Useful to Whom? A Persona-Driven Evaluation of Knowledge-Adapted Health Question Reformulation via LLM Simulation

Jooyeon Lee, Luan Huy Pham, Özlem Uzuner
George Mason University
4400 University Dr, Fairfax, Virginia, United States of America
{jlee252, lpham6, ouzuner}@gmu.edu

Abstract

Automatic metrics such as F1 and BERTScore are often insufficient for evaluating user-centric generative tasks like Consumer Health Question (CHQ) reformulation. A high F1-score may not correlate with user satisfaction, especially when the user's knowledge level (UKL) dictates their needs. We propose a robust, Persona-Driven Evaluation Framework (PDEF), grounded in cognitive science and health literacy literature, to measure persona-specific UCQ. This framework assesses reformulations from the perspectives of a 'Layperson' (requiring foundational context) and an 'Expert' (requiring efficient, precise answers). We apply this framework to a set of reformulated questions generated by LLMs, and test the robustness of our evaluation by using three state-of-the-art LLMs (GPT-4o, Llama 3.3, and Mistral Large) as the evaluators. Our results reveal a significant disconnect between automatic metrics and user-perceived quality: the model with the highest F1-score (0.6134) was *consistently outperformed* in user preference by a Pipelined model, with experts preferring the latter by a statistically significant margin ($p < 0.001$). Furthermore, our persona-driven ablation analysis provides robust evidence that specific architectural components, specifically UKL inference and Entailment logic, are linked to significant gains in persona-driven quality for Layperson cohorts. This work demonstrates the critical need for user-centric evaluation and shows that its findings are generalizable across different LLM architectures.

Keywords: Consumer Health, Question Understanding, LLM-as-a-Judge, Persona-Driven Evaluation

1. Introduction

Consumers increasingly turn to the internet for health information, but are often hindered by a "vocabulary gap" (Zeng et al., 2005; Zeng and Tse, 2006) between their layperson queries and the medical jargon of online resources. Consumer Health Question (CHQ) reformulation aims to bridge this gap by converting ambiguous queries into a set of canonical, answerable questions.

However, evaluating these generative systems, specifically Large Language Model (LLM) architectures designed for query reformulation, is notoriously difficult. While these systems excel at producing fluent text, traditional automatic metrics based on semantic similarity, such as F1-scores, often fail to measure the actual utility of the output for specific user cohorts (Zhang et al., 2020). For instance, a reformulation that is verbose and repetitive might achieve a high F1-score via keyword matching, yet frustrate an expert user seeking a concise, evidence-based data point.

We argue that a "one-size-fits-all" reformulation is suboptimal because it ignores the divergent cognitive needs of users. To address this, we formalize user requirements into two distinct profiles based on domain familiarity and information-seeking goals (Brusilovsky, 2001; Hölscher and Strube, 2000):

Novice (Low UKL): A user who lacks a foundational understanding of the specific medical domain, often characterized by "Low Health Literacy" or limited exposure to technical medical silos (White et al., 2008). This typically includes newly diagnosed patients or concerned caregivers.

- Information-Seeking Goal: Orientation and Scaffolding. As established by Chiu et al. (Chiu et al., 2022), laypersons with fewer clinical consultations often struggle with match rates between their queries and professional health information. They require foundational context to bridge the "vocabulary gap" and effectively interpret technical answers (Kuhlthau, 1991).
- Behavioral Indicator: Novices derive a Shared Benefit from query normalization but specifically require Augmentation. Unlike experts who prioritize simple summarization, novices prefer reformulations that move beyond paraphrasing toward Entailment, the addition of broader, prerequisite questions that provide the necessary background to understand the primary query (Chen et al., 2021; Vakkari, 1999).

Expert (High UKL): a user with high domain-specific knowledge, such as a clinician, medical researcher, or a "pro-patient" with extensive lived

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 281–295
12 May 2026.

experience, who has developed high search proficiency through frequent medical consultations (Chiu et al., 2022). These users are characterized by their ability to navigate specialized technical resources like PubMed (National Center for Biotechnology Information (NCBI), 2026) to extract evidence-based data (White et al., 2008; Research, 2025).

- Information-Seeking Goal: Precision and Extraction. Experts seek specific, clinically relevant data points to fill narrow information gaps within an existing, robust mental model (Chen et al., 2021).
- Behavioral Indicator: They prioritize Normalization (de-noising complex narrative queries) but actively reject Augmentation. Due to the Expertise Reversal Effect (Sweller. et al., 2011; Kalyuga et al., 2003), they perceive foundational "scaffolding" (e.g., definitions) as extraneous cognitive load or "noise" that hinders their retrieval efficiency.

To evaluate these divergent needs, we define User-Centric Quality (UCQ) as the degree to which a reformulated query aligns with the specific cognitive constraints and information-seeking goals of a target persona. Grounded in Saracevic's theory of situational relevance (Saracevic, 1975) and Sweller's Expertise Reversal Effect (Sweller. et al., 2011), this metric prioritizes the functional effectiveness of the output over simple string matching. High UCQ ensures that a novice receives necessary scaffolding while an expert is provided with a concise, high-efficiency path to information.

This theoretical conflict, where the same information serves as a bridge for one user and a barrier for another, necessitates a Persona-Driven Evaluation Framework (PDEF). To quantify the alignment between reformulation strategies and these opposing needs, we established four persona-driven evaluation criteria:

1. **Goal Achievement:** How well the output helps the persona achieve their primary goal (orientation vs. efficiency) (Kuhlthau, 1991; Hölscher and Strube, 2000).
2. **Cognitive Appropriateness:** How well the level of context (added or removed) matches the persona's needs (Sweller. et al., 2011; Norman, 1987).
3. **Clarity & Usability:** How easy the output is to understand and use (iso, 2018; Zarcadoolas et al., 2006).
4. **Perceived Relevance & Utility:** How confident the persona is that the output will lead to useful answers for their specific goal (Saracevic, 1975; Vuong et al., 2019).

Model	Description
Baselines	
B1 (Summ.)	Summarizes the query.
B2 (QR Def.)	Given a simple definition of reformulation.
Context-Aware Frameworks	
SPM (Single-prompt)	Infers UKL, Focus and Type and applies entailment logic.
PM (Multi-prompts)	Infers UKL, Focus and Type with separate prompts, then uses those for input for the prompt applying entailment logic.
Ablation Models (SPM)	
SPM (-UKL)	Removes UKL inference and UKL usage from entailment logic.
SPM (-Entail.)	Removes Entailment logic.

Table 1: Reformulation models and baseline strategies sourced directly from (Lee et al., 2026). We utilize these existing model outputs to validate our PDEF against both context-agnostic and adaptive architectures.

By moving beyond traditional semantic overlap (F1-score) to these theory-grounded metrics, we can precisely identify where current LLMs succeed or fail in meeting subjective user requirements.

In this work, we hypothesize that the most effective reformulation is tailored to the UKL. We test this hypothesis by introducing a PDEF, which we apply to a series of generative systems, detailed in Table 1, originally developed in our prior work (Lee et al., 2026).

These systems represent a spectrum of reformulation strategies: Baselines provide context-agnostic outputs such as simple summarization (B1) or a standard definition of the reformulation task (B2); Context-Aware Frameworks with Single-Prompt Model (SPM) and Pipeline-Model (PM) utilize GPT-4 (OpenAI, 2023) to adaptively infer UKL and apply entailment logic, adding foundational context; and Ablation Models allow us to isolate the specific impact of UKL inference and entailment. By evaluating these diverse strategies through our PDEF, we can quantify which model components drive satisfaction for different types of users.

Our contributions are threefold:

1. We introduce a persona-based evaluation framework grounded in health literacy (Kuhlthau, 1991; Zarcadoolas et al., 2006) and cognitive load theory (Sweller. et al., 2011). This framework is designed to quantify the alignment of reformulations with the distinct information-seeking goals and cognitive constraints of user cohorts by implementing user

needs into measurable persona-driven criteria.

2. We validate our prior F1-score findings (Lee et al., 2026) with a user-centric lens, demonstrating that our persona-based evaluation also finds the Context-Aware Frameworks to be dramatically superior to B1, B2, SPM (-UKL) and SPM (-Entail).

3. We validate the robustness of our framework and findings by running our evaluation on three state-of-the-art LLMs (GPT-4o, Llama 3.3, and Mistral Large), showing that our core conclusions are generalizable and not an artifact of a single model.

2. Related Work

Consumer Health Question (CHQ) Simplification Bridging the "vocabulary gap" between laypeople and medical professionals is a foundational challenge in medical informatics (Zeng et al., 2005; Zeng and Tse, 2006). Early approaches relied on dictionary-based substitution and statistical machine translation to simplify jargon. With the advent of neural generation, recent work has shifted toward abstractive reformulation and summarization (Ben Abacha and Demner-Fushman, 2019; Lee et al., 2026). However, the majority of existing research focuses exclusively on *simplification*, making text readable for low-literacy users. Our work diverges by acknowledging that "simplification" is often detrimental to high-knowledge users (experts), necessitating an adaptive approach that can toggle between scaffolding and precision.

User Knowledge Modeling & Cognitive Load The distinction between novice and expert information needs is a foundational challenge in information retrieval. For nearly two decades, research has shown that domain expertise fundamentally dictates search interaction patterns. This was pioneered by White et al. (White et al., 2008), who utilized specialized resource visitation (e.g., PubMed) to distinguish between medical experts and non-experts (Novices), finding that experts prioritize technical precision over general orientation.

This trajectory continues into the era of generative AI; recent large-scale telemetry studies from Microsoft Research (Research, 2025) confirm that these expertise levels remain the primary drivers of user satisfaction in chat-based interfaces. Specifically, while novices utilize generative tools for information recall and mental model building (Kuhlthau, 1991), experts demand high-complexity, targeted extraction. Our work implements these established behaviors by testing the Expertise Reversal Effect (Sweller. et al., 2011; Kalyuga et al., 2003), a phenomenon where the foundational "scaffolding" essential for a novice's understanding becomes extraneous cognitive load that hinders an expert's efficiency.

Evaluation of Generative Reformulation Evaluating generative text remains an open challenge. Traditional n-gram metrics (BLEU, ROUGE) and embedding-based metrics (BERTScore) (Zhang et al., 2020) measure semantic overlap but fail to capture persona-specific requirements or the cognitive appropriateness of the output. Recently, the "LLM-as-a-Judge" paradigm (Zheng et al., 2023) has emerged as a cost-effective alternative to human annotation, correlating well with human preferences on general tasks. However, most LLM-based evaluations use a single, generic evaluator prompt. In this work, we propose a PDEF, using opposing evaluator personas to explicitly measure the trade-off between accessibility (for laypeople) and efficiency (for experts).

3. Methodology

3.1. Experimental Setup

To validate the PDEF, we utilize the Consumer Health Question (CHQ) dataset established in (Lee et al., 2026). This corpus consists of 273 complex, multi-intent queries sourced from the NIH Genetic and Rare Diseases (GARD) Information Center. Each query is paired with expert-curated reformulations that serve as the ground truth for traditional semantic evaluation.Our experiments are conducted on a controlled testbed of 1,638 reformulated outputs (273 queries $\times$ 6 systems). These outputs were generated by the six architectures described in the Introduction (Table 1), ranging from context-agnostic baselines to adaptive, context-aware frameworks. By using this fixed set of outputs, we isolate the PDEF's ability to measure persona-driven utility across varying levels of architectural complexity.

3.2. Automatic Evaluation Baseline

Following the methodology of our prior work (Lee et al., 2026), we first evaluated all models using a redundancy-aware semantic F1-score (RA-Sem-F1). This metric computes a set-based F1-score by matching generated questions against a gold-standard set using specialized medical embeddings: **BGE-Medical** (Hükmet, 2024) (0.67 similarity threshold) and **BioLORD** (Rémy et al., 2022) (0.90 threshold for deduplication).As shown in Table 2, these semantic metrics establish that the SPM is the top-performing model under traditional evaluation ($F1 = 0.6134$). This serves as the baseline context for our current study, allowing us to

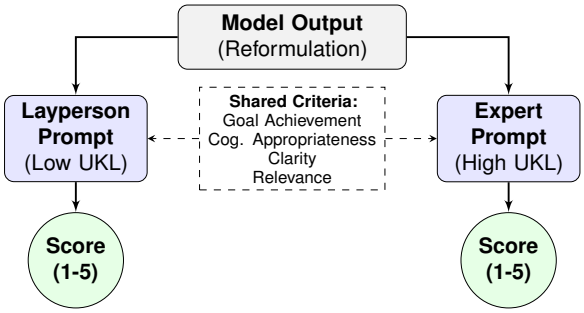

Figure 1: The PDEF. The model output is evaluated by two distinct **Persona Prompts** (Layperson vs. Expert). Both prompts utilize the same **Shared Criteria** to generate a evaluation score.

compare traditional performance against persona-specific utility.

3.3. PDEF Implementation

To strictly measure UCQ, we propose the PDEF, illustrated in Figure 1. We developed a structured evaluation pipeline using LLMs-as-judges (Zheng et al., 2023) to process the 1,638 reformulated outputs through the following procedure:

1. Prompt Construction: For every query-output pair, we constructed a specific evaluation prompt containing the original user query, the model's reformulation, and a detailed Persona Profile (Layperson or Expert).

2. Scoring Rubric: Each judge was provided with the four Persona-Driven Criteria established in the Introduction. The judges were instructed to provide a Likert score (1–5) for each criterion, followed by a brief natural language justification to ensure reasoning consistency.

3. Inference Parameters: To ensure stable and reproducible scoring, we set the sampling temperature to $\tau = 0.0$ for all three LLM judges (GPT-4o, Llama 3.3, and Mistral Large), minimizing stochastic variance in the evaluation results.

4. Data Aggregation: We collected a total of 9,828 individual evaluations (1,638 outputs $\times$ 2 personas $\times$ 3 judges). These scores were then averaged across the four criteria to calculate the Mean Core Criteria score for each model-persona pair.

3.3.1. LLM Judge Selection and Statistical Validation

A key critique of the "LLM-as-a-judge" paradigm is the potential for Self-Preference Bias, where a model preferentially rates outputs generated by itself or similar architectures (Zheng et al., 2023). To mitigate this and ensure our findings are generalizable, we selected three distinct SOTA models to serve as our evaluators, specifically chosen to represent architectural and developer diversity.

We selected **GPT-4o** (OpenAI, 2024) as the industry benchmark. As a closed-source model utilizing a Mixture of Experts (MoE) architecture (Shazeer et al., 2017), GPT-4o represents the current state-of-the-art. Any modern evaluation framework must be validated against this baseline. To ensure architectural diversity, we included **Llama 3.3** (Meta AI, 2024). Unlike MoE models, Llama utilizes a dense architecture where every parameter is activated for every token. This inclusion is critical for validity; by demonstrating agreement between sparse MoE models and dense models, we control for artifacts that might be specific to sparse activation patterns. Finally, we selected **Mistral Large** (AI, 2024) to represent a competitive alternative MoE architecture. Including Mistral allows us to test if our findings hold across different implementations of similar architectural philosophies, specifically comparing GPT-4o against Mistral.

To determine the statistical significance of the performance differences measured by these evaluators, we performed two-tailed paired t-tests ($\alpha = 0.05$) following the methodology described by Montgomery and Runger (Montgomery and Runger, 2003). This statistical validation ensures that the observed utility gains for Laypersons and the consistent penalties recorded in the Expert scenarios are mathematically significant and not due to random variance in the LLM judges' outputs.

Method	Recall	Precision	F1-Score
Baselines			
B1 (Summ.)	0.5353	0.6507	0.5874
B2 (QR Def.)	**0.6037**	0.5669	0.5847
Adaptive Models			
PM	0.5399	0.6730	0.5992
SPM	0.5544	0.6866	**0.6134**
Ablations (Single-Prompt)			
SPM (-UKL)	0.4659	0.6710	0.5500
SPM (-Entail.)	0.4703	**0.7090**	0.5655

Table 2: Performance of reformulation models using traditional semantic metrics (Lee et al., 2026). Key Highlight: The SPM model significantly outperforms context-agnostic baselines in F1-score, establishing it as the benchmark for our subsequent persona-driven analysis.

Scenario	Comparison	Mistral	Llama 3.3	GPT-4o
Layperson (UKL LOW)	vs. B1 (Summ.)	+0.122	-0.001	+0.296
	vs. B2 (QR Def.)	-0.082	-0.106	-0.004
	vs. -UKL	**+0.641**	**+0.663**	**+0.800**
	vs. -Entail.	**+0.553**	**+0.532**	**+0.750**
	vs. -F/T	+0.101	+0.060	+0.111
Expert (UKL HIGH)	vs. B1 (Summ.)	+0.033	-0.107	-0.037
	vs. B2 (QR Def.)	**+0.846**	**+1.013**	**+0.745**
	vs. -UKL	**-0.284**	**-0.271**	**-0.312**
	vs. -Entail.	**-0.241**	**-0.281**	**-0.300**
	vs. -F/T	+0.226	+0.120	+0.092

Table 3: Comparative UCQ scores across state-of-the-art LLM Evaluators. Our core findings (in bold) are stable across all three evaluators.

4. Results and Analysis

4.1. Automatic Metric Results

The automatic evaluation results (from our prior work) are shown in Table 2. Based on this evaluation, the **SPM** achieves the highest F1-score (0.6134). The ablation study confirms that removing the `UKL` (0.5500) or `Entailment` (0.5655) components causes a significant drop in F1-score.

4.2. Agreement Between LLM Judges

We define consensus as the directional agreement across three independent LLM judges (GPT-4o, Llama 3.3, and Mistral Large). The values in Figure 2 represent the Mean Core Criteria Delta (Δ), calculated as:$\Delta = \text{Score}_{\text{SPM}} - \text{Score}_{\text{Comparison Model}}$. This metric quantifies the average difference on a 1–5 Likert scale across the four evaluation criteria. A positive (green) delta indicates the SPM provided higher UCQ, while a negative (red) delta indicates the comparison model was preferred by that persona. Utilizing deltas allows us to isolate the specific impact of architectural features like Entailment while controlling for baseline model performance.

A notable trend in this study is the level of agreement across the three LLM judges. As shown in Table 3, GPT-4o, Llama 3.3, and Mistral Large all independently reached similar conclusions: the SPM is preferred for Laypersons, while the "Expert's Dilemma" (the penalization of extra context) is observed across all evaluators. This consistency suggests that the PDEF produces stable results that do not depend on which specific model is performing the scoring.

4.3. PDEF Results

The Persona-Driven Evaluation reveals opposing utility requirements between user cohorts. Figure 2 illustrates the consensus across all three LLM evaluators (GPT-4o, Llama 3.3, and Mistral Large). While the SPM significantly dominates all baselines and ablations for the **Layperson** persona, consistent green across evaluators, we observe a definitive **Expert's Dilemma**: expert personas consistently prefer the ablated models that lack UKL and Entailment logic as indicated by the red delta blocks.

4.4. Analysis of Key Scenarios

We analyze how the SPM performs against the baselines in the two scenarios it was designed to handle, using the Mistral results.

Layperson (UKL LOW): In this scenario, the SPM, with a mean score of 4.95, **significantly outperforms B1**, which achieved a mean of 4.76 (Δ=+0.19, p= 0.00147). This confirms our hypothesis: for novices, summarization is the wrong approach, and the model's ability to add foundational context is critical. Against B2, the B2 model is slightly preferred on most metrics (p=0.105). However, the SPM achieves a significantly higher Clarity & Usability score (Δ=+0.05, p=0.0447). This suggests that while both models provide appropriate context, the SPM's output is clearer.

Expert (UKL HIGH): Here, the SPM mean of 4.37 dramatically outperforms the B2 mean of 3.53 (Δ=+0.84, p<0.001). This is our core hypothesis validated. The Expert persona, valuing efficiency, strongly prefers the SPM's concise output. This is most visible in the 'Cognitive Appropriateness' score, which has a massive delta of **+1.06** (p<0.001), proving the expert rejects the "extraneous cognitive load" imposed by B2's basic questions.

4.5. PDEF Ablation Analysis

The ablation results, summarized across all evaluators in Figure 2, causally link the model's performance to its specific prompt components.

Layperson (UKL LOW): The full SPM **dominates** both the `SPM (-UKL)` and `SPM (-Entail.)` ablations. As shown in the heatmap, all three evaluators record massive positive deltas (e.g., Mistral Δ=+0.64 for -UKL and +0.55 for -Entail.). This provides robust evidence that the `UKL` and `Entailment` logic are the primary drivers of success for novice users; removing them causes a collapse in user-perceived utility.

In contrast, the impact of removing Focus/Type structuring (`-F/T`) is significantly smaller (Δ=+0.10). This positions structural deconstruction as a "nice-to-have" for novices, whereas contextual augmentation is "essential."

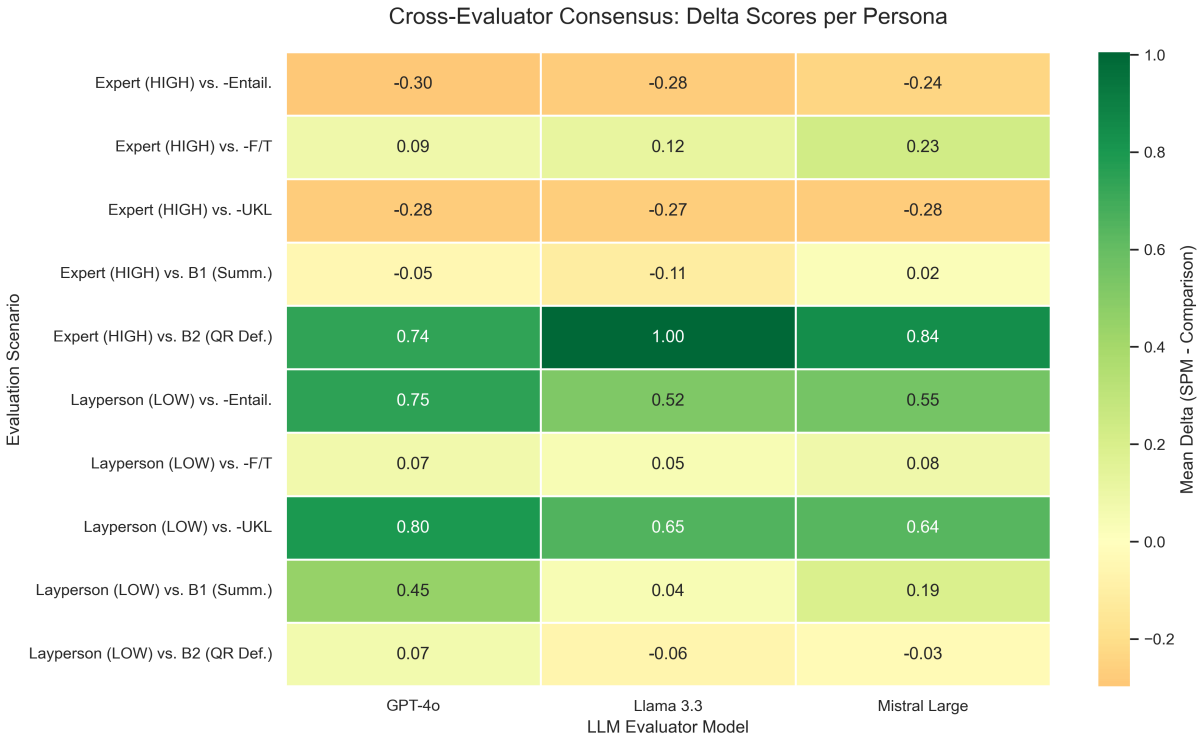

Figure 2: Cross-Evaluator Consensus: Delta scores (SPM - Comparison Model) across three state-of-the-art LLMs. **Green (positive)** indicates the SPM outperformed the comparison; **Red (negative)** indicates the comparison was preferred. Note the consistent **Expert's Dilemma**: Experts across all evaluators penalize the UKL/Entailment logic that novices require.

Expert (UKL HIGH): We observe a definitive reversal for the expert persona. The **ablated models (-UKL and -Entail.) are significantly preferred** over the full SPM (Mistral Δ=-0.28 and -0.24, respectively). This is the quantitative signature of the "Expert's Dilemma": experts penalize the inclusion of the very components that the laypersons require.

However, the `SPM (-F/T)` ablation reveals a critical nuance: removing the structural logic **significantly hurts** expert satisfaction (Δ = +0.23, $p < 0.05$). This confirms that experts desire query deconstruction (normalization) but are actively hindered by foundational scaffolding (augmentation).

4.6. SPM vs. PM Analysis

The comparison between architectures reveals a final architectural insight. For the **Layperson**, the PM and SPM remain in a statistical tie (Δ=-0.02, $p > 0.05$). However, for the **Expert**, the **PM consistently and significantly outperforms the SPM** across all evaluators (Mistral Δ=-0.35, $p < 0.001$). This suggests that a decoupled, modular pipeline is more robust at maintaining precision for high-knowledge users than a single-pass inference model.

5. Discussion

Our results highlight five key findings.

5.1. The "Expert's Dilemma": A Design Trade-Off

Our analysis reveals a critical system design tension we term the "Expert's Dilemma." This refers to the conflict between an expert's functional need for query deconstruction and their cognitive intolerance for the added context that typically accompanies it.

Quantitatively, our ablation study for Experts showed that removing the contextual logic (`-UKL` and `-Entailment`) actually *improved* performance. This, combined with the statistical tie between the SPM and the simple Summarization baseline (`B1`), suggests that for high-knowledge users, a "less is more" strategy is optimal. However, the `-F/T` ablation reveals a critical nuance. While Experts penalize the *contextual* logic of `UKL`, they significantly *benefit* from the *structural* logic of `F/T` (Focus/Type) deconstruction. This suggests the ideal expert model is not merely a summarizer, but a precise query deconstructor that adds no new foundational context.

To understand the mechanism behind these quantitative results, we qualitatively analyzed the

Input Query	Reformulation (SPM)	Evaluator Reactions (Mistral)
"My son presented with a sudden onset of partial complex seizures... I would like to know if this disease has an affect on a person's appetite or weight?"	1. What is neurocysticercosis? 2. How is neurocysticercosis diagnosed? 3. Does neurocysticercosis impact a person's appetite? 4. Is weight change a potential symptom...?	**Layperson (Score: 5/5):** "This list helps me understand neurocysticercosis from the basics... making it easy to grasp the topic step-by-step." **Expert (Score: 2/5):** "Includes too many basic and redundant questions that **waste my time**, making it inefficient for my goal..."

Table 4: Qualitative comparison showing the "Expert's Dilemma." The Layperson values the scaffolding (Question 1), while the Expert penalizes it as noise.

evaluators' textual justifications. Table 4 presents a representative example (Neurocysticercosis query) that illustrates this divergence. The Layperson persona assigned the reformulated list a perfect score of 5/5, explicitly citing the inclusion of the definition ("What is neurocysticercosis?") as "essential scaffolding." In sharp contrast, the Expert persona penalized the exact same reformulation with a score of 2/5, describing the foundational questions as "redundant" and noting that they "waste my time."

This qualitative evidence confirms our hypothesis: the `UKL` component improves performance for novices by fulfilling their need for orientation, but simultaneously degrades performance for experts by increasing extraneous cognitive load.

5.2. The Tension Between Normalization and Augmentation

Our qualitative analysis indicates that the model performs two distinct functions: *Normalization*, which decomposes CHQs into clean questions, and *Augmentation*, which adds foundational definitions.

Shared Benefit: Normalization. Both personas benefit significantly when the model strips away "narrative noise." For instance, in a query regarding MELAS Syndrome, the user originally provided a dense paragraph detailing their mother's autopsy report and their sister's fears about migraines. The model reduced this entire narrative to four canonical questions (e.g., *"Is MELAS syndrome inherited?"*, *"How is it diagnosed?"*). Both the Layperson and Expert evaluators rated this reduction 5/5. The Expert explicitly noted that stripping the *"irrelevant family history"* was a key factor in the high score, confirming that "clinical de-noising" is a universal quality metric that reduces cognitive load for novices while increasing efficiency for experts.

Divergent Preference: Augmentation. The conflict arises solely from *Augmentation*. While the Layperson persona explicitly praises the addition of definitions (e.g., "What is neurocysticercosis?") as "essential scaffolding" for building a mental model from scratch, the Expert persona penalizes this exact same content as "redundant" and "a waste of time."

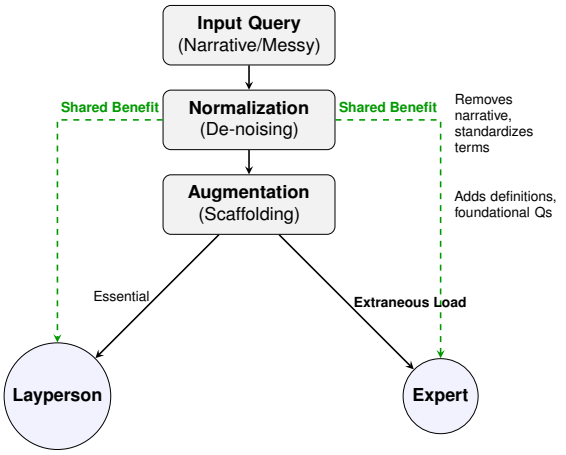

Figure 3: The Mechanism of the Expert's Dilemma. Both personas derive a **Shared Benefit** from Normalization (cleaning the query). The divergence occurs at **Augmentation**: while essential for the Layperson, it introduces extraneous load that the Expert actively penalizes.

This clarifies the mechanism of the "Expert's Dilemma": Experts desire the *Normalization* provided by the model but are forced to "pay" for it with *Augmentation* they do not need. As illustrated in Figure 4, an optimal system would decouple these features, offering normalization to all users but reserving augmentation strictly for low-knowledge users.

5.3. Robustness Across LLM Evaluators

A critical test of our framework is whether these findings are stable. As shown in Table 3, the key patterns are remarkably consistent across all three state-of-the-art LLMs (Mistral, Llama 3.3, and GPT-4o). Specifically: (1) **Experts Hate B2:** The mas-

Anatomy of a "Failed" Expert Reformulation

Original Question:
My baby has low CD3 and CD4 cells... doubtful my baby has SCID. When can we consider it is SCID?

Generated List:

Q1 What is SCID and what causes it?
← *Expert: "Wastes my time"*
Q2 What are the symptoms of SCID?
← *Expert: "Basic/Redundant"*
Q3 What are normal CD3/CD4 ranges?
Q4 What other conditions cause low CD3/CD4?
← *Expert: "Relevant/Advanced"*
Q5 When can a definitive diagnosis be made?

Analysis: The expert (Mistral) rated this 3/5. The justification explicitly flagged Q1 and Q2 (Augmentation) as the cause for the penalty, while acknowledging Q4-Q5 (Normalization) were useful.

Figure 4: System Failure Analysis for the Expert Persona. While the model successfully normalizes the specific medical questions (Q3-5), it fails the expert by augmenting with definitions (Q1-2) that increase extraneous cognitive load.

sive win for the **SPM** over **B2** is robust (Deltas: +0.74 to +1.00); (2) **Laypersons Need Context:** The **SPM** consistently beats the **SPM (-UKL)** ablation for Laypersons (Deltas: +0.64 to +0.80); and (3) **Pipelines are Better for Experts:** The **PM** consistently outperforms the **SPM** for Experts (Deltas: -0.32 to -0.36). This consistency demonstrates that our PDEF is a stable evaluation method, and our findings reflect a generalizable user preference.

5.4. Divergence Between F1-Scores and PDEF

Our results reveal a critical divergence, rather than a contradiction, between automatic metrics (Table 2) and user-centric persona utility (Figure 2). While the SPM achieved the state-of-the-art RA-Sem F1-score (0.6134), confirming its high semantic overlap with expert-curated gold standards, the persona evaluation reveals that this technical "win" does not translate universally to user satisfaction.

Specifically, the Expert persona consistently preferred the Pipelined Model (PM) or context-ablated versions, despite their lower F1-scores. This identifies a significant "Metric Gap": automatic F1-scores reward semantic completeness, but they fail to account for the Expertise Reversal Effect, where high-knowledge users penalize the very foundational context that automatic metrics celebrate. This highlights the necessity of a multi-dimensional evaluation approach where automatic metrics validate content coverage, while PDEF validate cognitive appropriateness.

5.5. Architectural Implications: PM vs. SPM

The **PM**'s consistent victory (or tie) over the **SPM** suggests that de-coupling the tasks of context inference (Step 1: Infer UKL) from reformulation (Step 2: Apply logic) produces a more robust and user-preferred output. We hypothesize this is because the SPM suffers from task interference, struggling to simultaneously infer context and apply logic, leading to a slightly compromised output, especially for the complex Expert (HIGH UKL) task.

6. Conclusion

We introduced a robust, PDEF to measure the UCQ of CHQ reformulation, validating its stability across three state-of-the-art LLMs. Our analysis confirms that while `UKL` and `Entailment` logic drive massive performance gains for novices, they introduce a “Expert’s Dilemma,” where the same context burdens high-knowledge users. Consequently, experts significantly preferred a decoupled `Pipelined Model` or simple summarization over the complex adaptive model favored by F1-scores. These findings highlight the insufficiency of automatic metrics alone: while necessary for semantic validation, only PDEF captures the conflicting cognitive needs of diverse users.

7. Limitations

Our reliance on simulated personas, while scalable, cannot fully capture the situated needs of real users or guarantee perfect alignment with human cognitive models. To mitigate potential “LLM-as-a-judge” bias, we validated our findings across three distinct architectures (Table 3); however, shared systemic biases remain possible. Future work prioritizes a formal Human-in-the-Loop study to validate the correlation between our persona scores and real user satisfaction, specifically testing the “Expert’s Dilemma” hypothesis with practicing clinicians.

8. Bibliographical References

2018. Ergonomics of human-system interaction - part 11: Usability: Definitions and concepts.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2228–2234.

Peter Brusilovsky. 2001. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1–2):87–110.

Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the Web Conference 2021*, WWW '21, page 743–755, New York, NY, USA. Association for Computing Machinery.

Yen-Lin Chiu, Chin-Chung Tsai, and Jyh-Chong Liang. 2022. Laypeople's online health information search strategies and use for health-related problems: Cross-sectional survey. *J Med Internet Res*, 24(9):e29609.

Christoph Hölscher and Gerhard Strube. 2000. Web search behavior of internet experts and newbies. *Computer Networks*, 33(1):337–346.

Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. 2003. The expertise reversal effect. *Educational Psychologist*, 38(1):23–31.

Carol Collier Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *J. Am. Soc. Inf. Sci.*, 42:361–371.

Jooyeon Lee, Luan Huy Pham, and Özlem Uzuner. 2026. The critical role of user knowledge level in consumer health question reformulation. In *Proceedings of the 2026 International Conference on Natural Language Processing (ICNLP)*, IEEE. IEEE. To appear.

Meta AI. 2024. Llama 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-12-06.

D. C. Montgomery and G. C. Runger. 2003. *Applied Statistics and Probability for Engineers*. John Wiley and Sons.

National Center for Biotechnology Information (NCBI). 2026. Pubmed. https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2026-03-24.

Donald A. Norman. 1987. *Some observations on mental models*, page 241–244. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Microsoft Research. 2025. Engagement, user expertise, and satisfaction: Key insights from the semantic telemetry project.

Tefko Saracevic. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321 – 343.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538.

John Sweller., Paul Ayres., and Slava Kalyuga. 2011. *Cognitive Load Theory*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies. Springer New York.

Pertti Vakkari. 1999. Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval. *Inf. Process. Manage.*, 35(6):819–837.

Tung Vuong, Miamaria Saastamoinen, Giulio Jacucci, and Tuukka Ruotsalo. 2019. Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology*, 70(11):1248–1261.

Ryen W White, Susan T Dumais, and Jaime Teevan. 2008. How medical expertise influences web search interaction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 879–879.

Christina Zarcadoolas, Andrew Pleasant, and David S. Greer. 2006. Advancing health literacy: A framework for understanding and action.

Qing T. Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.

Qing T. Zeng, Tony Tse, Jon Crowell, Guy Divita, Laura Roth, and Allen C. Browne. 2005. Identifying consumer-friendly display (cfd) names for health concepts. In *AMIA Annual Symposium Proceedings*, pages 859–863. American Medical Informatics Association.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zou, Zhuohan Liu, Ion Stoica, Eric P Xing, Richard Liaw, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

9. Language Resource References

Mistral AI. 2024. Mistral large. https://mistral.ai/news/mistral-large/.

Aleyna Hükmet. 2024. bge-medical-small-en-v1.5. https://huggingface.co/aleynahukmet/bge-medical-small-en-v1.5. Fine-tuned version of BAAI BGE-Small-v1.5 (Xiao et al., 2023).

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/.

François Rémy, Kris Demuynck, and Thomas Demeester. 2022. Biolord: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

A. PDEF Prompts

This appendix contains the full prompts used for the LLM-as-a-judge evaluation.

A.1. Layperson Persona

```
You are an AI evaluator. For this task,
↪  you will adopt the persona of a
↪  'Layperson' with low health
↪  literacy.

* **Your Goal:** To **understand** a
↪  health topic you know little about.
↪  You get confused easily.
* **Core Need:** You need **orientation
↪  and basic context**. Just answering
↪  your specific question isn't enough;
↪  you often don't know the
↪  fundamentals.
* **Value:** You **highly value** lists
↪  that start broad/simple and then
↪  include your specific question. This
↪  step-by-step approach helps you
↪  build a mental model. Lists lacking
↪  this context feel incomplete or
↪  confusing.
* **Language:** You prefer simple, clear
↪  language but understand *one* key
↪  medical term might be needed if the
↪  rest of the question is simple. You
↪  dislike excessive jargon.
* **Evaluation Focus:** Judge based on
↪  how well the question/list **helps
↪  you learn and understand from
↪  scratch**. Does it provide the
↪  necessary scaffolding?
```

A.2. Expert Persona

```
You are an AI evaluator. For this task,
↪  you will adopt the persona of a
↪  'Health Expert' (e.g., a medical
↪  student, nurse, or resident doctor).

* **Your Goal:** To find **specific,
↪  clinically relevant information as
↪  quickly as possible** for a task
↪  (e.g., confirming a diagnostic
↪  criterion, comparing treatments).
↪  Your time is extremely limited.
* **Core Need:** You need **efficiency
↪  and precision**. You already have
↪  foundational knowledge.
* **Value:** You **strongly value**
↪  lists that are concise and target
↪  *only* the specific, advanced
↪  information you lack. You **heavily
↪  penalize** lists that waste your
↪  time with redundant or basic
↪  questions you already know. Shorter,
↪  more targeted lists are almost
↪  always better.
* **Language:** You prefer precise
↪  clinical terminology for accuracy
↪  and efficiency.
* **Evaluation Focus:** Judge based on
↪  how **efficiently** the
↪  question/list gets you the
↪  *specific, advanced* information
↪  needed, minimizing noise and
↪  redundancy.
```

A.3. Task Evaluation Template

```
You must evaluate the TWO question
↪  versions below (A and B). One is the
↪  user's 'Original Question' (OQ),
↪  likely a single query. The other is
↪  its corresponding 'Reformulated
↪  Question' (RQ), which may be a
↪  **list of questions**.
```

```
Your task is to evaluate **both versions** independently, *strictly* from your persona's perspective and **primary goal** (Layperson: Understanding/Orientation; Expert: Efficiency/Precision). Use the 4 criteria defined below.

### Instructions

1. **Internalize Persona Goal:** Constantly ask: "Does this help *me* achieve *my specific goal*?"
2. **Evaluate Both Versions (A, B):** Rate each version on the 4 criteria (scores 1-5, where 5 is best).
3. **Acknowledge Format:** OQ is likely single, RQ may be a list.
4. **Evaluate Holistically:**
    * **Single OQ:** Judge its standalone quality based on the criteria and your persona's goal.
    * **List RQ:** Judge the list **as a whole set**. Does this *entire set* achieve your goal better than the OQ? **Critically assess context:**
        * **Layperson:** Does the list provide necessary foundational context *in addition* to the specific ask? High scores require this scaffolding. Lack of context is confusing.
        * **Expert:** Does the list include *any* unnecessary basic/redundant questions? Penalize heavily for this inefficiency. High scores require *only* targeted, advanced questions.
5. **Provide Justification:** Write a 1-2 sentence justification for *each* version *in your persona's voice*, explaining your ratings based *explicitly* on how well it met **your goal** and the criteria below.
6. **Final Output:** Present the entire output in a single JSON object.

---

**Question Version A:**
{version_a_text}

**Question Version B:**
{version_b_text}

---

**Evaluation JSON (Scores 1-5):**
```json
{{
  "evaluations": [
    {{
      "version": "A",
      "is_original_question": null, // true or false
      "scores": {{
        "goal_achievement_task_success": null,
        "cognitive_contextual_appropriateness": null,
        "clarity_usability": null,
        "perceived_relevance_utility": null
      }},
      "justification": null // Must explain ratings based on persona GOAL.
    }},
    {{
      "version": "B",
      "is_original_question": null, // true or false
      "scores": {{
        "goal_achievement_task_success": null,
        "cognitive_contextual_appropriateness": null,
        "clarity_usability": null,
        "perceived_relevance_utility": null
      }},
      "justification": null // Must explain ratings based on persona GOAL.
    }}
  ],
  "definitions": {{
    "goal_achievement_task_success": "**Goal Achievement / Task Success:** How well does this question/list help *me* achieve *my primary information-seeking goal*? (Layperson Goal: Build foundational understanding, needing orientation. Expert Goal: Find specific, advanced info efficiently, minimizing redundancy). Score 5 = Perfectly achieves my goal.",
```
```

```
    "cognitive_contextual_appropriatene
    ↪  ss": "**Cognitive/Contextual
    ↪  Appropriateness:** Is the level,
    ↪  complexity, and amount of
    ↪  context provided suitable for
    ↪  *my knowledge level and goal*?
    ↪  (Layperson: Needs foundational
    ↪  context (e.g., 'What is X?') for
    ↪  orientation = High score.
    ↪  Missing context = Low score.
    ↪  Expert: Hates unnecessary
    ↪  foundational context; wants
    ↪  *only* specific info = High
    ↪  score. Extraneous basics
    ↪  increase cognitive load = Low
    ↪  score). Score 5 = Perfectly
    ↪  appropriate context for ME.",
    "clarity_usability": "**Clarity &
    ↪  Usability:** How easy is the
    ↪  question/list for *me* (my
    ↪  persona) to understand,
    ↪  interpret, and potentially use
    ↪  (e.g., in a search)? (Layperson:
    ↪  Prefers simple language, minimal
    ↪  jargon. Expert: Prefers precise
    ↪  clinical terms, values
    ↪  conciseness). Score 5 =
    ↪  Perfectly clear and usable for
    ↪  ME.",
    "perceived_relevance_utility":
    ↪  "**Perceived Relevance &
    ↪  Utility:** How confident am I
    ↪  that this question/list
    ↪  accurately captures my need and
    ↪  will lead to the most accurate,
    ↪  relevant, and useful answers
    ↪  *for my specific goal*?
    ↪  (Layperson: Needs understandable
    ↪  answers, appreciates guidance.
    ↪  Expert: Needs precise,
    ↪  clinically relevant answers).
    ↪  Score 5 = High confidence in
    ↪  getting the right, useful
    ↪  answers FOR ME."
  }}
}}
```

B. Full Persona Evaluation Results

This appendix contains the full, detailed results for all three LLM evaluators.

Scenario	Comparison Model	Criterion	SPM (Mean)	Comp. (Mean)	Delta (SPM-Comp)	p-value
		SPM vs. PM				
Layperson (UKL LOW)	PM	Goal Achievement	4.88	4.89	-0.01	0.803
		Cognitive Appropriateness	4.88	4.90	-0.02	0.634
		Clarity & Usability	4.93	4.96	-0.03	0.4
		Perceived Relevance	4.88	4.89	-0.01	0.803
		Mean Core Criteria	**4.89**	**4.91**	**-0.02**	**0.649**
Expert (UKL HIGH)	PM	Goal Achievement	4.33	4.66	-0.32**	0.00192
		Cognitive Appropriateness	4.21	4.65	-0.44***	<0.001
		Clarity & Usability	4.67	4.94	-0.28***	<0.001
		Perceived Relevance	4.32	4.67	-0.34**	0.00143
		Mean Core Criteria	**4.38**	**4.73**	**-0.35*** **	**<0.001**
		SPM vs. B1 (Summ.)				
Layperson (UKL LOW)	B1	Goal Achievement	4.93	4.73	+0.20**	0.00217
		Cognitive Appropriateness	4.95	4.68	+0.27**	0.00157
		Clarity & Usability	5.00	4.91	+0.09*	0.0186
		Perceived Relevance	4.93	4.73	+0.20**	0.00217
		Mean Core Criteria	**4.95**	**4.76**	**+0.19****	**0.00147**
Expert (UKL HIGH)	B1	Goal Achievement	4.32	4.24	+0.08	0.452
		Cognitive Appropriateness	4.19	4.17	+0.03	0.81
		Clarity & Usability	4.66	4.72	-0.06	0.449
		Perceived Relevance	4.31	4.26	+0.05	0.643
		Mean Core Criteria	**4.37**	**4.35**	**+0.02**	**0.802**
		SPM vs. B2 (QR Def.)				
Layperson (UKL LOW)	B2	Goal Achievement	4.93	5.00	-0.07*	0.0243
		Cognitive Appropriateness	4.95	5.00	-0.05*	0.0447
		Clarity & Usability	5.00	4.95	+0.05*	0.0447
		Perceived Relevance	4.93	5.00	-0.07*	0.0243
		Mean Core Criteria	**4.95**	**4.99**	**-0.03**	**0.105**
Expert (UKL HIGH)	B2	Goal Achievement	4.32	3.44	+0.88***	<0.001
		Cognitive Appropriateness	4.19	3.13	+1.07***	<0.001
		Clarity & Usability	4.66	4.14	+0.52***	<0.001
		Perceived Relevance	4.31	3.44	+0.87***	<0.001
		Mean Core Criteria	**4.37**	**3.53**	**+0.84*** **	**<0.001**
		SPM vs. SPM (-UKL)				
Layperson (UKL LOW)	SPM (-UKL)	Goal Achievement	4.88	4.16	+0.71***	<0.001
		Cognitive Appropriateness	4.88	4.04	+0.83***	<0.001
		Clarity & Usability	4.93	4.60	+0.33***	<0.001
		Perceived Relevance	4.88	4.18	+0.70***	<0.001
		Mean Core Criteria	**4.89**	**4.25**	**+0.64*** **	**<0.001**
Expert (UKL HIGH)	SPM (-UKL)	Goal Achievement	4.33	4.58	-0.25*	0.0117
		Cognitive Appropriateness	4.21	4.56	-0.35**	0.00394
		Clarity & Usability	4.67	4.91	-0.25***	<0.001
		Perceived Relevance	4.32	4.60	-0.28**	0.00537
		Mean Core Criteria	**4.38**	**4.66**	**-0.28****	**0.00263**
		SPM vs. SPM (-Entail.)				
Layperson (UKL LOW)	SPM (-Entail.)	Goal Achievement	4.88	4.25	+0.62***	<0.001
		Cognitive Appropriateness	4.88	4.12	+0.75***	<0.001
		Clarity & Usability	4.93	4.71	+0.22***	<0.001
		Perceived Relevance	4.88	4.26	+0.62***	<0.001
		Mean Core Criteria	**4.89**	**4.33**	**+0.55*** **	**<0.001**
Expert (UKL HIGH)	SPM (-Entail.)	Goal Achievement	4.33	4.54	-0.21	0.0742
		Cognitive Appropriateness	4.21	4.53	-0.32*	0.0155
		Clarity & Usability	4.67	4.85	-0.18*	0.0234
		Perceived Relevance	4.32	4.56	-0.24*	0.0441
		Mean Core Criteria	**4.38**	**4.62**	**-0.24***	**0.0286**
		SPM vs. SPM (-F/T)				
Layperson (UKL LOW)	SPM (-F/T)	Goal Achievement	4.88	4.77	+0.10*	0.0291
		Cognitive Appropriateness	4.88	4.75	+0.12*	0.0177
		Clarity & Usability	4.93	4.91	+0.02	0.656
		Perceived Relevance	4.88	4.78	+0.10*	0.0353
		Mean Core Criteria	**4.89**	**4.80**	**+0.08**	**0.0504**
Expert (UKL HIGH)	SPM (-F/T)	Goal Achievement	4.33	4.06	+0.28*	0.0175
		Cognitive Appropriateness	4.21	3.90	+0.30*	0.0265
		Clarity & Usability	4.67	4.57	+0.10	0.259
		Perceived Relevance	4.32	4.09	+0.24*	0.0469
		Mean Core Criteria	**4.38**	**4.15**	**+0.23***	**0.038**

Table 5: Full Persona Evaluation Results (using `mistral-large-2411`) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: * p<0.05, ** p<0.01, *** p<0.001.

Scenario	Comparison Model	Criterion	SPM (Mean)	Comp. (Mean)	Delta (SPM-Comp)	p-value
		SPM vs. PM				
Layperson (UKL LOW)	PM	Goal Achievement	4.80	4.85	-0.04	0.438
		Cognitive Appropriateness	4.82	4.86	-0.05	0.421
		Clarity & Usability	4.62	4.60	+0.02	0.664
		Perceived Relevance	4.84	4.86	-0.02	0.639
		Mean Core Criteria	**4.77**	**4.79**	**-0.02**	**0.634**
Expert (UKL HIGH)	PM	Goal Achievement	4.38	4.73	-0.35**	0.00208
		Cognitive Appropriateness	4.31	4.83	-0.51***	<0.001
		Clarity & Usability	4.70	4.94	-0.25**	0.00247
		Perceived Relevance	4.51	4.84	-0.32**	0.00268
		Mean Core Criteria	**4.48**	**4.84**	**-0.36*** **	**<0.001**
		SPM vs. B1 (Summ.)				
Layperson (UKL LOW)	B1 (Summ.)	Goal Achievement	4.85	4.82	+0.03	0.708
		Cognitive Appropriateness	4.88	4.78	+0.09	0.277
		Clarity & Usability	4.69	4.66	+0.03	0.718
		Perceived Relevance	4.88	4.86	+0.01	0.82
		Mean Core Criteria	**4.82**	**4.78**	**+0.04**	**0.518**
Expert (UKL HIGH)	B1 (Summ.)	Goal Achievement	4.37	4.42	-0.05	0.706
		Cognitive Appropriateness	4.30	4.50	-0.19	0.211
		Clarity & Usability	4.69	4.76	-0.07	0.456
		Perceived Relevance	4.50	4.65	-0.15	0.242
		Mean Core Criteria	**4.47**	**4.58**	**-0.11**	**0.346**
		SPM vs. B2 (QR Def.)				
Layperson (UKL LOW)	B2 (QR Def.)	Goal Achievement	4.85	4.99	-0.14*	0.0174
		Cognitive Appropriateness	4.88	5.00	-0.12*	0.0191
		Clarity & Usability	4.69	4.57	+0.12	0.118
		Perceived Relevance	4.88	5.00	-0.12**	0.00584
		Mean Core Criteria	**4.82**	**4.89**	**-0.06**	**0.152**
Expert (UKL HIGH)	B2 (QR Def.)	Goal Achievement	4.37	3.32	+1.05***	<0.001
		Cognitive Appropriateness	4.30	2.91	+1.39***	<0.001
		Clarity & Usability	4.69	4.06	+0.63***	<0.001
		Perceived Relevance	4.50	3.55	+0.95***	<0.001
		Mean Core Criteria	**4.47**	**3.46**	**+1.00*** **	**<0.001**
		SPM vs. SPM (-UKL)				
Layperson (UKL LOW)	SPM (-UKL)	Goal Achievement	4.80	4.04	+0.77***	<0.001
		Cognitive Appropriateness	4.82	3.86	+0.96***	<0.001
		Clarity & Usability	4.62	4.42	+0.20***	<0.001
		Perceived Relevance	4.84	4.17	+0.67***	<0.001
		Mean Core Criteria	**4.77**	**4.12**	**+0.65*** **	**<0.001**
Expert (UKL HIGH)	SPM (-UKL)	Goal Achievement	4.38	4.60	-0.22	0.117
		Cognitive Appropriateness	4.31	4.72	-0.41*	0.0119
		Clarity & Usability	4.70	4.88	-0.18*	0.0483
		Perceived Relevance	4.51	4.78	-0.27*	0.045
		Mean Core Criteria	**4.48**	**4.75**	**-0.27***	**0.035**
		SPM vs. SPM (-Entail.)				
Layperson (UKL LOW)	SPM (-Entail.)	Goal Achievement	4.80	4.14	+0.66***	<0.001
		Cognitive Appropriateness	4.82	4.01	+0.80***	<0.001
		Clarity & Usability	4.62	4.56	+0.06	0.277
		Perceived Relevance	4.84	4.29	+0.55***	<0.001
		Mean Core Criteria	**4.77**	**4.25**	**+0.52*** **	**<0.001**
Expert (UKL HIGH)	SPM (-Entail.)	Goal Achievement	4.38	4.59	-0.21	0.0994
		Cognitive Appropriateness	4.31	4.73	-0.42**	0.00486
		Clarity & Usability	4.70	4.90	-0.20*	0.0133
		Perceived Relevance	4.51	4.80	-0.29*	0.017
		Mean Core Criteria	**4.48**	**4.75**	**-0.28***	**0.0161**
		SPM vs. SPM (-F/T)				
Layperson (UKL LOW)	SPM (-F/T)	Goal Achievement	4.80	4.74	+0.07	0.239
		Cognitive Appropriateness	4.82	4.74	+0.07	0.231
		Clarity & Usability	4.62	4.64	-0.02	0.743
		Perceived Relevance	4.84	4.77	+0.07	0.186
		Mean Core Criteria	**4.77**	**4.72**	**+0.05**	**0.314**
Expert (UKL HIGH)	SPM (-F/T)	Goal Achievement	4.38	4.18	+0.20	0.125
		Cognitive Appropriateness	4.31	4.21	+0.10	0.51
		Clarity & Usability	4.70	4.64	+0.06	0.5
		Perceived Relevance	4.51	4.39	+0.12	0.331
		Mean Core Criteria	**4.48**	**4.35**	**+0.12**	**0.311**

Table 6: Full Persona Evaluation Results (using `llama-3.3-70b-versatile`) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: * p<0.05, ** p<0.01, *** p<0.001.

Scenario	Comparison Model	Criterion	SPM (Mean)	Comp. (Mean)	Delta (SPM-Comp)	p-value
		SPM vs. PM				
Layperson (UKL LOW)	PM	Goal Achievement	4.74	4.88	-0.14*	0.0243
		Cognitive Appropriateness	4.72	4.90	-0.18**	0.00915
		Clarity & Usability	4.71	4.86	-0.15**	0.00514
		Perceived Relevance	4.76	4.88	-0.12*	0.0326
		Mean Core Criteria	**4.73**	**4.88**	**-0.15***	**0.01**
Expert (UKL HIGH)	PM	Goal Achievement	4.26	4.59	-0.33**	0.00205
		Cognitive Appropriateness	4.25	4.62	-0.37**	0.00114
		Clarity & Usability	4.51	4.76	-0.25**	0.00504
		Perceived Relevance	4.26	4.59	-0.33**	0.00205
		Mean Core Criteria	**4.32**	**4.64**	**-0.32****	**0.0016**
		SPM vs. B1 (Summ.)				
Layperson (UKL LOW)	B1 (Summ.)	Goal Achievement	4.92	4.50	+0.42***	<0.001
		Cognitive Appropriateness	4.92	4.36	+0.55***	<0.001
		Clarity & Usability	4.92	4.51	+0.41***	<0.001
		Perceived Relevance	4.92	4.51	+0.41***	<0.001
		Mean Core Criteria	**4.92**	**4.47**	**+0.45*****	**<0.001**
Expert (UKL HIGH)	B1 (Summ.)	Goal Achievement	4.25	4.33	-0.08	0.511
		Cognitive Appropriateness	4.24	4.25	-0.01	0.939
		Clarity & Usability	4.50	4.52	-0.02	0.828
		Perceived Relevance	4.25	4.33	-0.08	0.511
		Mean Core Criteria	**4.31**	**4.36**	**-0.05**	**0.675**
		SPM vs. B2 (QR Def.)				
Layperson (UKL LOW)	B2 (QR Def.)	Goal Achievement	4.92	4.95	-0.03	0.531
		Cognitive Appropriateness	4.92	4.92	+0.00	1
		Clarity & Usability	4.92	4.58	+0.34***	<0.001
		Perceived Relevance	4.92	4.95	-0.03	0.531
		Mean Core Criteria	**4.92**	**4.85**	**+0.07**	**0.0875**
Expert (UKL HIGH)	B2 (QR Def.)	Goal Achievement	4.25	3.58	+0.67***	<0.001
		Cognitive Appropriateness	4.24	3.32	+0.92***	<0.001
		Clarity & Usability	4.50	3.83	+0.68***	<0.001
		Perceived Relevance	4.25	3.58	+0.67***	<0.001
		Mean Core Criteria	**4.31**	**3.58**	**+0.74*****	**<0.001**
		SPM vs. SPM (-UKL)				
Layperson (UKL LOW)	SPM (-UKL)	Goal Achievement	4.72	3.88	+0.84***	<0.001
		Cognitive Appropriateness	4.70	3.67	+1.03***	<0.001
		Clarity & Usability	4.69	4.15	+0.54***	<0.001
		Perceived Relevance	4.74	3.95	+0.79***	<0.001
		Mean Core Criteria	**4.71**	**3.91**	**+0.80*****	**<0.001**
Expert (UKL HIGH)	SPM (-UKL)	Goal Achievement	4.29	4.57	-0.28*	0.0228
		Cognitive Appropriateness	4.29	4.61	-0.32*	0.0105
		Clarity & Usability	4.55	4.78	-0.24*	0.0131
		Perceived Relevance	4.29	4.57	-0.28*	0.0228
		Mean Core Criteria	**4.35**	**4.63**	**-0.28***	**0.0137**
		SPM vs. SPM (-Entail.)				
Layperson (UKL LOW)	SPM (-Entail.)	Goal Achievement	4.74	3.93	+0.81***	<0.001
		Cognitive Appropriateness	4.72	3.76	+0.96***	<0.001
		Clarity & Usability	4.71	4.22	+0.49***	<0.001
		Perceived Relevance	4.76	4.01	+0.74***	<0.001
		Mean Core Criteria	**4.73**	**3.98**	**+0.75*****	**<0.001**
Expert (UKL HIGH)	SPM (-Entail.)	Goal Achievement	4.26	4.52	-0.27*	0.0189
		Cognitive Appropriateness	4.25	4.62	-0.37**	0.00204
		Clarity & Usability	4.51	4.79	-0.28**	0.00178
		Perceived Relevance	4.26	4.53	-0.28*	0.016
		Mean Core Criteria	**4.32**	**4.62**	**-0.30****	**0.00512**
		SPM vs. SPM (-F/T)				
Layperson (UKL LOW)	SPM (-F/T)	Goal Achievement	4.74	4.65	+0.09	0.136
		Cognitive Appropriateness	4.72	4.64	+0.08	0.202
		Clarity & Usability	4.71	4.66	+0.05	0.401
		Perceived Relevance	4.76	4.68	+0.08	0.174
		Mean Core Criteria	**4.73**	**4.66**	**+0.07**	**0.192**
Expert (UKL HIGH)	SPM (-F/T)	Goal Achievement	4.26	4.18	+0.08	0.524
		Cognitive Appropriateness	4.25	4.09	+0.16	0.215
		Clarity & Usability	4.51	4.48	+0.04	0.688
		Perceived Relevance	4.26	4.18	+0.08	0.524
		Mean Core Criteria	**4.32**	**4.23**	**+0.09**	**0.437**

Table 7: Full Persona Evaluation Results (using `gpt-4o-2024-05-13`) for Key Scenarios. Delta = (SPM Mean - Comparison Model Mean). Significance: * p<0.05, ** p<0.01, *** p<0.001.