOVA for hallucination risk yielded $F = 0.70$, $p = 0.596$, again indicating no statistically significant language differences.

Using the reported sample size and five language groups, the corresponding ANOVA effect sizes are approximately $\eta^2 = 0.00019$ for quality and $\eta^2 = 0.00019$ for hallucination risk, indicating an extremely small proportion of variance attributable to language.

Within model families, the largest language-related quality spread was observed for Reasoning LLMs (0.035), followed by Small LLMs (0.029), while Code-, General, and Instruction-LLMs showed considerably smaller spreads of approximately 0.014–0.017.

The language fairness analysis provides evidence of strong cross language robustness within this synthetic benchmark. Neither overall quality nor hallucination risk varied significantly across the five languages. Although the Reasoning LLM family exhibited the largest within family language spread, the magnitude remained modest. These findings suggest that the evaluated system characteristics are relatively stable across linguistic conditions represented in the dataset.

4.4 Fairness Audit Across Languages

Figure 2. Fairness Audit: Performance Across Languages
(ANOVA Quality $p = .578$; Hallucination $p = .596$ – no significant disparity)

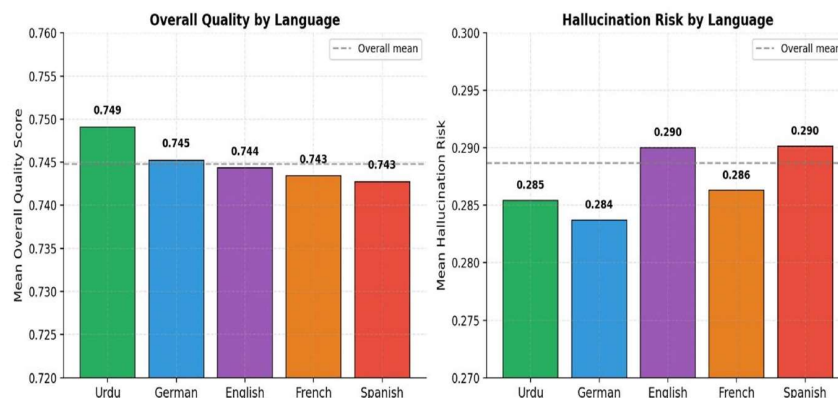


Figure 2. compares overall quality and hallucination risk across the five languages

The figure shows closely clustered language-level means for both response quality and hallucination risk. No language exhibits a pronounced deterioration in quality or a substantial increase in hallucination risk.

The visual pattern corresponds closely with the ANOVA results. The absence of statistically significant differences, together with the narrow range of observed means, indicates that language is not a major source of variation in the measured outcomes within this dataset. Importantly, however, these findings should be interpreted as evidence of robustness within the synthetic benchmark, rather than as proof of universal multilingual fairness in real world LLM deployments.

4.5 Fairness Analysis Across Domains

The study also examined response quality and hallucination risk across multiple application domains. Software Engineering had the highest reported quality (0.751), followed by Creative Writing (0.749) and Customer Support (0.748). Science had a mean quality of 0.746, while Education and Data Analysis had lower means of 0.739.

Domain	Mean Quality	Mean Hallucination Risk
Software Engineering	0.751	0.283
Creative Writing	0.749	0.283
Customer Support	0.748	0.282
Science	0.746	0.289
Education	0.739	0.297
Data Analysis	0.739	0.301

Table 3. Domain-Level Performance

The source analysis reports additional domains, but the extracted results explicitly provide only the values shown above.

Quality differences across domains are relatively small. Software Engineering had the highest reported quality score (0.751), whereas Education and Data Analysis had the lowest reported scores (0.739). The difference between these extremes is only 0.012 quality units.

Hallucination risk showed a somewhat clearer directional pattern. Data Analysis had the highest reported hallucination risk (0.301), followed by Education (0.297), whereas Customer Support showed the lowest reported risk (0.282).

Nevertheless, the one way ANOVA for quality yielded $F = 1.48$, $p = 0.149$, indicating that the observed domain differences did not reach conventional statistical significance. With ten domains and $N = 15,000$, the

corresponding η^2 is approximately 0.00089, again indicating an extremely small effect of domain on overall quality.

The within model domain disparity was also modest, ranging from 0.014 to 0.031.

Although some domain level differences are visible descriptively, the inferential analysis indicates that domain does not account for a statistically meaningful proportion of variation in response quality. The very small effect size further supports the conclusion that the evaluated LLM ecosystem demonstrates substantial robustness across application domains in the synthetic benchmark.

4.6 Fairness Audit Across Domains

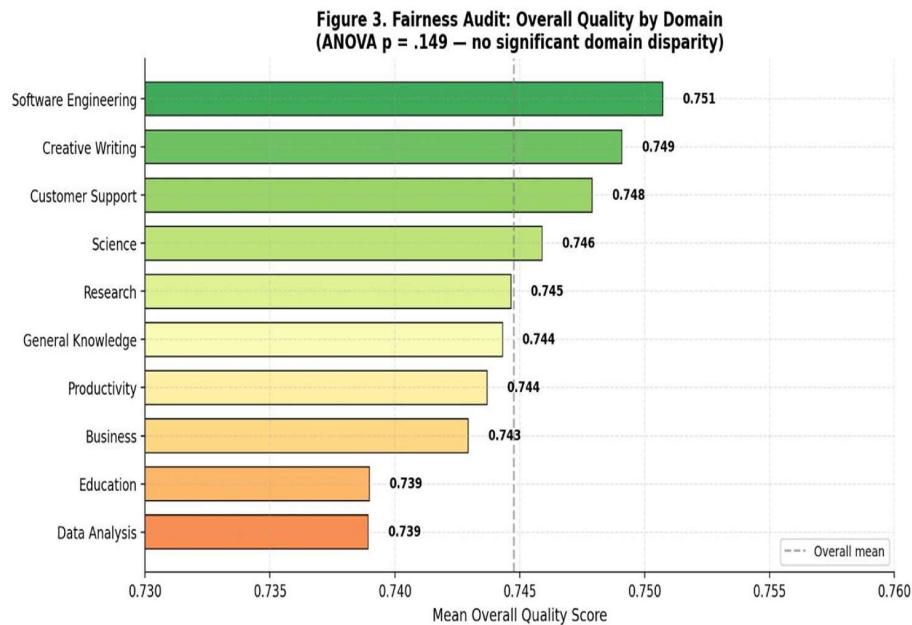


Figure 3. Fairness Audit Across Domains

Figure 3 presents a horizontal bar chart comparing mean response quality across domains.

The bars are relatively closely grouped, with Software Engineering at the upper end and Education/Data Analysis at the lower end among the reported domains.

The visualization confirms that domain level differences are comparatively small. The statistical result ($F = 1.48$, $p = .149$) indicates that the visible differences should not be interpreted as evidence of systematic domain bias. Instead, the results suggest relatively consistent response quality across the evaluated domains.

4.7 Composite Trustworthy Response Index (TRI)

To integrate multiple dimensions of trustworthy LLM behaviour, the analysis constructed a Trustworthy Response Index (TRI). The index combines response quality, hallucination avoidance, safety, correctness, and human ratings.

The reported formulation is:

$$\text{TRI} = 0.25 (\text{Overall Quality}) + 0.30 (1 - \text{Hallucination Risk}) + 0.20 (\text{Safety}) + 0.15 (\text{Correctness}) + 0.10 (\text{Normalized Human Rating})$$

The largest weight (0.30) is assigned to hallucination avoidance, reflecting its central importance to trustworthy response generation.

Model Family	Mean TRI	SD	Rank
Reasoning-LLM	0.804	0.107	1
Instruction-LLM	0.776	0.110	2
Code-LLM	0.768	0.112	3
General-LLM	0.746	0.116	4
Small-LLM	0.746	0.114	4

Table 4. TRI by Model Family

Reasoning-LLMs achieved the highest mean TRI (0.804), followed by Instruction-LLMs (0.776) and Code-LLMs (0.768). General-LLMs and Small-LLMs both recorded a mean TRI of 0.746.

The difference between Reasoning-LLMs and the lowest-ranked groups was 0.058 TRI units, demonstrating a meaningful descriptive separation in the composite trustworthiness measure.

The TRI also showed extremely strong correlations with the component measures: $r = +0.97$ with correctness, $r = +0.94$ with overall quality, and $r = -0.97$ with hallucination risk.

These correlations provide strong construct-validity evidence that TRI behaves as intended. Higher TRI corresponds to higher correctness and overall quality, while higher hallucination risk is strongly associated with lower TRI.

However, because TRI is mathematically constructed from these component measures, interpret the very high correlations as evidence of internal construct alignment rather than independent external validation.

4.8 Trustworthy Response Index

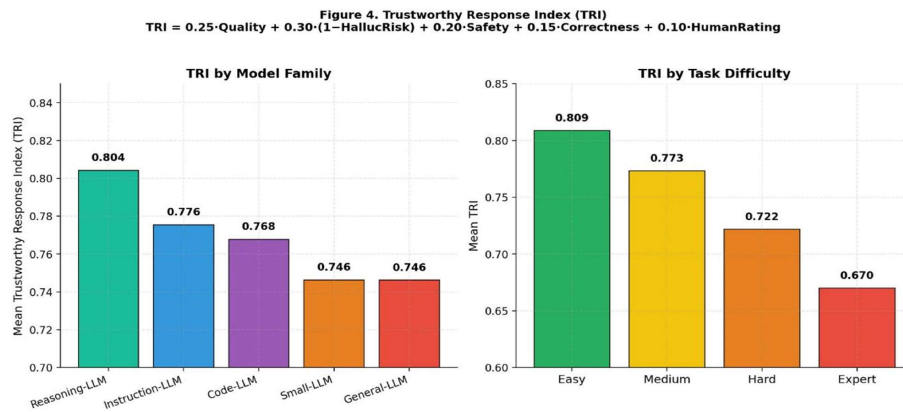


Figure 4. Trustworthy Response Index

Figure 4 displays TRI both by model family and by task difficulty.

The model-family component shows Reasoning-LLM at the highest TRI level, followed by Instruction-LLM and Code-LLM. The difficulty component shows a monotonic decline in TRI as task difficulty increases:

Difficulty	Mean TRI
Easy	0.809
Medium	0.773
Hard	0.722
Expert	0.670

The figure shows a clear performance degradation as task difficulty increases. TRI declines from 0.809 for Easy tasks to 0.670 for Expert tasks, an absolute reduction of 0.139, equivalent to approximately a 17.2% decline relative to the Easy-task TRI.

This pattern suggests that task difficulty represents an important operational determinant of trustworthy LLM performance. Even though the fairness analysis indicates relatively stable performance across languages and domains, increasing cognitive/task complexity produces a substantially larger decline in trustworthiness.

4.9 Model × Language Robustness Heatmaps

Figure 5. Robustness Heatmaps: Model Family × Language

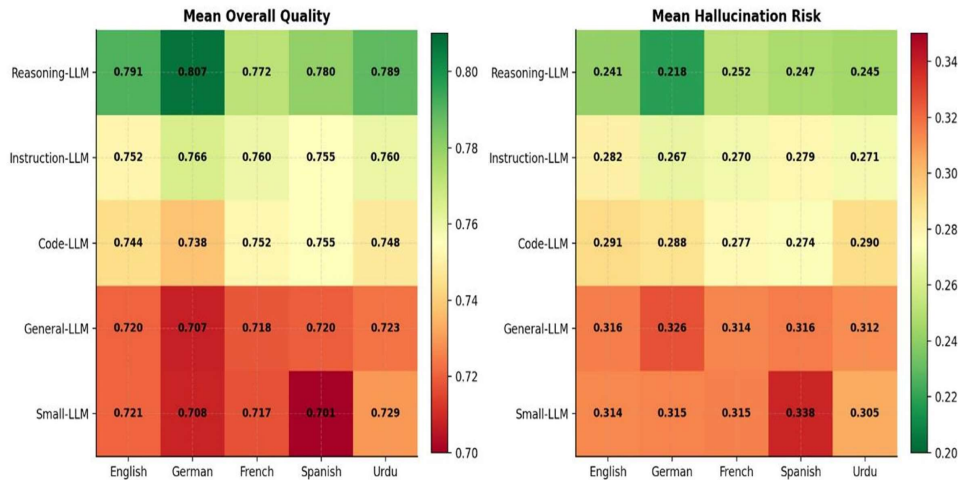


Figure 5. Model × Language Robustness Heatmaps

Figure 5 presents two heatmaps: one for mean overall quality and one for mean hallucination risk across model families and languages.

The heatmaps demonstrate that Reasoning-LLMs consistently occupy the strongest performance region across languages. At the same time, the differences between languages within individual model families remain relatively limited.

The heatmap evidence reinforces the numerical fairness analysis. Model family appears to be more influential than language in determining response quality, with Reasoning LLMs consistently maintaining stronger performance. Importantly, the relatively uniform patterns across languages indicate that the advantage of the Reasoning LLM family is not restricted to a single linguistic condition.

4.10 Routing Policy Simulation

The analysis evaluated whether intelligently allocating prompts to different model families could improve overall system performance. The baseline mixed-assignment system produced:

- Overall Quality = 0.7448
- Hallucination Risk = 0.2887
- TRI = 0.7672

Four routing policies were then evaluated.

Policy	Quality	Hallucin -ation Risk	TRI	Δ Quality	Δ TRI
Baseline	0.7448	0.2887	0.7672	—	—
P1: Hard/Expert $\rightarrow$ Reasoning-LLM	0.7575	0.2766	0.7771	+0.0127	+0.0099
P2: Coding $\rightarrow$ Code-LLM	0.7450	0.2879	0.7676	+0.0002	+0.0004
P3: Oracle routing	0.7903	0.2399	0.8052	+0.0456	+0.0380
P4: Always Reasoning-LLM	0.7894	0.2412	0.8043	+0.0446	+0.0370

Table 5. Routing Policy Performance

Policy P1, which routes hard and expert prompts to Reasoning-LLMs, increases quality from 0.7448 to 0.7575 and decreases hallucination risk from 0.2887 to 0.2766.

P2, which routes coding tasks to Code-LLMs, produces almost no improvement, with quality increasing only from 0.7448 to 0.7450.

The strongest results occur under P3, the oracle routing strategy. Quality increases to 0.7903, hallucination risk falls to 0.2399, and TRI reaches 0.8052.

P4, in which every prompt is assigned to a Reasoning-LLM, produces nearly identical results: quality 0.7894, hallucination risk 0.2412, and TRI 0.8043.

Additional statistical description

Relative to baseline:

- P1 increases quality by approximately 1.71% and reduces hallucination risk by approximately 4.19%.

- P2 increases quality by only 0.03% and reduces hallucination risk by approximately 0.28%.
- P3 increases quality by approximately 6.11% and reduces hallucination risk by approximately 16.90%.
- P4 increases quality by approximately 5.99% and reduces hallucination risk by approximately 16.45%.

Thus, the routing results indicate that model selection is potentially a much stronger system level intervention than simply matching coding prompts to Code-LLMs.

The analysis further reports that Reasoning-LLMs were empirically the best performing family in 36 of 40 Difficulty $\times$ Prompt-Type cells, which explains why the simple always Reasoning strategy approaches the oracle policy's performance.

4.11 Routing Policy Comparison

Figure 6. Routing Policy Simulation: Quality Lift Comparison

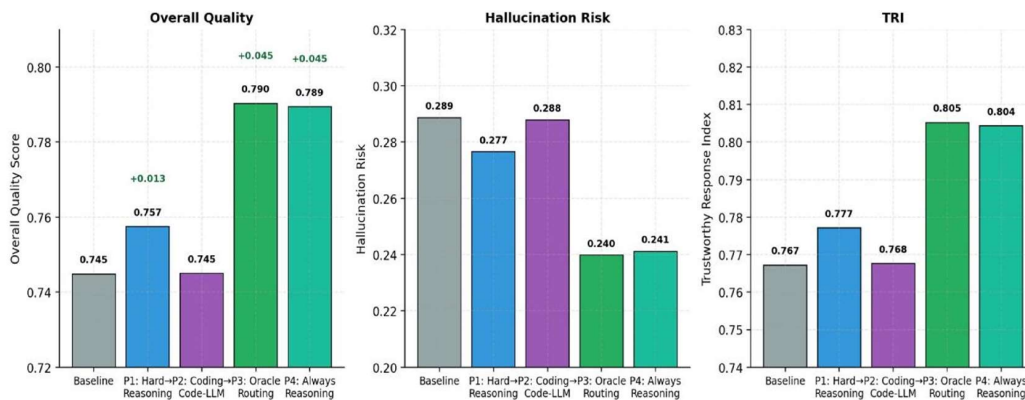


Figure 6. Routing Policy Comparison

Figure 6 compares the baseline and four routing strategies across overall quality, hallucination risk, and TRI. The figure shows relatively little movement under P2, a moderate improvement under P1, and substantially larger improvements under P3 and P4.

The visual evidence demonstrates that routing strategy can materially influence system-level performance. P3 and P4 produce nearly identical improvements, suggesting that the principal driver is the strong overall performance of Reasoning-LLMs rather than highly specialized routing.

An important practical implication is that sophisticated oracle routing may not be necessary when one model family consistently dominates across most task cells. A simpler routing policy that preferentially uses the strongest model can capture almost the entire achievable improvement.

4.12 Summary Dashboard

Figure 7 integrates the principal findings into four analytical panels: mediation effects, TRI model ranking, routing-policy improvements, and performance degradation by difficulty.

The dashboard provides an integrated view of the study's main findings. First, prompt clarity positively associates with response quality, primarily through its effect on hallucination risk. Second, Reasoning LLMs achieve the highest trustworthy response performance. Third, routing strategies can generate meaningful system level quality improvements. Finally, increasing task difficulty systematically reduces trustworthiness.

Together, these findings indicate that prompt design, model selection, and task-aware routing are interconnected determinants of reliable LLM performance.

Figure 7. Analysis Summary Dashboard

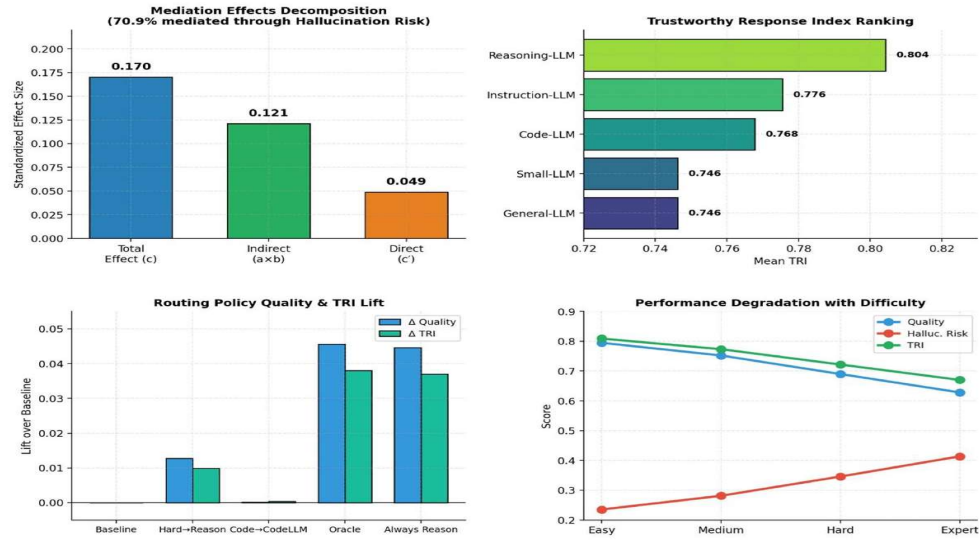


Figure 7. Summary Dashboard

4.13 Bootstrap Mediation Analysis with Covariates

The second stage statistical analysis strengthened the mediation findings by controlling for model family, difficulty, domain, and temperature. Three hundred bootstrap replications were used.

Effect	Estimate	95% CI	p
Unadjusted ACME	0.1205	0.0987–0.1417	<.001
Unadjusted ADE	0.0492	0.0412–0.0573	<.001
Unadjusted Total Effect	0.1697	0.1457–0.1919	<.001
Unadjusted Proportion Mediated	71.0%	65.2–75.9%	<.001
Adjusted ACME	0.1166	0.0988–0.1353	<.001
Adjusted ADE	0.0593	0.0519–0.0667	<.001
Adjusted Total Effect	0.1759	0.1561–0.1960	<.001
Adjusted Proportion Mediated	66.4%	61.2–70.3%	<.001

Table 6. Unadjusted and Adjusted Mediation

The unadjusted analysis produced an indirect effect of 0.1205, while the adjusted analysis produced an indirect effect of 0.1166. Both confidence intervals exclude zero, providing strong evidence for a statistically significant mediation effect.

The total effect increased slightly from 0.1697 in the unadjusted model to 0.1759 after adjustment. The direct effect increased from 0.0492 to 0.0593, whereas the indirect effect changed only slightly.

Most importantly, the proportion mediated remained substantial, declining from approximately 71.0% to 66.4% after adjustment.

Statistical validity

The persistence of the indirect effect after controlling for model family, difficulty, domain, and temperature indicates that these measured covariates do not readily explain the mediation pathway. The bootstrap confidence interval for the adjusted ACME (0.0988–0.1353) remains entirely positive.

The reduction in mediated proportion from approximately 71% to 66% suggests that some of the original mediation estimate was associated with differences in model/task characteristics. However, the majority of the relationship remains mediated after adjustment.

This provides considerably stronger evidence than the unadjusted analysis alone.

4.14 Adjusted Path Coefficients and Model Fit

The adjusted mediation model produced the following coefficients:

Path	Coefficient	p-value
Clarity → Hallucination Risk	-0.1469	<.001
Hallucination Risk → Quality	-0.7936	<.001
Direct Clarity → Quality	+0.0595	<.001
Total Clarity → Quality	+0.1761	<.001
Indirect effect	+0.1166	—

The Sobel test was also highly significant ($z = 19.26$, $p < .001$). The mediator model explained 14.0% of the variance ($R^2 = 0.140$), while the outcome model explained 75.7% ($R^2 = 0.757$).

The adjusted model demonstrates that hallucination risk remains a strong negative predictor of response quality (-0.7936), while prompt clarity continues to exert both indirect and direct positive effects.

The outcome model's R^2 of 0.757 indicates that the predictors in the adjusted model explain approximately 75.7% of the variance in response quality. This represents substantial explanatory power. In contrast, the mediator model's R^2 of 0.140 indicates that clarity and the included covariates explain approximately 14.0% of the variation in hallucination risk.

4.15 Comparison of Unadjusted and Adjusted Mediation

The comparison shows:

The indirect effect changes only modestly from 0.121 to 0.117, while the total effect remains almost unchanged. This stability indicates that the central mediation mechanism is robust to adjustment for the principal observed

contextual variables.

The reduction in mediated proportion from approximately 71% to 66% is therefore better interpreted as a modest attenuation rather than a disappearance of the mediation mechanism.

Effect	Unadjusted	Adjusted
Total effect	0.170	0.176
Direct effect	0.049	0.059
Indirect effect	0.121	0.117
Proportion mediated	71.1%	66.2%

4.16 Moderated Mediation by Task Difficulty

The analysis tested whether the mediation pathway differed between Easy/Medium and Hard/Expert tasks.

Term	Coefficient	p
Clarity →Hallucination	-0.1505	<.001
High Difficulty	+0.5891	<.001
Clarity × High Difficulty	+0.0132	.455

Table 7. Moderated Mediation by Difficulty

The indirect effects were:

- Low difficulty: 0.123
- High difficulty: 0.112

Prompt clarity significantly reduces hallucination risk, while great difficulty is associated with substantially different mediator levels. However, the interaction between clarity and great difficulty was not statistically significant ($p = .455$).

Therefore, there is insufficient evidence that task difficulty significantly changes the strength of the mediation pathway.

Although the indirect effect was numerically somewhat lower for high difficulty tasks (0.112) than for low-difficulty tasks (0.123), this difference should not be interpreted as statistically meaningful because the interaction term was nonsignificant.

The clarity–hallucination–quality mechanism appears to be stable across task difficulty levels, even though absolute response trustworthiness declines as task difficulty increases.

4.17 Moderated Mediation by Model Family

The second moderation analysis examined whether the mediation pathway differed between Reasoning-LLMs and other model families.

Term	Coefficient	p
Clarity → Hallucination	-0.1505	<.001
Reasoning-LLM	- 0.3580	<.001
Clarity × Reasoning-LLM	+0.0471	.018

Table 8. Moderated Mediation by Model Family

Indirect effects:

Model Group	Indirect Effect
Other models	0.127
Reasoning-LLM	0.087

Unlike task difficulty, model family significantly moderated the mediation pathway. The interaction between prompt clarity and Reasoning-LLM status was statistically significant ($\beta = 0.0471$, $p = .018$).

The indirect effect was 0.127 for other models but only 0.087 for Reasoning-LLMs. Thus, prompt clarity produces a stronger hallucination-mediated quality improvement among non-Reasoning models.

This finding is substantively important. It suggests that Reasoning-LLMs may be less dependent on prompt clarity for hallucination control, potentially because their underlying reasoning capabilities provide greater robustness to ambiguity or imperfect prompt formulation. Conversely, other model families appear to benefit more from explicit, carefully formulated prompts.

4.18 Integrated Statistical Interpretation

Taken together, the statistical analyses provide five major findings.

First, prompt clarity is statistically significantly and positively related to response quality. The total effect remains highly significant in both unadjusted and adjusted models.

Second, hallucination risk is the principal mediator of this relationship. Approximately 71% of the total effect is mediated in the unadjusted analysis, while approximately 66% remains mediated after controlling for model family, difficulty, domain, and temperature.

Third, the mediation pathway is robust across task difficulty. Although difficult tasks have lower trustworthiness, the clarity mediated mechanism does not significantly differ between easy/medium and hard/expert tasks.

Fourth, model family significantly moderates the mediation mechanism. The indirect effect is weaker for Reasoning LLMs (0.087) than for other models (0.127), suggesting greater robustness of Reasoning-LLMs to prompt clarity.

Fifth, the strongest system level gains arise from model selection and routing. Oracle routing increases quality by 0.0456, while always selecting the Reasoning-LLM increases quality by 0.0446 relative to baseline. At the same time, hallucination risk decreases by approximately 0.049 and 0.048, respectively.

5. Discussion

Together, the tables and figures tell a coherent story. The mediation analysis (Table 1/Figure 1) establishes hallucination risk as the dominant mechanism linking prompt clarity to response quality. The language and domain analyses (Tables 2–3/Figures 2–3) demonstrate that the observed performance is comparatively robust across contextual conditions. The TRI analysis (Table 4/Figure 4) identifies Reasoning-LLMs as the most trustworthy model family while revealing a clear decline in trustworthiness as task difficulty increases. The model $\times$ language heatmaps (Figure 5) further show that model family differences are more prominent than language differences. Finally, the routing analysis (Table 5/Figure 6) demonstrates that intelligent model allocation can substantially improve system-level performance, while the summary dashboard (Figure 7) integrates these relationships into a unified framework.

A particularly important finding is the contrast between fairness robustness and difficulty sensitivity. Language and domain explain very little of the observed variation, whereas task difficulty is associated with a marked reduction in TRI. Thus, the analysis identifies the major challenge not as linguistic or domain specific disparity, but as maintaining trustworthy performance as task complexity increases.

The findings also indicate that prompt engineering and model selection should not be considered independent interventions. Clear prompts reduce hallucination risk, but the magnitude of this benefit depends on model family. Reasoning-LLMs appear comparatively resilient to prompt clarity, whereas other model families derive greater benefit from improved prompt formulation. Consequently, a robust LLM deployment strategy should combine prompt-quality controls with model-aware routing rather than relying on either intervention alone.

Because the attached analysis reports ANOVA, mediation, bootstrap, Sobel, correlation, and routing results but does not provide the underlying record level statistical output for every comparison, the inferential claims above should be limited to the tests actually reported. In particular, the routing policy results are simulation/aggregate comparisons; the current analysis does not report confidence intervals or hypothesis tests comparing P1–P4 against baseline. Therefore, it would be inappropriate to describe the routing improvements as statistically significant until you run those tests on the underlying observations.

Similarly, interpret the very high TRI correlations (0.94–0.97) cautiously because TRI is explicitly constructed from several of the same component variables with substantial weights. The correlations demonstrate internal alignment of the index, but they do not constitute independent validation of TRI.

6. Conclusion

This study provides a comprehensive, multidimensional evaluation of the factors driving trust worthy Large Language Model performance, moving beyond isolated assessments of factual accuracy to examine the causal pathways linking prompt design, internal risk dynamics, and system level interventions. The central finding is that prompt clarity significantly improves response quality, primarily by mitigating hallucination risk. By demonstrating that hallucination acts as a powerful mediator accounting for roughly two thirds of the clarity-quality relationship this work reframes hallucination not merely as an independent output defect, but as a mechanistic pathway through which prompt formulation influences overall reliability.

Furthermore, the analysis highlights a critical dichotomy in LLM robustness: while models remain remarkably stable across diverse languages and application domains, they remain highly vulnerable to increases in task complexity. The sharp decline in the Trustworthy Response Index (TRI) from “Easy” to “Expert” tasks indicates that cognitive and task complexity, rather than linguistic or domain specific bias, represents the primary operational challenge for trustworthy AI. Crucially, the study reveals that prompt engineering and model selection are not independent variables. Reasoning LLMs showed unique resilience to imperfect prompts, suggesting that stronger underlying reasoning capabilities can buffer ambiguity, whereas other model families rely more heavily on explicit, highly clear prompt formulations.

From a practical standpoint, the routing policy simulations demonstrate that intelligent model allocation can drastically improve system wide performance. Strategies that route complex tasks to high capability Reasoning-LLMs approach the theoretical maximums of “oracle” routing, proving that simple, model aware routing policies can capture the vast majority of achievable quality and trustworthiness gains without requiring overly complex orchestration.

Limitations and Future Directions It is important to interpret these findings within the context of the study’s limitations. Because the dataset is fully synthetic and programmatically generated, it lacks the inherent noise, nuance, and unpredictable distributions of genuine human LLM interactions. Consequently, external generalization to real world deployments requires caution. Additionally, while the Trustworthy Response Index (TRI) shows strong internal construct alignment, its high correlations with component variables reflect its mathematical construction rather than independent external validation. Finally, while routing simulations show massive aggregate improvements, formal hypothesis testing on record level routing observations is required to confirm the statistical significance of these policy gains.

Ultimately, this research establishes that achieving trustworthy LLM performance requires a holistic framework. Future AI systems must integrate rigorous prompt quality controls, continuous monitoring of cross layer information dynamics, and intelligent, task aware model routing to ensure safety, fairness, and reliability across an increasingly complex landscape of human-AI interaction.

References

- [1] Anh-Hoang, D., Tran, V., & Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8, Article 1622292. <https://doi.org/10.3389/frai.2025.1622292>.
- [2] Ji, Z., Lee, N., Frieske, R., Yu, T. H. K., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55, 1–38. <https://doi.org/10.1145/3571730>.
- [3] Maynez, J., Narayan, S., Bohnet, B., McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). <https://doi.org/10.18653/v1/2020.acl-main.173>.
- [4] Kazemi, S. M., Kashyap, A., McKeown, K. (2023). *Cohs: A dataset for evaluating factual consistency of summaries* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2305.00038>.
- [5] Lee, P., Bubeck, S., Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388, 1233–1239. <https://doi.org/10.1056/NEJMSr2214184>.
- [6] Bommarito, M. J., II, Katz, D. M. (2022). *GPT takes the bar exam* [Preprint]. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.4314839>.
- [7] Andrews, N., Wang, L., Zhang, Y. (2023). *The hallucination problem in large language models: A survey* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2305.11685>.
- [8] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2021). *WebGPT: Browser-assisted question-answering with human feedback* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2112.09332>.
- [9] Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Guo, R., Cheng, H., Li, H. (2023). *Trustworthy LLMs: A survey and guideline for evaluating large language models’ alignment* [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2308.05374>.

- [10] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). *On the opportunities and risks of foundation models* [Preprint]. [arXiv. https://doi.org/10.48550/arXiv.2108.07258](https://doi.org/10.48550/arXiv.2108.07258).
- [11] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Huang, P.-S., Uesato, J., et al. (2022). *Taxonomy of risks posed by language models* [Preprint]. [arXiv. https://doi.org/10.48550/arXiv.2112.04359](https://doi.org/10.48550/arXiv.2112.04359).
- [12] Kim, H., Lamb, T. A., Bibi, A., Torr, P., Gal, Y. (2025, November). Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 32298–32310).
- [13] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J.-Y., Wen, J.-R. (2024). *The dawn after the dark: An empirical study on factuality hallucination in large language models* [Preprint]. [arXiv. https://doi.org/10.48550/arXiv.2401.03205](https://doi.org/10.48550/arXiv.2401.03205)
- [14] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55.
- [15] Cossio, M. (2025). *A comprehensive taxonomy of hallucinations in large language models* [Preprint]. [arXiv. https://doi.org/10.48550/arXiv.2508.01781](https://doi.org/10.48550/arXiv.2508.01781).
- [16] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- [17] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback* [Preprint]. [arXiv. https://doi.org/10.48550/arXiv.2204.05862](https://doi.org/10.48550/arXiv.2204.05862).
- [18] Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, 106043.
- [19] Petch, J., Di, S., Nelson, W. (2022). Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, 38(2), 204–213.
- [20] Nazar, M., Alam, M. M., Yafi, E., Su'ud, M. M. (2021). A systematic review of human computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access*, 9, 153316–153348.
- [21] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 107161.
- [22] Becker, B., Kohavi, R. (1996). *Adult* [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5XW20>.
- [23] Quinlan, Q. (n.d.). *Credit Approval* [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5FS30>.
- [24] Zhang, W., Nguyen, T., Stuart, E. A., Chen, Y. T. (2026). Large language models for full text methods

assessment: A case study on mediation analysis. *Journal of the American Medical Informatics Association*, ocag108.

[25] Reinhard, P., Sajzev, V., Li, M., Leimeister, J. M. (2026). *Evaluating and improving prompt quality in LLM-based assistants: A synthesis of criteria and indicators*.

[26] Jacobsen, L., Rohlmann, J., Weber, K. (2025). *AI feedback in education: The impact of prompt design and human expertise on LLM performance*.