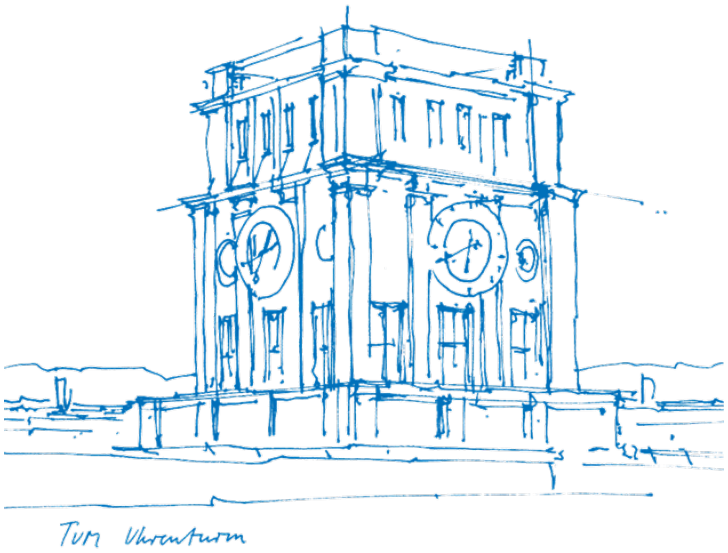
Chair of Software Engineering & AI
School of Computation, Information and Technology (CIT)
Technical University of Munich

Mitigating Hallucination Potential in User Prompts Through AI-Guided Iterative Refinement

Bachelor Thesis

Mohamed Nejjar



Mitigating Hallucination Potential in User Prompts Through AI-Guided Iterative Refinement

Bachelor Thesis

Mohamed Nejjar

Chair of Software Engineering & AI
School of Computation, Information and Technology (CIT)
Technical University of Munich

Mitigating Hallucination Potential in User Prompts Through AI-Guided Iterative Refinement

Bachelor Thesis

Mohamed Nejjar

Thesis for the attainment of the academic degree

Bachelor of Science (B.Sc.)

at the School of Computation, Information and Technology (CIT) of the Technical University of Munich.

Examiner:
Prof. Dr. Chunyang Chen

Supervisor:
Dr. Mark Huasong Meng

Submitted:
Munich, 18.12.2025

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 18.12.2025 Mohamed Nejjar

Dedication

To my parents and my little brother,
whose love and strength carried me through every hardship.

And to my younger self, who never stopped dreaming, even when I did.

Preface

A Personal Note

Writing this thesis marks the end of a journey that has shaped me far beyond the boundaries of academic work. Between 2023 and 2024, I went through one of the most difficult periods of my life: facing illness, repeated setbacks in my studies, and moments where I no longer believed in myself.

---

*For a long time, it felt as though everything in my life was moving in the wrong direction, and there were days when **I struggled to see a way forward**.*

Yet through persistence, discipline, and countless days of piecing myself back together, I slowly grew stronger. By the end of 2024, something changed: effort began turning into progress, and it started to bear fruit. I went from a student who once only dreamed of studying at TUM from afar in Morocco to someone who contributed to research with real impact, collaborated with Fraunhofer, Allianz, and BCG, and rediscovered confidence in his own abilities.

Working on this thesis deepened that transformation.

It taught me to think more clearly, to question assumptions, and to approach complex systems with both creativity and rigor. It reminded me that progress is not linear, but with perseverance, growth is inevitable.

Along the way, I also learned to accept that each person moves forward at a different rhythm, and that *my own pace was perfectly enough.*

I am grateful for every challenge, because each one contributed to the person I am now and the person I am becoming.

This thesis is therefore not only an academic milestone, but a personal one — a testament to resilience, self-belief, and the power of new beginnings.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Mark Huasong Meng, for his exceptional guidance throughout this thesis. His openness to ideas, the freedom he allowed me to experiment, his ability to challenge my assumptions, and his constant availability for discussion made this work not only possible but deeply enriching. His mentorship shaped both the technical quality of this thesis and my growth as a researcher.

I also wish to thank the Chair of Software and Systems Engineering for providing an environment that encouraged intellectual curiosity and independence.

My heartfelt thanks go to Prof. Dr. Ingo Weber, whose trust and support were instrumental at the starting stages of my academic journey. He selected me for the DevOps and LLM4Science seminar, offered me a position in the highly competitive Fraunhofer practical course on retrieval-augmented generation, and later became my co-author in my first research paper. His confidence in my potential opened doors that defined my trajectory in AI research.

I am equally grateful to Dr. Jan Niederreiter and the AIML team at Allianz SE for giving me my first opportunity as a working student. Their support, professionalism, and the sense of belonging i have felt during my employment year played a major role in my development as an AI engineer but also as a person.

I extend my thanks to my fellow students, colleagues, and friends who created an atmosphere of excellence, an environment in which I constantly strived to become the best version of myself.

Finally, I wish to thank my family for their unwavering belief, their constant encouragement, and their unconditional support. Their strength and love have been the foundation upon which every achievement in my life is built.

Abstract

Faithfulness-related hallucinations in large language models (LLMs) stem not only from model-side limitations like training and model architecture but also from inconsistencies in user prompts: ambiguity, missing essentials, vague constraints, conflicting instructions, and structural weaknesses can all degrade an LLM's output quality by forcing the model to make unstated assumptions, overfill informational gaps, or prioritize incompatible objectives which ultimately increases the likelihood of unsupported, incoherent, or fabricated content. Additionally, due to their compliant nature, LLMs rarely warn users about gaps or contradictions in their prompts. Instead of seeking clarification, they improvise filling in missing details in ways that can subtly distort meaning or misrepresent the user's intent.

Although prior work revealed many of these issues, existing research and practitioner guidelines are spread across disparate sources, leaving no shared framework for detecting problems at the token level. This thesis fills that gap through a Design Science Research (DSR) approach that turns user-side risks into a two-dimensional **Prompt/Meta Risk Taxonomy** and makes it operational through a modular XML guideline set that can be applied across annotation and analysis.

Building on this foundation, the thesis introduces ECHO, a multi-agent artifact designed for shift-left, prompt-time hallucination mitigation. Developed through iterative design-space exploration, ECHO combines guideline-driven risk detection with a targeted clarification mechanism and a conversational refinement loop that resolves ambiguity and missing information. Change in model quality is then approximated using a normalized metric, **Prompt Risk Density (PRD)**, which quantifies prompt-time risk and enables before/after comparison across multiple refinements.

A 316-prompt benchmark, including lexical-variation pairs, evaluates detection accuracy, stability under rephrasing, and qualitative refinement effectiveness. Results show reliable token-level classification, stable detections across surface variations, and consistent PRD reductions after refinement. Qualitative examples further illustrate decreases in ambiguity and clearer alignment with user intent.

The thesis contributes: (i) a structured Prompt/Meta taxonomy which contains (ii) a reproducible and modular guideline set, (iii) PRD as a prompt-time risk metric, (iv) the ECHO artifact implementing the taxonomy, and (v) an evaluation framework for prompt-time risk analysis. Together, these results demonstrate that formalizing user-sided risks and guiding the user through conversational, iterative refinement can mitigate faithfulness issues *before generation*, complementing factuality as well as model-sided mitigation strategies.

Keywords: hallucinations, prompt engineering, faithfulness, taxonomy, design science research, LLM evaluation, risk detection, prompt refinement, prompt risk density

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation & Context

The surfacing of Large Language Models (LLMs) has denoted a paradigm shift in Natural Language Processing. They have shown impressive capabilities across a wide array of tasks[1]–[4] and their ability to generate coherent, context-aware and most importantly plausible text has led to deployments in various sectors of society e.g. in science, industry and every-day applications.

One of the critical limitations of these advancements are hallucinations. Broadly, hallucinations are model-generated statements that are factually incorrect, logically inconsistent, or unfaithful to user instructions or provided context [5]. Hallucinations are categorized into two types [6] : *faithfulness hallucinations* where outputs that deviate from the original user prompt or supplied data while *factuality hallucinations* emphasize the discrepancy between generated content and verifiable real-world facts. Such deviations raise concerns since it can undermine reliability in critical or high-stakes settings such as insurance, healthcare and law.

Across domains, the consequences of hallucinations are concrete and well documented. In law, error rates remain high: citation and attribution tasks fail in over 80% of outputs across GPT-4, PaLM-2, and LLaMA-2 [7], and even commercial tools marketed as “hallucination-free” still return false or misleading results in 17–33% of queries [8]. In healthcare, hallucinations can distort diagnoses or therapeutic reasoning, requiring strong governance and oversight for safe deployment [9]. In personal finance, LLMs systematically increase portfolio risk, raising concentration, sector exposure, and fee burdens, while prompt debiasing only partially mitigates these effects [10]. Together, these cases show that hallucinations are not rare glitches but everyday reliability failures that can meaningfully shape outcomes in high-stakes domains.

While the literature identifies numerous causes—including training data limitations, objective misalignment, retrieval failures, and interface design [5], [6], [11]—*prompt design remains an under-addressed source of faithfulness errors*. Ambiguous referents, missing constraints, vague quantifiers, conflicting instructions, and implicit assumptions frequently force models to guess, interpolate, or “fill in” missing details. Due to their compliant nature, LLMs rarely request clarification. Instead, they attempt to produce confident answers even when the prompt lacks essential information, which can distort meaning or misrepresent user intent.

This thesis focuses specifically on these *user-sided, prompt-time sources of faithfulness drift*. Rather than treating hallucinations only after they appear in the output, we adopt a *shift-left* perspective: identifying and mitigating hallucination potential at the level of the prompt itself, *before* generation occurs.

Bridging: Where Hallucinations Come From

Hallucinations arise from several sources across the LLM pipeline: data and pretraining artefacts, objective and reward misspecification, retrieval and knowledge limitations, and prompt design issues [6].

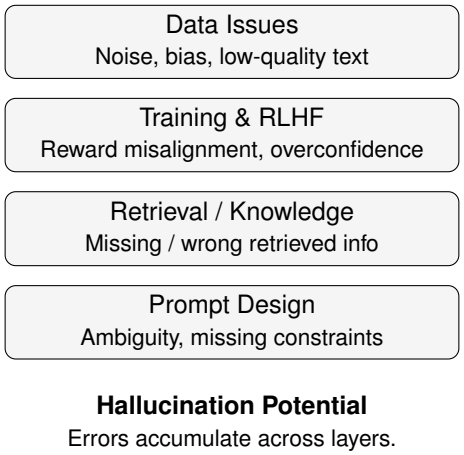
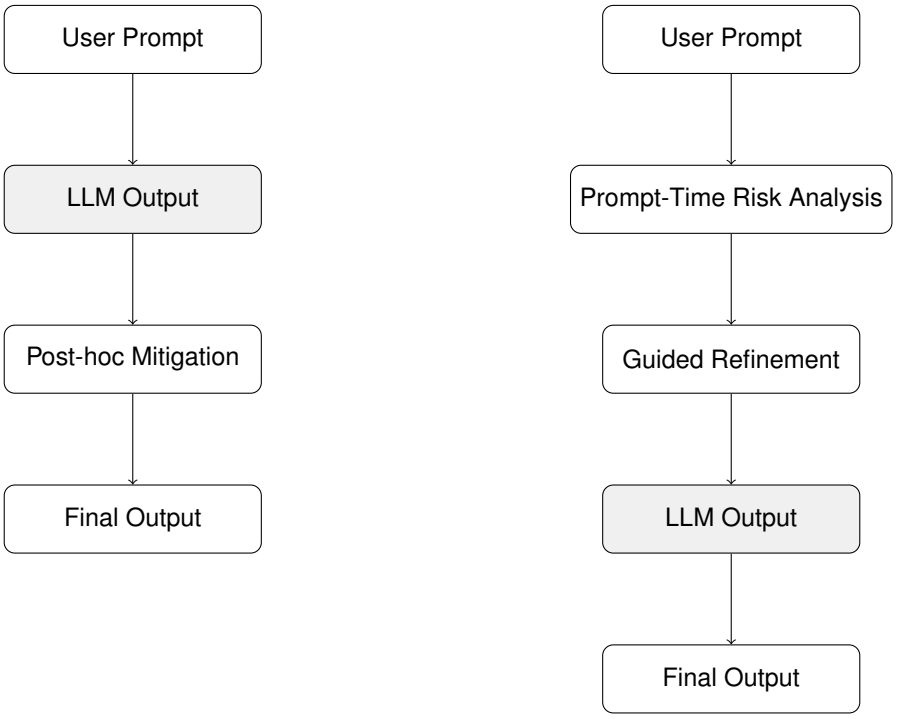


Figure 1.1 Hallucination sources across independent layers of the generation pipeline.

As illustrated above, different stages contribute to different kinds of errors. Model- and data-side factors primarily drive *factual* hallucinations, whereas prompt-related issues like ambiguous goals, missing constraints, implicit premises, and instruction conflicts are a major source of *faithfulness* deviations [5].

Prompt-time failures are particularly important because they occur *before* generation and cannot be repaired by downstream mitigation alone. Accordingly, although we briefly discuss factual risks (e.g., uncertainty cues, citation requirements), the focus of this thesis is on identifying and mitigating *prompt-related faithfulness risks*. We formalize this problem in section 1.2.

However, identifying where hallucinations originate is only half of the picture. The other half concerns *when* interventions act. While most hallucination mitigation approaches operate *after* model generation (e.g., RAG validation, post-hoc detectors, output rewriting), these methods remain reactive and cannot fix ambiguities already present in the prompt. In contrast, this thesis adopts a *shift-left* perspective: hallucination risks are surfaced and corrected *before* generation, directly at the prompt level.



Conventional (After-Generation).
Hallucinations appear in the model *output*.
Risk is detected *after* an incorrect answer is produced.

Shift-Left (Prompt-Time).
Risks are surfaced *before* generation by exposing unclear
or conflicting prompt segments (hallucination *potential*).

Figure 1.2 Comparison of after-generation mitigation (left) and shift-left, prompt-time mitigation (right).

As the figure illustrates, the dominant mitigation strategies act too late in the pipeline: they evaluate or correct the output, but they do not address the prompt segments that created the hallucination potential in the first place.

This observation leads directly to the problem statement in the following section.

1.2 Problem Statement and Objectives

The essence of the problem is a structural mismatch: users know what they mean, but the prompt rarely expresses this fully. Any unspoken assumption becomes a place where the model must infer intent, which introduces faithfulness hallucinations.

The following figure visualizes this gap.

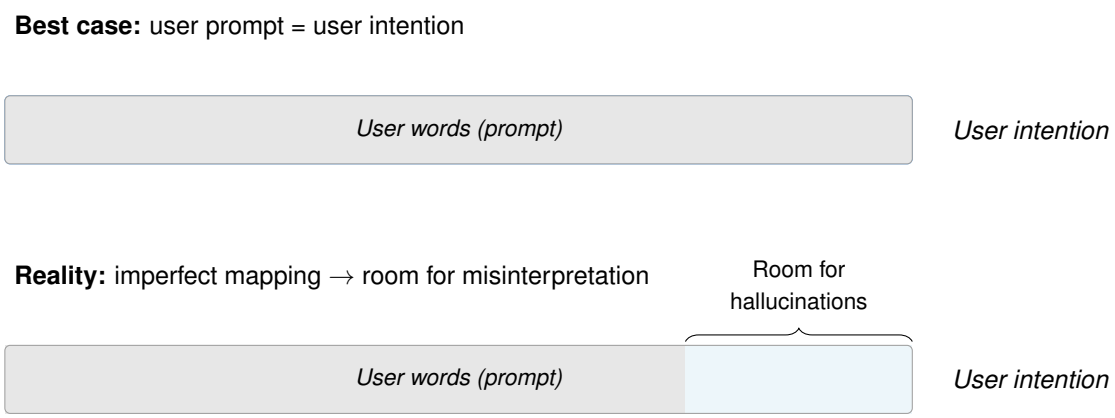


Figure 1.3 Prompt–intention alignment. Top: ideal alignment. Bottom: realistic right-side gap that can induce hallucinations; left underspecification is indicated subtly (grey) without call-out.

Problem. Faithfulness hallucinations are *majoritarily prompt-related*: ambiguity, imitative falsehoods [12] vague quantifiers, unverifiable asks, and instruction conflicts frequently cause instruction-deviating outputs [5], [6]. These issues arise at the surface form of the user prompt, yet current guidance (blog posts, best-practice lists, UI-based rewriting tools) remains largely *reactive*: they tell users *what* to do but rarely show *where* the prompt is underspecified or internally inconsistent.

Core idea. We introduce a *token-level* analysis that exposes prompt-surface defects with explicit criteria (e.g., ambiguous referents, missing units/sources, unverifiable claims), links each finding to spans, and provides guided rewrites. Instead of generic advice, users see *exactly which tokens* drive risk and how to fix them.

Reactive approaches are insufficient. Post-hoc checks, RAG, and prompt frameworks reduce some errors but act *after* generation, add latency/cost, and cannot recover information that was never specified. When the prompt lacks clarity, structure, or constraints, no downstream mitigation can guarantee instruction fidelity [6], [11].

Objectives.

O1) Develop a taxonomy of faithfulness-related prompt risks and a metric for estimating prompt risk severity.

O2) Detect faithfulness-risk patterns at the token level *before* generation.

O3) Provide guided, actionable fixes derived from explicit span-level analysis.

O4) Support iterative prompt refinement through a structured, role-based interaction loop.

Success criteria.

S1) Accurate detection of hallucination-prone patterns at the token level.

S2) Increased instruction alignment and constraint compliance relative to baseline prompts.

S3) Improved user understanding of prompt-specification pitfalls and failure modes.

S4) Maintain or improve output quality while reducing prompt-time risk.

Artifact teaser. *Echo* is a lightweight, model-agnostic assistant that runs pre-generation: it detects risks, highlights spans, and proposes guided refinements with transparent, deterministic scoring (see chapter 3).

1.3 Design Science Research Lens

Environment. The environment of this research includes practitioners who design natural language prompts for large language models across domains such as healthcare, law, and finance, as well as learners and researchers seeking to understand how prompt formulation affects faithfulness and output quality. Practitioners often lack formal training in prompt engineering and must work under constraints of latency, cost, and domain-specific compliance, while students and researchers face the challenge of reasoning about how prompt structure interacts with model behavior.

Across all these groups, tasks demand high factual and instructional faithfulness, as errors can propagate into medical misinformation, legal misjudgment, or financial risk (See 1.1). Current workflows provide little prompt-time guidance, leaving ambiguity and unverifiable instructions unaddressed, which motivates the need for a proactive instead of a reactive artifact [7], [13].

Knowledge base. This work builds on existing surveys and taxonomies of hallucinations and prompt-risk factors [5], [6], [11]. Prior research identifies risk categories but offers limited operationalization at the token span level. No unified guideline set exists for consistent, model-agnostic detection of prompt-related faithfulness risks, which motivates the construction of a formal taxonomy and rule set.

Artifact. *Echo* is a multi-agent system that operationalizes the faithfulness-risk taxonomy into token-level, prompt-time detection and iterative, guided refinement.

Utility goal. Reduce hallucination risk *at the source* by enforcing clarity and constraints, thereby aligning instructions before inference rather than attempting to repair outputs post hoc.

Process. We follow a Design Science Research cycle comprising problem identification, objective definition, system design and development of the artifact followed by demonstration through realistic, domain relevant scenarios, and evaluation. The artifact and its underlying taxonomy were iteratively refined following the design space exploration methodology, and the resulting knowledge aims to improve prompt reliability and faithfulness.

1.4 Application of DSR in This Thesis

1.4.1 Problem Identification & Motivation

Prompt ambiguity is both prevalent and consequential in high-stakes settings: legal systems wrongly cite or fabricate authorities, and medical summarization exhibits non-zero error/omission rates even under curated conditions [7], [8], [13]. Stakeholders include developers, analysts, clinicians, legal professionals, and students who rely on long, sensitive prompts and require faithful, instruction-aligned model behavior without the need for multiple testing iterations. This thesis therefore focuses on prompt-surface risks: ambiguity,

vagueness, conflicting instructions, or unverifiable asks while full fact verification, adversarial robustness, and model retraining remain out of scope.

1.4.2 Objectives of a Solution

The objectives follow directly from the challenges described in the environment (section 1.3). They are grouped into functional requirements, non-functional requirements, and measurable targets guiding evaluation.

Functional objectives. The artifact must (F1) detect hallucination-prone spans at the token level using transparent, reproducible criteria; (F2) provide actionable rewriting suggestions that help users improve their prompts; (F3) ensure token span to risk traceability so that each highlighted risk is linked to an explicit guideline and (F4) support an iterative refinement loop in which users analyze, revise, and re-analyze prompts.

Non-functional objectives. The solution must satisfy several quality constraints: (N1) *usability*, achieved via progressive display so that users are not overwhelmed; (N2) *determinism*, ensuring consistent scoring under repeated runs and preventing model drift from affecting rule application; (N3) *transparency*, so that risk assessments remain interpretable; and (N4) *modularity*, allowing new taxonomies, models or agents to be newly integrated with minimal friction.

Measurable targets. Success is assessed through reduction in hallucination-prone spans after refinement, increased compliance with explicit user constraints, and maintenance or improvement of the overall output quality relative to baseline prompts.

1.4.3 Design & Development (Echo)

Rationale. Surfacing risks before generation rather than repairing them post hoc which reduces downstream cost and improves reliability.
Principles. Echo is designed around determinism (stable scoring), progressive disclosure (from token span highlighting to overall metrics to span-level evidence), and traceability (each warning cites the rule and the affected token span).
Linkage. All criteria derive from the Prompt/Meta taxonomy introduced in earlier sections (section 2.1, chapter 4) and are operationalized through the XML guideline set described in chapter 3.

1.4.4 Demonstration

The artifact is demonstrated on prompts from legal, medical, and financial scenarios, as well as developer and student use cases. Utility is shown by token span highlighting and before vs after comparisons: reduced risky spans, clearer instructions, and higher constraint alignment, without dependence on a specific model provider (see chapter 3).

1.4.5 Evaluation

We evaluate the tagging quality of Echo compared to a human tagger using values such as recall, precision, and ablation (for lexically altered texts) as well as before/after qualitative comparisons between outputs of the same model when faced with the original prompt versus the refined one in chapter 5.

1.4.6 Communication

The complete implementation of ECHO, including code, guideline sets, configuration files, and evaluation prompts, is available in the public repository[1].

1.5 Research Questions

This thesis investigates user-sided, prompt-induced faithfulness risks for large language models through three research questions.

RQ1 — Landscape and Gap

Which types of user-sided prompt risks that can lead to faithfulness-related hallucinations are described in existing research and practitioner guidelines, and to what extent are these risks already organised into a structured, operational taxonomy?

RQ1 is addressed through a structured literature and practitioner review that collects and categorises prompt-related risk patterns from academic work, provider documentation, and engineering blogs. The outcome is a consolidated view of user-sided risks and an explicit identification of missing or only partially structured taxonomies, motivating the need for Echo's taxonomy.

RQ2 — Taxonomy and Detection

Can these literature-derived risks be consolidated into a two-dimensional prompt/meta taxonomy that supports reliable classification and token-span-level detection of user-sided faithfulness risks in real prompts?

RQ2 is addressed by deriving a prompt/meta taxonomy from the collected risks, encoding it as XML guidelines, and implementing Echo's analyser to perform span-level classification on a curated prompt set. Its answer is based on how well the taxonomy covers observed issues and on quantitative detection metrics (e.g., precision, recall, false positives/negatives) against manually annotated ground truth.

RQ3 — Refinement Effectiveness

Does Echo's interactive refinement loop, based on this taxonomy and its detections, measurably reduce prompt risk and qualitatively improve the faithfulness and completeness of user prompts compared to their original versions?

[1] https://github.com/MoNejjar/echo-hallucination-detect

RQ3 is addressed by evaluating Echo's multi-step refinement workflow on real prompts, comparing original and refined versions. The analysis combines quantitative changes in Prompt Risk Density (PRD) with qualitative before/after examples to assess whether prompts become more explicit, less ambiguous, and better aligned with the apparent user intent.

1.6 Contributions

This thesis makes the following contributions:

C1 A consolidated taxonomy of faithfulness-related prompt risks, structured into prompt-level and meta-level categories, together with explicit, token-level criteria for operational use [6], [11].

C2 ECHO, a model-agnostic, multi-agent artifact for prompt-time hallucination mitigation, combining span-level risk detection, guided refinement, and iterative clarification.

C3 A deterministic analysis and scoring pipeline with progressive disclosure and span-to-criterion traceability, enabling transparent inspection of hallucination-prone prompt segments.

C4 A quantitative and qualitative evaluation demonstrating improved instruction alignment, reduced risk-span density, and stable output quality across lexical variants.

C5 Prompt Risk Density (PRD), a length- and severity-normalized metric for estimating prompt-time faithfulness risk and supporting before/after refinement comparisons.

1.7 Scope, Assumptions, & Ethics

In scope. English/German text prompts; single-turn conversation-like interactions; developer/analyst workflows; token-level risk analysis and a guided approach to mitigating the hallucination prone sections.

Out of scope. Full fact verification, adversarial jailbreak defense, model retraining/fine-tuning, RAG.

Assumptions. API access or access to a hosted model; prompt length $<$8k tokens (to prevent the LLM from being lost in the context [14]); non-malicious use.

Ethics. No clinical/legal deployment claims; uncertainty is surfaced; limitations disclosed; datasets/prompts curated to avoid sensitive data [13].

1.8 Thesis Structure

In this thesis, Chapter 2.1 formalizes concepts and related work. Chapter 3 presents Echo's design based on the methodology in Chapter 4. Chapter 5 reports results and ablations. Chapter 6 reflects on implications and paves the way for future work. An appendix provides the public repository used for reproducibility.

2 Background and Related Work

2.1 Background & Theoretical Foundations

This chapter clarifies the conceptual foundations of hallucinations in large language models (LLMs) and situates this thesis within the landscape of detection and mitigation methods. We introduce the distinction between *faithfulness* and *factuality*, summarize major causes of hallucinations, and review three families of approaches: prompt-time controls, post-hoc retrieval and verification, and self-reflective or tool-augmented strategies. We then motivate why prompt-time defects deserve explicit treatment and identify gaps that lead to our focus on prompt-level risk detection.

2.1.1 Hallucination: Faithfulness vs. Factuality

Hallucinations are fluent but unfounded model outputs. Modern surveys separate two complementary dimensions [6]. **Factuality** concerns alignment with verifiable real-world information (external knowledge). A response is factually hallucinated if it asserts something false or unsupported by external evidence. **Faithfulness** concerns adherence to the given instructions or source context. A response is unfaithful if it contradicts the prompt, ignores constraints, or introduces details not derivable from the provided material.

Recent work further refines faithfulness into instruction inconsistency, context inconsistency, and logical self-contradiction [6]. Thus, hallucinations may arise because the world model is wrong (factuality) or because the instruction-following behavior is wrong (faithfulness).

Causes. Two widely observed mechanisms are (i) the “yes-man” effect: LLMs are trained to produce plausible continuations rather than calibrated uncertainty, leading to confident answers even under ambiguity [5]; and (ii) knowledge limitations: models interpolate beyond their training data, especially in domains such as law or medicine where coverage is sparse or outdated [7]. Further failures arise from vague, ambiguous, or conflicting prompts which provide insufficient structure for reliable instruction following. Because this thesis focuses on these *user-side* factors, the next subsection formalizes why prompt-time risks deserve explicit treatment.

2.1.2 Why Prompt-Time Risks Matter

Ambiguity, vague constraints, missing units, and conflicting instructions are prevalent sources of *faithfulness* errors because they create *degrees of freedom* in the task specification. When the prompt does not fully pin down referents, scope, baselines, or output constraints, the model must choose an interpretation. Those choices can be fluent yet misaligned with the user’s intent, producing instruction-inconsistent content.
Crucially, these failures originate in the *input* rather than the model’s parameters. They are therefore partially *controllable*: for many tasks, clarifying the prompt reduces the need for the model to guess missing structure, and decreases the surface area for unfaithful interpretations.

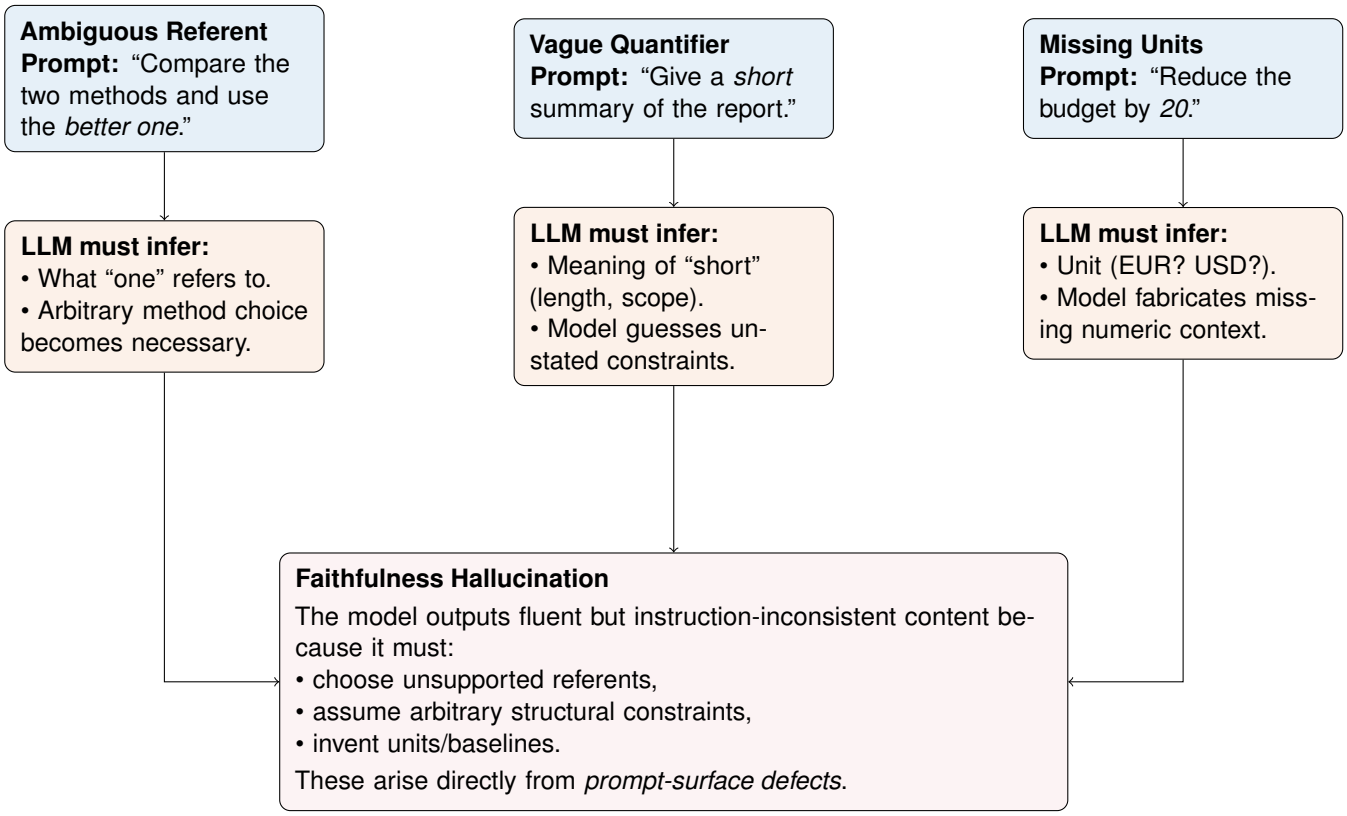


Figure 2.1 Prompt-surface defects (top, blue) force the model to infer missing information (middle, orange), producing faithfulness hallucinations (bottom, red).

Existing approaches like retrieval augmented generation (RAG) [15]–[17], post-hoc fact-checking , or model self-reflection [18], [19] operate *after* a response is generated. They improve reliability, yet none explicitly help users identify *where* their prompt is underspecified or likely to induce an unfaithful interpretation. This thesis adopts a shift-left approach: treat the prompt as a diagnosable artifact, surface risk-bearing spans at token granularity, and guide users in iterative prompt repair.

2.2 Detection & Mitigation Approaches

Prior work addressing hallucinations can be grouped into three families: (1) prompt-time controls, (2) retrieval and post-hoc verification, and (3) multi-pass reasoning and tool use. Each mitigates certain failure modes but leaves prompt-surface risks insufficiently treated.

2.2.1 Prompt-Time Controls & Guardrails

Prompt-time control frameworks improve alignment by constraining or structuring the *input*. Patterns such as “Role–Task–Output” scaffolds [20] and constraint-explicit prompting [21] reduce underspecification by clarifying scope, format, and audience. Declarative prompt programming frameworks (e.g., DSPy [22]) encode prompts as structured programs with deterministic parameters, improving traceability.

Guardrail systems (e.g., Guardrails AI [23]) validate inputs or outputs against rule sets. While effective for global safety conditions, these tools operate at a coarse granularity and do not identify *token-level* ambiguities inside the user’s prompt.

2.2.2 Post-hoc Fact-Checking & Retrieval

Retrieval-Augmented Generation (RAG) grounds model outputs in external documents which reduces factual hallucinations [15], [16]. However, retrieval does not fix an ambiguous or malicious prompt: if the query itself is vague or conflicting, retrieved results may be irrelevant or misused [24]. Post-hoc fact-checkers evaluate generated answers for support, but they intervene only *after* the model has produced a potentially flawed output. This adds latency and cannot recover missing constraints that should have been specified upfront.

2.2.3 Self-Consistency, Critique, & Tool Use

Self-consistency works by generating multiple reasoning paths and comparing them, helping the model filter out unstable or incorrect answers [25]. Reflective mechanisms such as Reflexion [18] or CriticGPT [26] prompt models to review and improve their own outputs. Tool-use methods (e.g., ReAct [27]) ground intermediate steps in external computation.

These strategies mitigate reasoning errors but remain *reactive*: they correct or evaluate outputs rather than addressing the *user prompt* as a potential source of risk. They therefore complement but cannot replace a proactive prompt-time diagnostic framework.

2.2.4 Comparative Summary and Boundary of This Thesis

Across the three families, prior work improves reliability at different stages of the pipeline. Prompt-time controls reduce underspecification by imposing structure in the instruction channel [20]–[22]. Retrieval and verification ground or validate claims against external evidence [15], [16], [28]. Multi-pass reasoning, critique, and tool use stabilize complex inference by sampling or revising reasoning trajectories and grounding intermediate steps in deterministic operations [18], [19], [25]–[27], [29].

A shared limitation is that these approaches rarely localize *where* the user prompt is underspecified. Templates and guardrails typically validate high-level constraints, retrieval and verification intervene after a response exists, and reflection primarily revises outputs rather than diagnosing prompt spans that force the model to infer missing referents, baselines, units, or task structure. This motivates the shift-left view taken in this thesis: prompts are treated as diagnosable artifacts, risks are surfaced at token granularity, and users are guided in iterative prompt repair.

2.3 Taxonomies of Hallucinations

Existing taxonomies distinguish intrinsic (faithfulness) and extrinsic (factuality) hallucinations [30]. Broader surveys categorize causes across data, training, inference, and prompting [6], [11], while theoretical analyses explain why hallucinations persist even as models scale [5].

A persistent gap is the underdevelopped treatment of *prompt-surface risks*. Existing frameworks acknowledge that ambiguous or underspecified prompts can induce hallucinations but do not define a comprehensive, rule-based taxonomy for identifying them.

This thesis fills that gap by developing a two-tier *Prompt/Meta Risk Taxonomy* (introduced in Chapter 4) that differentiates:

- prompt-localizable risks (e.g., vague quantifiers, ambiguous referents), and
- structural/meta risks (e.g., missing scope, multi-objective overload).

This taxonomy provides the foundation for token-level detection, span-level annotation, and the Prompt Risk Density (PRD) metric used throughout this thesis.

2.4 Evaluation Protocols and Benchmarks

Hallucinations are most commonly evaluated at the level of *model outputs*. In open-domain QA and truthfulness settings, benchmarks measure whether answers align with reference truth or resist common misconceptions, often using accuracy, EM/F_1, or rubric-based truthfulness scores [12], [31], [32]. In summarization, evaluation emphasizes whether generated statements are supported by the source, using human judgments, entailment-style checks, or QA-based metrics such as QUESTEVAL [30], [33], and meta-evaluation work highlights the fragility and failure modes of automatic factuality metrics [34]. For claim verification, datasets such as FEVER combine label correctness with evidence sufficiency [35]. Broader hallucination suites cover multiple task arenas, including dialogue, summarization, and QA [36], [37]. Instruction-following datasets further probe constraint satisfaction and adversarial failure modes [38], [39].

A complementary line of work evaluates hallucination risk through *uncertainty or self-agreement* signals. Semantic entropy measures dispersion of meaning across multiple samples and correlates high dispersion with hallucination likelihood [40]. Self-checking methods such as SELFCHECKGPT compare generations to detect inconsistency without requiring access to logits or an external knowledge base [41]. Contextual analyses also show that long prompts and effective context limitations can degrade reliability even when the nominal context window is large [14].

While these protocols are well suited for quantifying *output correctness* and downstream task performance, they provide limited visibility into *input quality*. When a prompt is ambiguous or underspecified, output-level metrics rarely indicate which spans introduced risk or how a user should repair the prompt. This thesis therefore complements output-oriented evaluation with prompt-side measurement: span-level risk localization and the Prompt Risk Density (PRD) metric, which quantify prompt risk directly and support targeted, iterative prompt repair.

2.5 Why Hallucinations Matter in High-Stakes Practice

Hallucinations are not only a benchmark phenomenon; they have measurable consequences in high-stakes workflows. In law, studies document fabricated citations and unreliable behavior under realistic legal queries, motivating strong safeguards and careful user interfaces [7], [8]. In medicine, clinical summarization and reporting studies highlight the need to assess hallucination rates and safety risks before deployment, particularly when outputs may influence downstream decisions [3], [9], [13]. In finance and decision support, work shows that model outputs can systematically reinforce biases and increase risk exposure, even when

responses appear coherent [10].

These findings reinforce the thesis motivation: reliability is shaped not only by model capabilities but also by the quality of the prompt that initiates the interaction. In practice, users frequently issue underspecified prompts under time pressure. A diagnostic framework that localizes prompt risks and guides repair can therefore serve as a low-friction "front line" complement to heavier-weight mitigations such as retrieval, verification, or multi-pass critique.

2.6 Gaps and Relevance to This Thesis

The literature indicates several gaps that motivate our focus on prompt-side diagnosis:

- **Gap 1: Limited prompt-time prevention.** Much work focuses on improving or verifying *outputs* after generation through retrieval or verification, rather than diagnosing defects in the input itself [15], [16], [28].
- **Gap 2: Limited token-level traceability.** Many taxonomies and tools operate at sentence or task level, which limits directed feedback for repairing specific prompt spans [6], [11].
- **Gap 3: Fragmented treatment of local vs. structural prompt risks.** Prompt ambiguity and broader structural misalignment are often discussed separately, complicating unified diagnosis [6].
- **Gap 4: Few quantitative prompt-side metrics.** Standard evaluations prioritize output correctness and uncertainty signals rather than normalized measures of prompt risk [12], [40], [41].
- **Gap 5: Limited user-facing instructional tooling.** In practice, many mitigations require expert configuration or multi-pass pipelines, providing limited in-situ guidance for end users to repair prompts [20], [22].

This thesis addresses these gaps through a rule-based Prompt/Meta Risk Taxonomy, a token-level detection and refinement workflow, and the Prompt Risk Density (PRD) metric, implemented and evaluated in the ECHO artifact.

2.7 Chapter Summary

This chapter surveyed the conceptual foundations of hallucinations in large language models and reviewed the major strands of mitigation research. We distinguished factuality from faithfulness, examined why LLMs produce ungrounded or unfaithful outputs, and highlighted the role of prompt-surface defects such as ambiguity, vague constraints, and missing units as under-theorized causes of faithfulness hallucinations.

Across prompt-time controls, retrieval augmentation, post-hoc verification, and multi-pass reasoning, prior work offers valuable mitigation techniques but provides limited guidance for diagnosing *where* user prompts introduce risk. Existing taxonomies often lack a span-localizable, rule-based operationalization, and evaluation protocols largely target output quality rather than prompt quality.

These gaps motivate a shift-left perspective in which prompts are treated as analyzable artifacts. The next chapter develops the methodological foundations for this approach: a rule-based Prompt/Meta Risk Taxonomy, operational guidelines for deterministic span- level detection, and a quantitative Prompt Risk Density (PRD) metric that together underpin the ECHO artefact and its evaluation.

3 System Design: Echo

3.1 Chapter Purpose

This chapter documents the *design and evolution* of Echo as an artefact for proactive, prompt-time mitigation of faithfulness risks. Unlike a traditional system specification that presents a finalized architecture, we adopt a **design science** perspective [42] and trace the path from initial concepts through failed iterations to the final system. Accordingly, this chapter traces the evolution of Echo's taxonomy, workflow, and multi-agent architecture through successive design iterations rather than presenting only a finalized system.

The chapter serves three purposes:

1. **Research transparency**: document design decisions together with explicit rationales and discarded alternatives.
2. **Traceability**: link design choices back to the research questions (Section 1.5) and evaluation metrics (Chapter 5).
3. **Reproducibility**: Provide sufficient detail for independent implementation, and release the Echo system as an open-source artifact to enable inspection, verification, and reuse by other researchers.

In addition to documenting the design process, this chapter also clarifies the role of *Echo* as the artifact through which the research contribution is made observable.

Echo functions as an integrative research artifact: a single environment in which the taxonomy, guidelines, rule explanations, and iterative refinement mechanism are made accessible. This consolidated interface allows the reader and evaluator to inspect the behavior of the taxonomy directly in the library function of the application and to reproduce the analyses and the evaluation using the chat interface. In this sense, Echo acts as both the demonstrator and the empirical lens of the thesis.

Chapter organization

- Section 3.2 defines design goals, success criteria, and scope.
- Section 3.5 describes the design space exploration and failed iterations.
- Section 3.6 motivates the main design decisions and contrasts them with alternatives.
- Section 3.7 presents the final system architecture and agent responsibilities.
- Section **??** explains the data flow and state management across the pipeline.
- Section **??** describes the user interaction flow.
- Section 3.8 summarises architectural limitations and non-goals.

3.2 Design Objectives and Success Criteria

3.2.1 Design Goals

Echo is designed to address the three research questions (RQ1–RQ3, Section 1.5) through the following design goals:

1. **DG1 (Addresses RQ1 & RQ2):** Operationalize a structured, taxonomy-based view of user-sided hallucination risks.
   - Adopt a two-actor perspective in which hallucinations originate either from the LLM itself or from the user through the prompt. This motivates Echo's focus on user-sided risks.
   - Distinguish between two classes of user-sided causes: **prompt-related** risks that can be localized to token spans, and **meta-related** risks that arise from missing, conflicting, or structurally defective information not localizable to specific tokens.
   - Maintain compatibility with existing hallucination typologies by recognizing that both prompt- and meta-level issues can manifest as either **faithfulness** hallucinations (misinterpretations of user intent/context) or **factuality** hallucinations (incorrect world-knowledge claims).
   - Surface PROMPT risks as highlightable spans with a 1:1 token-to-risk mapping, and META risks as global warnings requiring structural analysis.

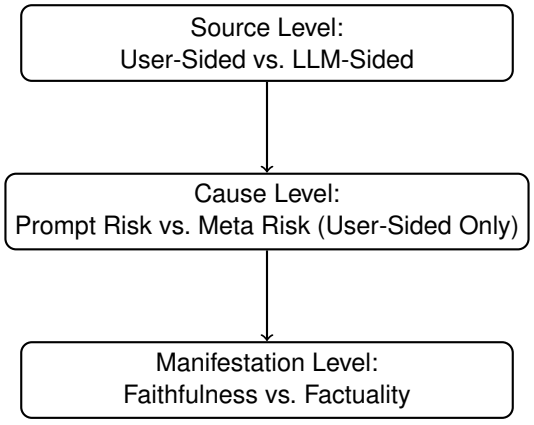


Figure 3.1 Abstraction layers motivating the taxonomy: hallucination *sources* (user vs. LLM), user-sided *causes* (prompt vs. meta), and downstream *manifestation types* (faithfulness vs. factuality).

2. **DG2 (Supports RQ2):** *Provide quantitative risk assessment before generation.*
   - Compute Prompt Risk Density (PRD), a severity-weighted, token-level measure that normalises total risk by prompt length. PRD expresses how much of the prompt is occupied by risky spans before generation.
   - Provide a stable, interpretable signal that allows investigation of whether higher PRD values are associated with an increased likelihood of hallucinations in subsequent model outputs.
   - Enable before vs after PRD value comparisons to measure the effect of refinements quantitatively.
3. **DG3 (Supports RQ3):** *Enable iterative, user-driven refinement rather than one-shot rewriting.*
   - Generate targeted clarifying questions (instead of silently rewriting the prompt).

- Preserve user intent while mitigating detected risks.

- Support re-analysis to validate that high-severity issues have been addressed and deal with potential remaining false negatives.

4. **DG4 (Cross-cutting):** *Ensure traceability and reproducibility.*

   - Each detected risk is assigned an explicit guideline rule ID and severity.

   - Scoring is deterministic given the same analysis output as it is calculated computationally, enabling stable evaluation.

   - The same prompt fed into the same configuration yields the same annotated output.

5. **DG5 (Cross-cutting; educational):** *Promote user learning.*

   - Expose rules, explanations, and examples in an integrated guideline library so that Echo not only fixes prompts but also teaches users how to avoid faithfulness risks in future prompts through constant references to rules and a dedicated library in the UI for guidance.

   - Although not part of the primary research questions, Echo's educational input supports long-term user understanding of RQ1's taxonomy and contributes to sustained reduction of faithfulness risks.

3.2.2 Success Criteria

The success of the design is evaluated through criteria that reflect Echo's intended purpose as a diagnostic, refinement, and educational artifact. These are not universal benchmarks; rather, they capture thesis-specific evidence that the system is usable, operational, and aligned with the taxonomy and design goals.

Criterion	Target (Thesis-Level)	Measurement	Rationale
SC1: Detection quality	Echo identifies a substantively meaningful portion of gold risk spans	Token-/span-level agreement with human annotations; qualitative inspection of typical successes and misses	Demonstrate that the taxonomy can be operationalised into a practical detector (supports RQ2).
SC2: Detection stability	Outputs remain broadly consistent under minor prompt rephrasings	Divergence between analyses of lexical variants of the same prompt; inspection of where instability arises	Ensure the detector is not overly sensitive to superficial changes and reflects structural risks rather than wording noise.
SC3: Refinement effectiveness	Prompts show reduced pre-generation risk and yield clearer, more faithful LLM answers	(i) Reduction in PRD after refinement; (ii) Blind qualitative comparison of LLM outputs for original vs. refined prompts	Show that iterative refinement improves both prompt structure and downstream answer quality (supports RQ3).
SC4: False-positive reasonableness	Echo avoids overflagging obviously clean prompts	Specificity estimated on manually validated clean prompts; qualitative review of false-positive patterns	Preserve usability—excessive warnings undermine trust and overwhelm users.
SC5: User learning	Users improve prompting literacy and understand the rationale behind required edits	Evidence of rule-linked feedback being followed; occasional use of the integrated guideline library	Supports DG5.

Table 3.1 Design success criteria aligned with the thesis goals and evaluation plan.

3.2.3 Scope and Assumptions

In scope

- **Risk type.** Detection and mitigation of *prompt-induced faithfulness risks* that arise from ambiguity, missing information, conflicting instructions, or structural defects in user prompts. Factuality risks appear in the taxonomy and UI but are not empirically evaluated.
- **Operationalisation of the taxonomy.** Full implementation of the Prompt/Meta taxonomy as an executable XML guideline set, covering rule IDs, pillars, severities, and span-level traceability.
- **Span-level analysis.** Extraction of prompt-related risks as token-indexed spans, plus structured meta-risk warnings. This includes deterministic post-processing, severity clamping, and reconstruction of annotated prompts.
- **Interactive refinement.** A multi-turn refinement workflow (Analyzer $\rightarrow$ Initiator $\rightarrow$ Conversation $\rightarrow$ Re-analysis $\rightarrow$ Preparator) that elicits missing context, resolves ambiguities, and integrates user-guided edits.

- **Prompt-time metrics.** Use of Prompt Risk Density (PRD) as a descriptive, length-normalised measure of risk concentration and as a before/after indicator during refinement.
- **Evaluation setting.** Span-level agreement against a gold set (single annotator) and qualitative QA comparisons using several English, text-only prompts ranging from short instructions to full system prompts.
- **Model usage.** A GPT-5–class model is used for all agents. The guideline format is model-agnostic, and alternative LLMs were tested informally for compatibility.
- **Interaction model.** Single-user, session-based usage without assumptions about user expertise; Echo supplies questions and alternatives to minimise user friction.

Out of scope

- **Deterministic detection guarantees.** Despite structured XML output, low-temperature decoding, and post-processing, LLM-based detection is inherently non-deterministic. Full output invariance across runs or model versions is not targeted.
- **PRD–hallucination causality.** PRD is used solely as a descriptive, prompt-time metric. This thesis does not establish a calibrated or causal relationship between PRD values and actual hallucination probabilities. Investigating this link remains an open and promising direction for future research.
- **Factuality verification.** Echo does not access external knowledge sources or fact-checking tools and therefore does not mitigate factuality hallucinations caused by incorrect world knowledge.
- **LLM-induced errors.** Errors originating from the underlying model (architecture, training data, decoding) are out of scope; Echo targets only user-sided risks.
- **Multimodal inputs.** Images, audio, video, and code are not analysed.
- **Model infrastructure.** Automatic model switching, multi-provider support, or local model hosting are not addressed.
- **User studies.** No controlled usability or longitudinal learning studies are performed; claims about long-term educational impact are not made.
- **Generalisation across domains.** Evaluation focuses on selected English QA tasks; broader cross-task, cross-language, or domain-specialised generalisation is beyond scope.
- **Inter-annotator reliability.** Gold annotations were produced by a single annotator; no inter-annotator agreement (IAA) statistics are available.

3.3 High-Level Solution Overview

Before presenting the detailed design choices and the final system architecture (Sections 3.7), this section provides a concise overview of Echo as an artifact. Its purpose is to bridge the gap between the design goals in Section 3.2 and the taxonomy-operationalisation details in Section 3.4 by introducing the system's overall workflow, the role of its agents, and the conceptual data flow that structures all subsequent design decisions.

Echo implements a shift-left, prompt-time hallucination mitigation workflow. Unlike most hallucination-detection systems, which analyze *outputs*, Echo acts *before generation*, treating the initial user prompt as a structured object that can be analyzed, annotated, and refined. To make this possible, Echo combines (i) a rule-based taxonomy encoded as XML guidelines, (ii) a multi-agent pipeline that analyzes, discusses and finally transforms the prompt, (iii) a dedicated section for a 1:1 mapping between token spans and hallucination risks that allows the user to evaluate the risk factor of a prompt before generation and (iiii) quantitative metrics such as Prompt Risk Density (PRD) that allow users to evaluate improvements across refinement cycles.

3.3.1 The Multi-Agent Pipeline

Echo is organised as a sequential, four-agent architecture:

1. **Analyzer Agent — Analysis.** Performs span-level and meta-level risk identification based on the XML taxonomy. Outputs an annotated prompt with <RISK_n> tags, structured violation objects with associated risk catageories and mitigation strategies, and PRD fields calculated deterministically based on the amount and severity of identified risks.

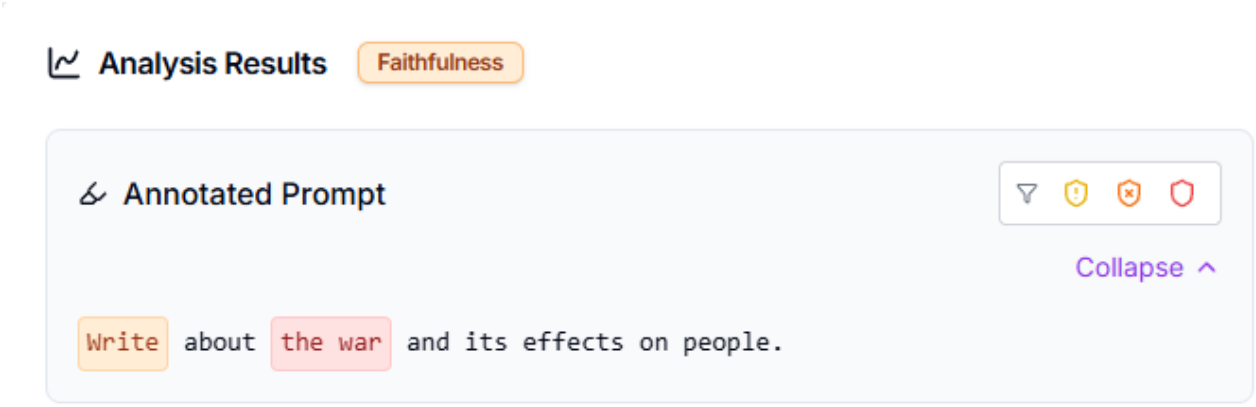


Figure 3.2 Echo UI: annotated prompt with span-level highlights.

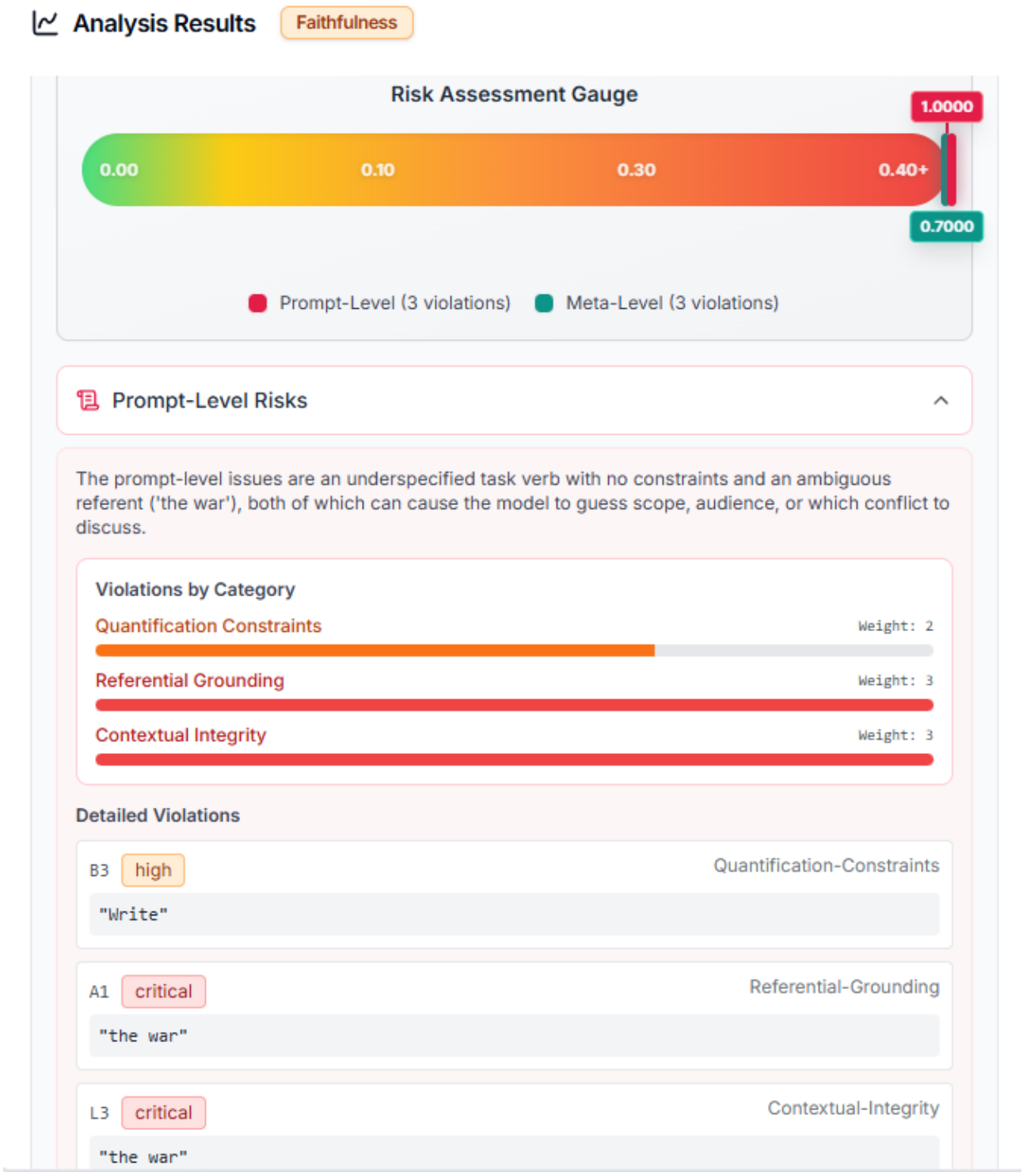


Figure 3.3 Echo UI: PRD gauge as well as an enumeration of prompt risks.

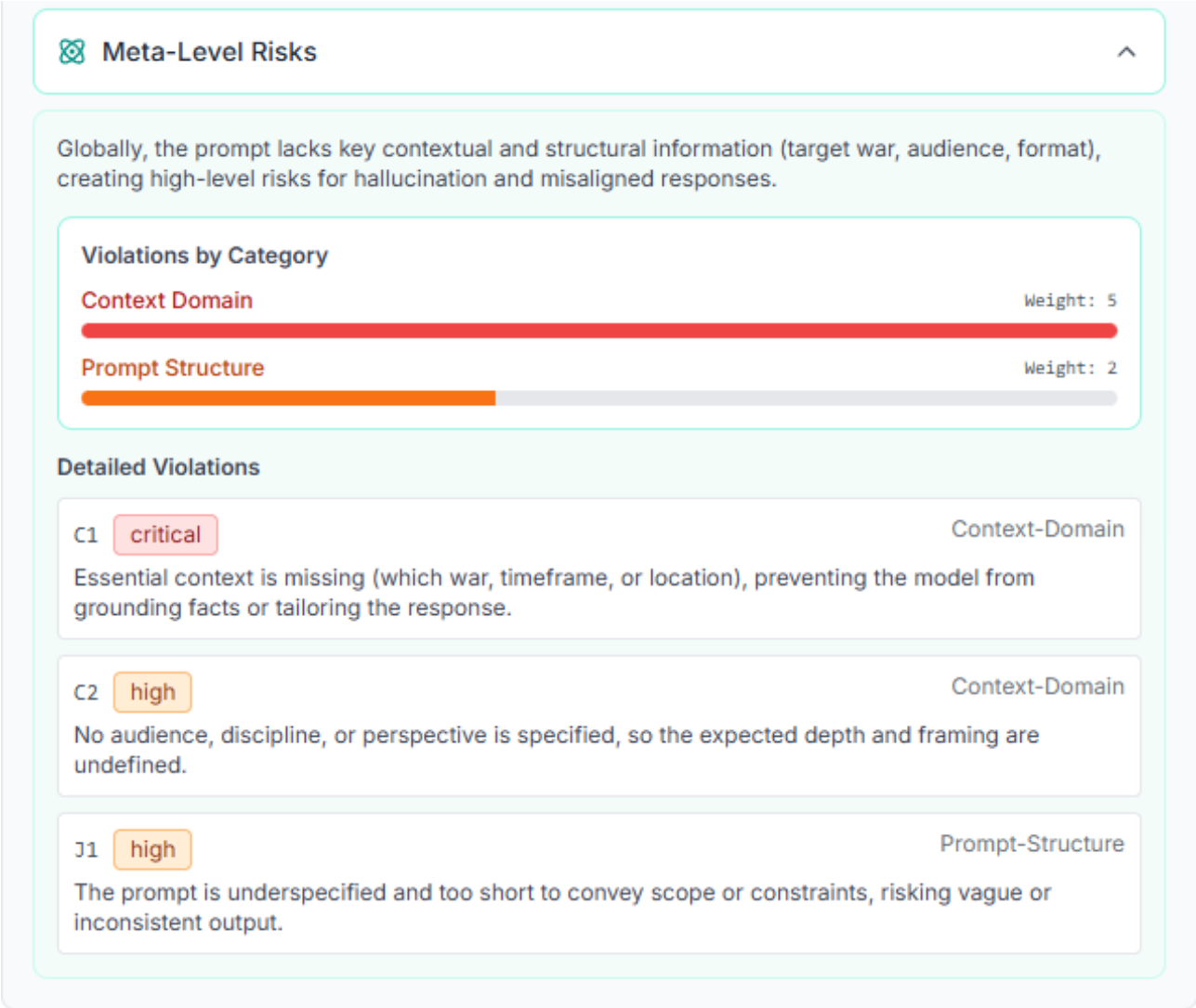


Figure 3.4 Echo UI: enumeration of global, meta risks.

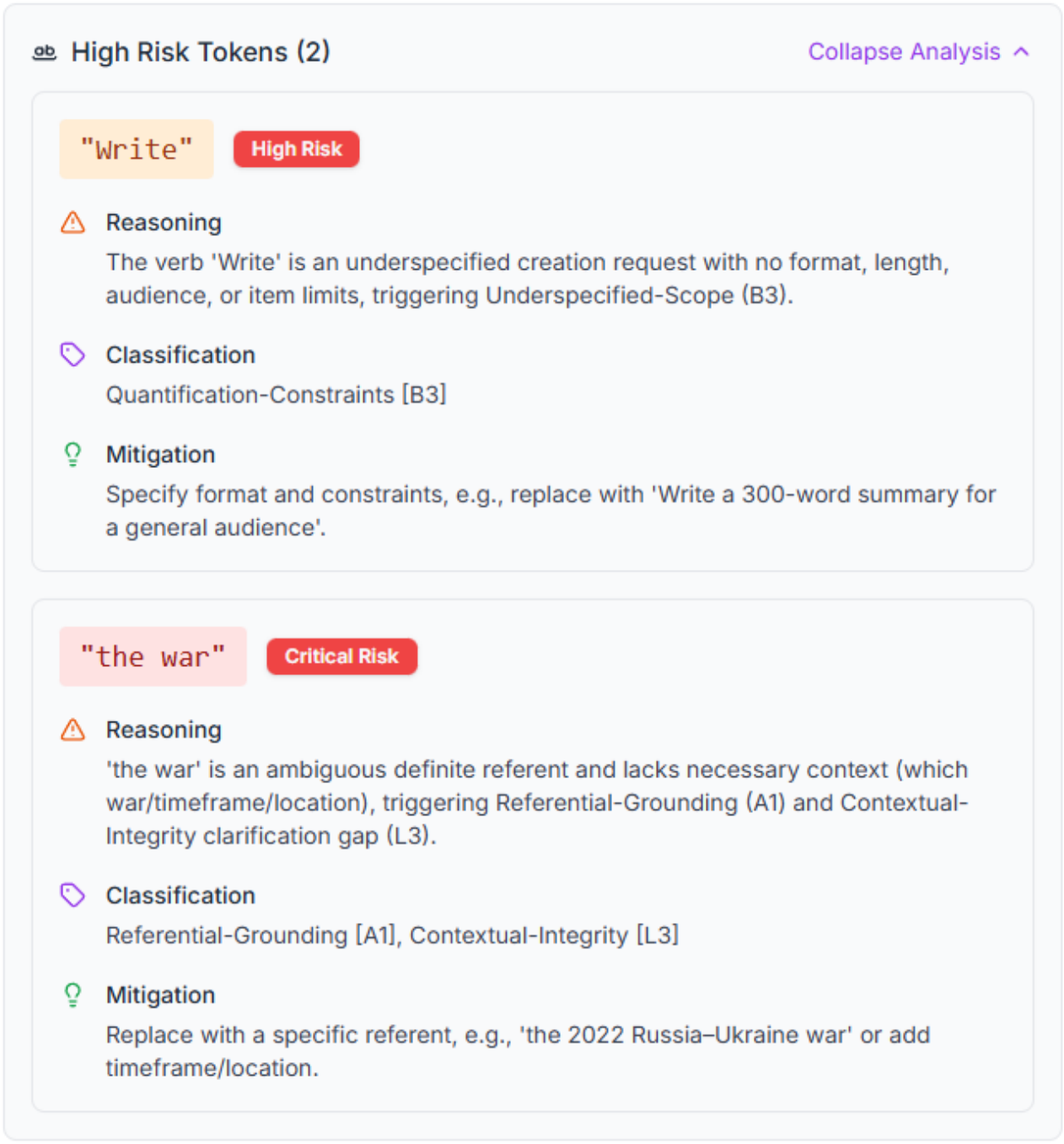


Figure 3.5 Echo UI: 1:1 mapping between token spans and hallucination risks with the risk category and a mitigation strategy.

2. **Initiator Agent — Clarification.** Transforms detected violations into targeted clarification questions. Each highlighted token span is turned into a specific question that aims at filling the gaps left by the initial user prompt. This is then followed by a section explaining to the user how answering these questions directly beneifts the final prompt and indirectly the LLM output.

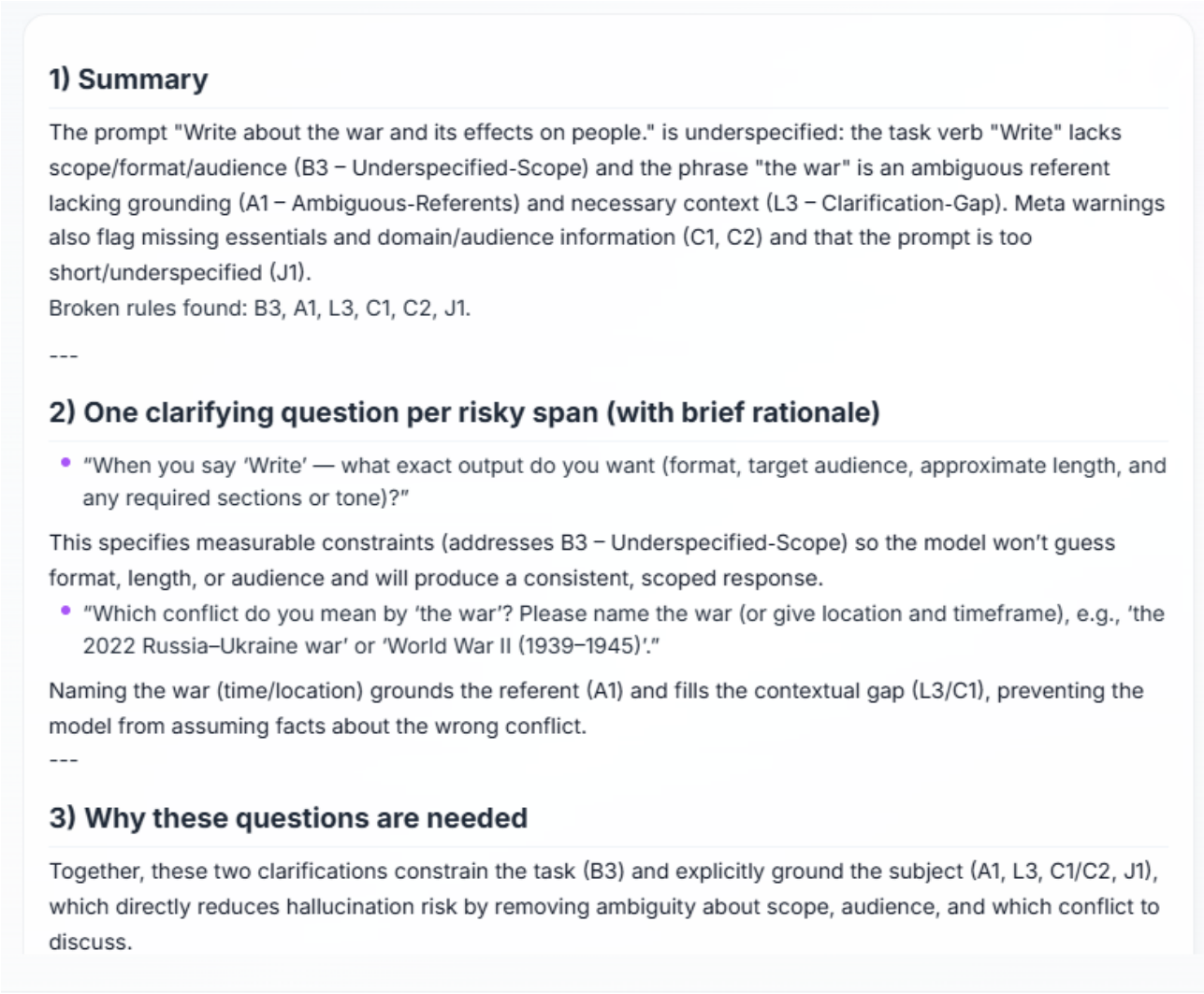


Figure 3.6 Echo UI: Prepared questions by the initiator agent to (i) minimize user input and effort by analyzing the risks and providing a clear, directed question for each of the token spans (ii) Providing a section explaining the integration of every answer into the final, refined prompt.

3. **Conversational Agent — Iterative Refinement.** Engages in a controlled, critical dialogue with the user. It evaluates proposed edits against the taxonomy, prevents regressions, and guides the user toward removing ambiguity, resolving conflicts, and restoring structural clarity. Provides intermediate prompts in each stage until the user is satisfied with the final prompt state.

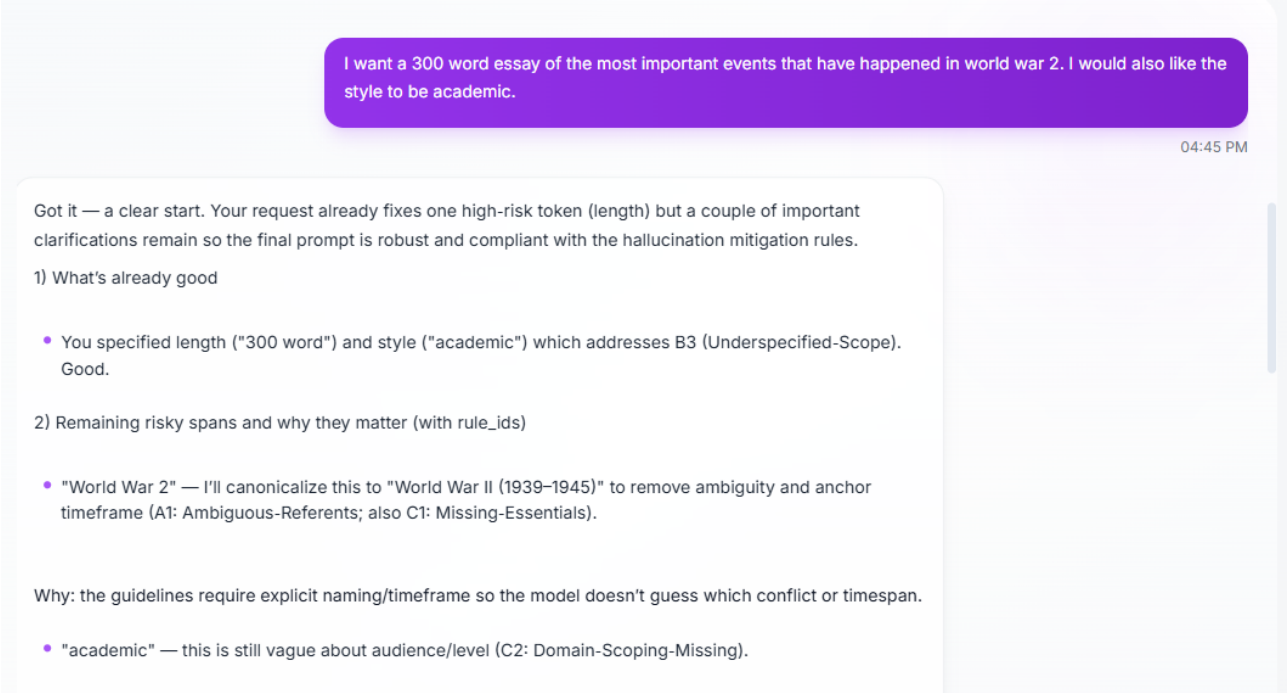


Figure 3.7 Echo UI: One conversation turn between the user and Echo after the user answers the questions of the initiator.

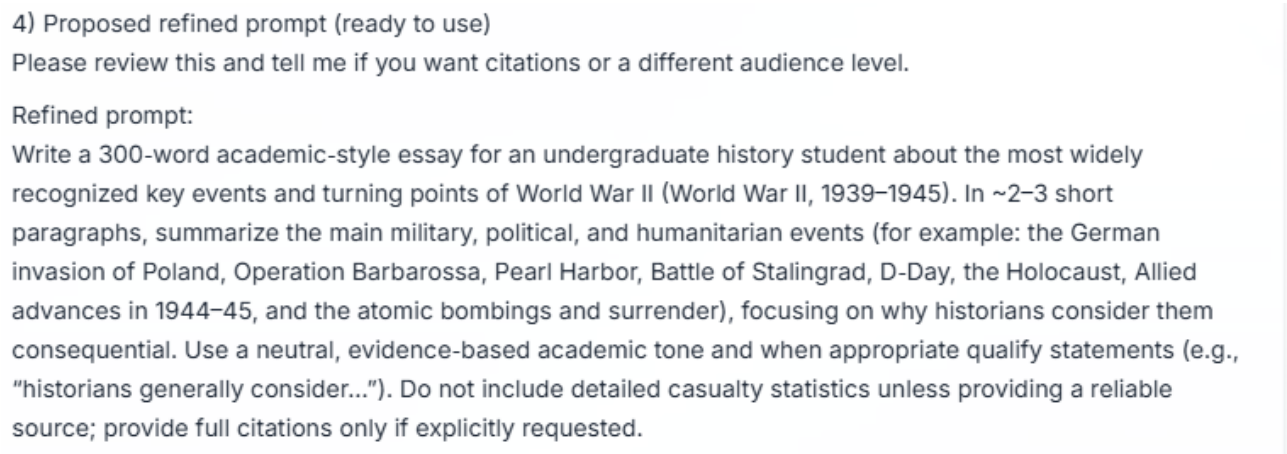


Figure 3.8 Echo UI: The final, rewritten prompt provided by Echo after the conclusion of one iteration loop between him and the user. This prompt is presented as the best option to minimize hallucinations.

4. **Preparator Agent — Re-analysis.** Consolidates conversation inputs into five prompt versions in the event that the final prompt proposed by the conversational agent is not at an acceptable level , prepares it for re-analysis, and enables iterative improvement until PRD and the violation list converge to acceptable levels.

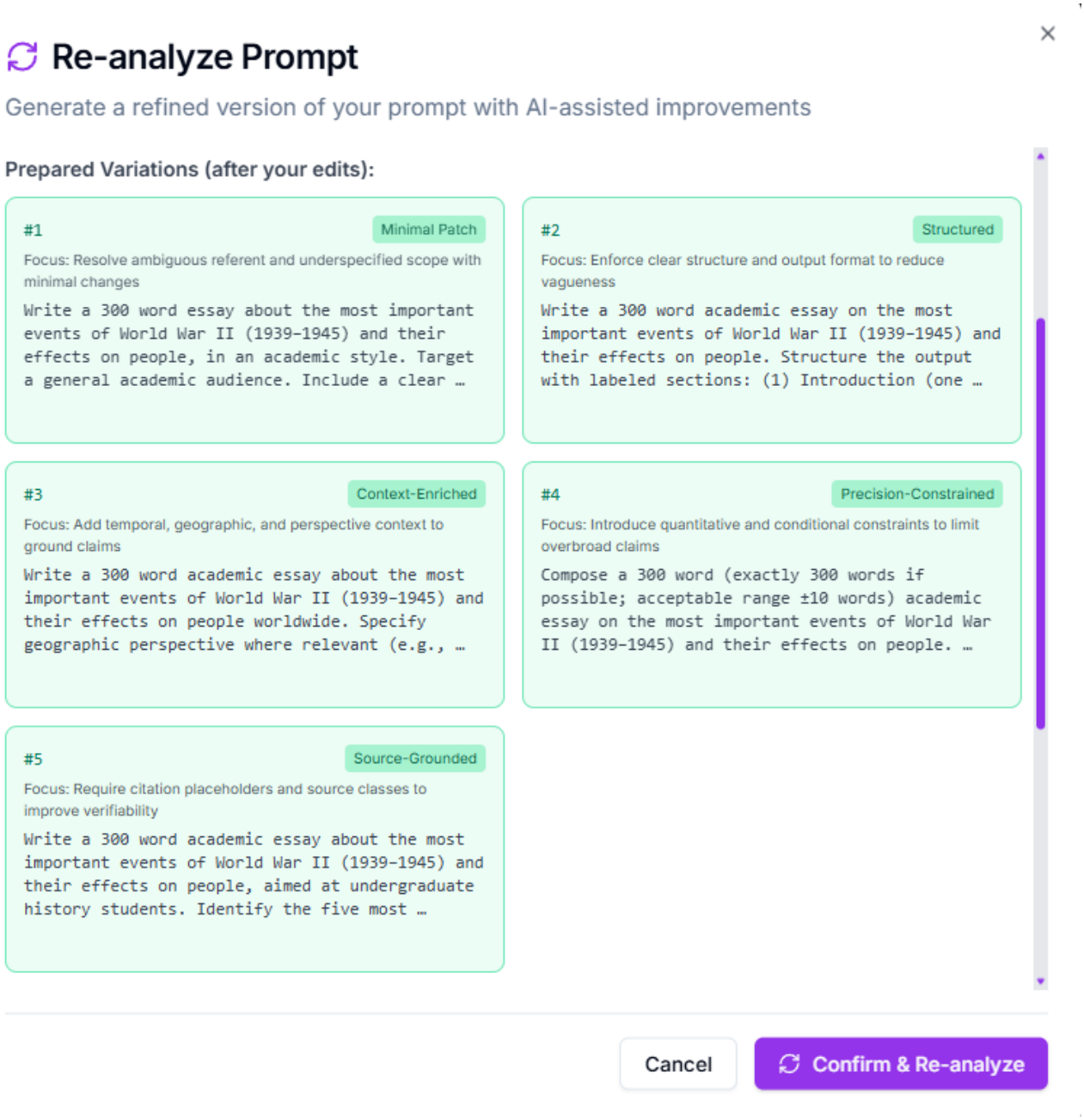


Figure 3.9 Echo UI: Proposed alternatives to the initial prompt proposed by the conversational agent in case the user wants to deviate from the verbosity, style or amount of change proposed by it.

This agentic workflow creates a controlled loop:

$$\text{Prompt} \rightarrow \text{Analysis} \rightarrow \text{Clarification} \rightarrow \text{Refinement} \rightarrow \text{Refined prompt } (\rightarrow \text{Re-analysis.})$$

Echo therefore operates not as a one-shot rewriting tool but as a structured refinement environment in which user intent remains central and traceable.

3.3.2 Prompt/Meta Taxonomy at a High Level

Echo's behaviour is governed entirely by the XML-encoded taxonomy described in Chapter 3. The taxonomy distinguishes between:

- **Prompt-level risks**: token-localisable issues that the Analyzer highlights inside the prompt.
- **Meta-level risks**: global issues that cannot be attributed to specific spans and are surfaced as structured warnings.

This dichotomy determines:

a) the output schema of the Analyzer (mixed span and global violations),

b) the refinement prompts generated by the Initiator (local vs. structural questions),

c) the refinement strategy enforced by the Conversational Agent,

d) the synthesis logic of the Preparator.

In this way, the taxonomy does not merely classify risks; it shapes Echo's architecture.

3.3.3 Prompt Risk Density (PRD) at a High Level

Echo summarises detected risks using the Prompt Risk Density (PRD) metric, a length-normalised, severity-weighted indicator of how much of the prompt is occupied by hallucination-prone spans. At a conceptual level, PRD serves two functions:

- **diagnostic**: a signal indicating whether a prompt contains significant ambiguity or structural defects;
- **process-oriented**: a before/after comparison metric that quantifies the effect of each refinement cycle.

PRD is not interpreted causally in this thesis (see Section 3.2.3), but it provides a consistent, interpretable measure that anchors Echo's iterative loop.

3.3.4 Overview Diagram

Figure 3.10 summarises the full solution flow, showing how the user prompt is transformed through the four agents and guided by the taxonomy.
This overview establishes the conceptual foundation for the following sections : Section 3.5 explores alternative architectures that were considered and discarded. Section 3.6 analyses the design decisions that led to Echo's final shape. Section 3.7 then provides a complete description of the final architecture, agent prompts, XML contracts, and system components.

3.4 Taxonomy Operationalisation

This section provides a design-oriented operationalization of the Prompt/Meta Taxonomy introduced in Chapter 3 and explains how its conceptual structure is realized in Echo's agent pipeline, data contracts, and user interface. Rather than repeating the full taxonomy, the goal here is to highlight the specific taxonomy features that directly shape (i) the Analyzer's detection behavior, (ii) the internal JSON/XML output schema, and (iii) the frontend's visualization components, including risk spans, PRD gauges, and meta-level warnings.

3.4.1 Prompt vs. Meta Risks: System-Level Implications

The taxonomy draws a strict distinction between *prompt-level* and *meta-level* risks (see Chapter 3), which plays a decisive architectural role.

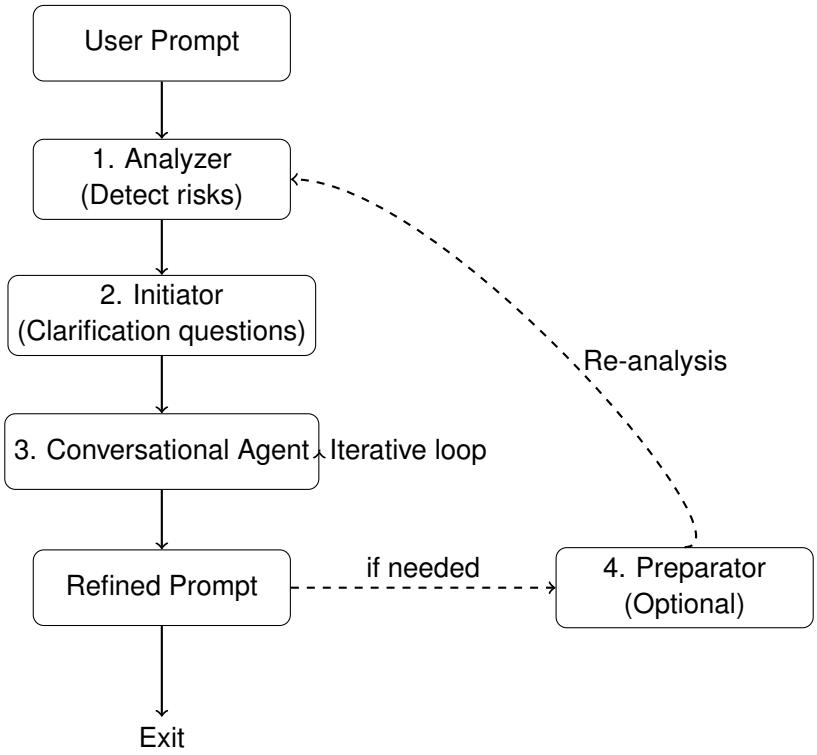


Figure 3.10 High-level agent workflow in Echo.

PROMPT-level risks (token-localisable). Prompt risks correspond to concrete spans in the user's input. They include ambiguous referents, quantifier vagueness, ungrounded temporal markers, or underspecified task verbs. Their local nature allows the Analyzer Agent to wrap affected text segments in structured XML tags:

<RISK_1> ... </RISK_1>.

Each tagged span is associated with exactly one highest-severity rule match (e.g. A1, B3, L3), ensuring deterministic span-to-rule mapping.

META-level risks (global, non-localisable). Meta risks are structural deficiencies that cannot be meaningfully mapped to spans: missing actors, missing domain constraints, overlong prompts, or multi-objective overload. These are returned as global warnings in the JSON schema, without span highlights. Meta risks guide the Initiator Agent in producing high-level clarifying questions that restore missing structure or context, setting the stage for the user to draw the complete picture of the prompt.

Architectural Implications. This dichotomy drives several system design decisions:

- The Analyzer *must* emit both XML span annotations and global meta warnings to be then parsed and displayed in the UI.
- The UI requires three presentation layers: (i) inline token highlights, (ii) a PRD container with global, meta warnings and localized, prompt warning, (iii) a section for a 1:1 mapping between token spans and hallucination risks with attributed categories and mitigation strategies.
- The Initiator as well as the conversational agents recognize the dichotomy and formulates the questions accordingly to draw a line between structural and lexical changes that have to be done by the user.

3.4.2 Severity, Rules, and the Detection Contract

Each rule in the XML guideline file encodes a severity level (`critical`, `high`, or `medium`). Echo treats these values as *non-negotiable*: the model is explicitly prevented from adjusting severities, upgrading or downgrading harms, or re-interpreting rules according to its own judgment. This constraint ensures stability and reproducibility of scoring.

Severity propagation. Severity levels influence three system components:

1. **UI highlight intensity**, applied directly to the XML-marked span. If two risks overlap, the more severe dictates the highlight intensity.
2. **Refinement prioritization**: critical and high-level risks are surfaced first.
3. **Contribution to Prompt Risk Density (PRD)** via deterministic severity weights.

XML–JSON output contract. The Analyzer produces a hybrid output:

- The **annotated prompt** is rendered as the user's original text decorated with <RISK_n> tags.
- A structured **JSON object** describes each risk token, its classification, severity, rationale, and mitigation.

Listing 3.1 Excerpt of the Analyzer's output schema and the contract enforced on the model

```
{
  "annotated_prompt": "The ORIGINAL prompt with <RISK_1>risky token</RISK_1
     >...",
  "analysis_summary": "...",
  "risk_tokens": [
    {
      "id": "RISK_1",
      "text": "risky token",
      "risk_level": "high",
      "classification": "Quantification-Constraints [B3]",
      "reasoning": "One-sentence explanation.",
      "mitigation": "One concrete local fix."
    }
  ],
  "risk_assessment": {
    "prompt": {
      "prompt_PRD": "",
      "prompt_violations": [...],
      "prompt_overview": "..."
    },
    "meta": {
      "meta_PRD": "",
      "meta_violations": [...],
      "meta_overview": "..."
    }
```

```
    }
}
```

3.4.3 UI Integration: Annotated Prompt and Risk Dashboard

For completeness, Figure 3.11 shows how the taxonomy materializes in the production interface. This screenshot corresponds to the input prompt “Write about the war and its effects on people.” and demonstrate: (i) span-level highlights; (ii) PRD computation; (iii) token-level and meta-level violations.

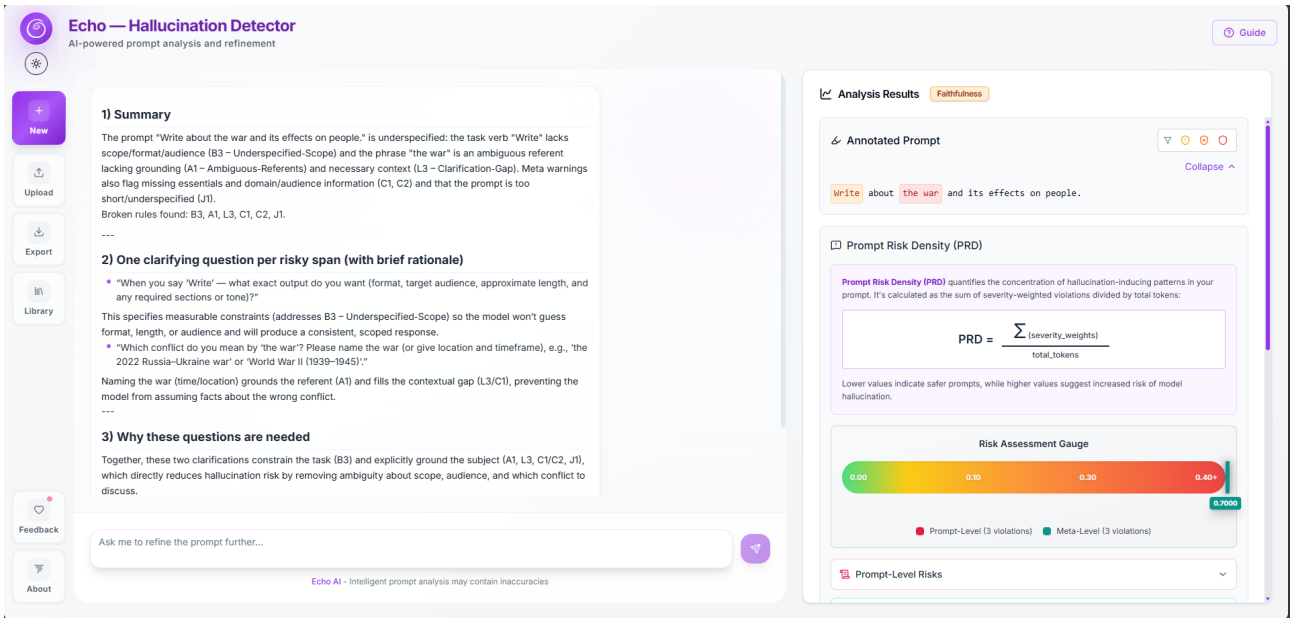


Figure 3.11 Overview of the complete Echo UI.

3.4.4 Purpose of This Operational Recap

This section establishes the concrete design commitments that arise when the taxonomy of Chapter 3 is implemented in a multi-agent shift-left refinement system:

- Echo must support a mixed span/global output format.
- All agents rely on a deterministic, severity-locked rule set encoded in XML.
- The UI requires both inline and panel-based visualisation channels.
- PRD acts as a normalised process indicator across refinement cycles.

Together, these elements form the operational substrate from which the analysis pipeline, refinement loop, and user feedback mechanisms in Sections 3.10 and 3.4 are derived and the baseline for the whole system to function.

3.5 Design Space Exploration

This section documents the iterative design process that led to Echo’s final architecture. Rather than presenting a *fait accompli*, it traces four major iterations, each addressing concrete limitations discovered in

the previous version. This sequence ties the final design to concrete empirical findings instead of relying solely on theory.

3.5.1 Iteration 0: One–Shot Rewriting vs. Iterative Clarification

Before any taxonomy or agent architecture existed, the first design fork concerned the overall *workflow paradigm*. The simplest baseline, which was aimed to mimic tools such as PromptPerfect or ChatGPT's "Improve Prompt" mode, was a **one-shot rewriting system**: the user provides a raw prompt, and the model returns a single improved version in one step, guided by hallucination-related prompting guidelines extracted from prior literature.

Approach

The initial prototype implemented exactly this: given a prompt, an LLM produced a rewritten version intended to be clearer, more structured, and less ambiguous, without asking the user for clarification or providing intermediate feedback. The difference with present tools would be that this model would be fed hallucination detection guidelines tailored for this use case instead of general prompting best practices, which would have made this solution specialized in rewriting prompts to make them more robust against hallucinations.

Outcome

Although superficially appealing, the one-shot rewriting paradigm failed on four central dimensions:

- **Missing information cannot be reconstructed.** If the user omits domain, actors, baselines, or assumptions, a rewrite *must guess*. This often injected incorrect context, introducing new hallucination risks (e.g., arbitrary dates, invented entities).
- **Ambiguities were resolved incorrectly rather than clarified.** Pronouns ("it"), deixis ("this", "above"), vague constraints ("brief"), and underspecified tasks ("summarise") were resolved by the model's heuristics rather than by user intent.
- **Risks were left untreated due to LLM stochasticity.** As shown later in the Evaluation chapter, first-pass recall was consistently lower than precision: the model reliably found some risks, but not all. One-shot rewriting therefore preserved false negatives that an iterative workflow could subsequently eliminate.
- **Rewriting introduced fresh risks.** Fixing a vague phrase frequently inserted new vague or stylistically inflated ones, shifted framing , or produced constraints misaligned with the user's actual goals.

In practice, rewritten prompts often appeared cleaner but were *less faithful* to the user's actual goal and only marginally more robust.

Learning

This baseline demonstrated a critical insight:

> **One-shot prompt rewriting does not mitigate prompt-induced hallucination risks.** Faithfulness requires recovering user intent, and this is only possible through iterative clarification.

Therefore, Echo must be built around:

- explicit detection of risky spans,

- user-driven clarification questions,
- iterative refinement through conversation and re-analysis,
- rather than opaque one-shot transformations.

Comparison of Workflows

Figure 3.12 contrasts the one-shot rewriting baseline with Echo's iterative refinement loop. The former is linear and inference-driven, while the latter is cyclic and clarification-driven, enabling recovery of missing intent and elimination of residual risks.

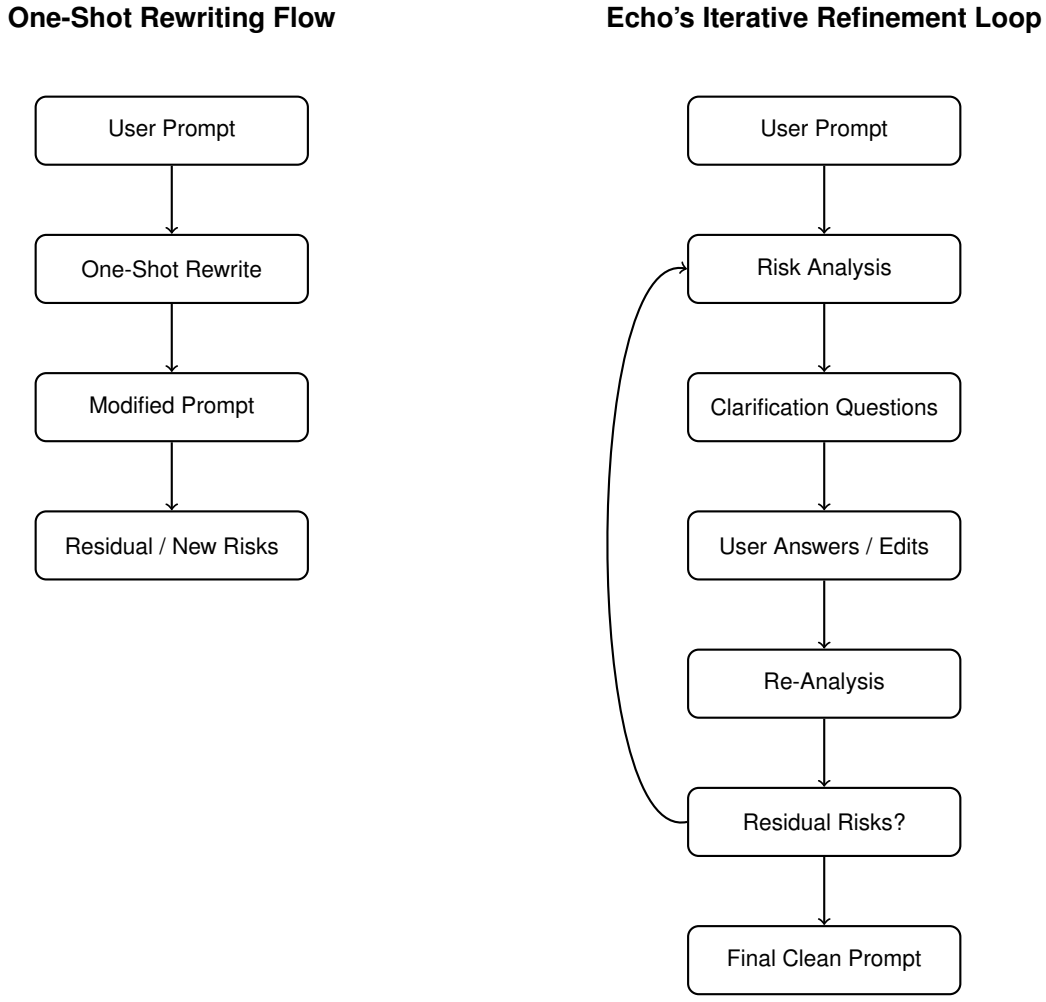


Figure 3.12 Comparison of design paradigms: one-shot rewriting (left), which cannot recover missing intent and frequently introduces new risks, versus Echo's iterative clarification and re-analysis loop (right), which progressively eliminates user-sided hallucination triggers.

3.5.2 Iteration 1: Flat Rule List and Unstructured Detection

Approach

The first attempt at operationalizing the literature review consisted of implementing a *flat list* of 32 hallucination-related rules derived from academic papers, provider guidelines, and practitioner engineering blogs (see Chapter 2.1). The rules were loosely grouped under broad headings such as "Ambiguity",

“Context”, and “Constraints”, but no deeper structure existed.

The prototype instructed an LLM to:

> *“Analyze the following prompt for violations of these 32 rules. For each violation, identify the token span and rule ID.”*

The model returned a list of rule IDs and corresponding token spans, which were then highlighted in the UI.

Outcome

Although superficially functional, this design surfaced three major flaws:

- **Lack of conceptual structure.** The flat list was an aggregation of heterogeneous rules with no principled explanation of how they related. From a research perspective, this did not constitute a contribution—only a consolidation effort.
- **Lack of traceability and clarity.** The system had no explicit mapping between rule categories and types of hallucination risks. Users frequently asked: *“Why is this a problem?” “Is this missing context or a wording issue?”*
- **Unclear remediation strategy.** Because rules were homogeneous and unstructured, the system could not decide:
  - which risks require *rewriting*,
  - which require *additional information*,
  - which were truly important vs. cosmetic.

Learning

The key insight from this iteration was the importance of **1:1 token-span traceability**. Users must be able to see *exactly which substring* causes hallucination-related risk, and why.
This established the foundational design requirement:

Every detected prompt-level risk must correspond to a precise and minimal token span.

This insight later motivated the XML ‘<RISK_n>‘ highlighting scheme and required moving beyond an unstructured rule list.

3.5.3 Iteration 2: Introducing the Prompt/Meta Taxonomy

Approach

Building on the lessons of Iteration 1, the next iteration is to provide *conceptual structure* by analyzing the nature of the 32 rules. A crucial observation emerged: the rules naturally split into two classes:

1. **Prompt-level risks** — issues that *exist in the text* and can be mapped to token spans (e.g., ambiguous pronouns, vague quantifiers, missing units).
2. **Meta-level risks** — issues arising from *absent or conflicting information* that cannot be localised (e.g., missing actors, fused multi-step instructions, contradictions).

These two categories required different remediation strategies:

- Prompt-level risks $\rightarrow$ local rewriting.
- Meta-level risks $\rightarrow$ clarifying questions or structural additions.

This motivated the definition of the **two-dimensional taxonomy** later formalised in Chapter 3: PROMPT vs. META risks as one layer, and faithfulness vs. factuality as the manifestation layer.

Outcome

This iteration produced two essential advancements:

- A principled categorisation of rules into *token-localisable* vs. *structural* risks.
- The insight that the UI and subsequent agents must treat these two classes differently.

However, this iteration still used a single-agent pipeline for analysis and refinement, leading to the problems documented in Iteration 3.

Learning

Two fundamental insights emerged:

- **Token-level and structural risks behave differently.** Detecting them, presenting them, and remediating them require distinct workflows.
- **The taxonomy is the conceptual backbone of the system.** It provides a research contribution, an organising principle for detection rules, and a way to explain risk origins to users.

These insights laid the foundation for Echo's current architecture and the design of the Analyzer's XML output schema.

3.5.4 Iteration 3: Two-Agent System Without Explicit Questioning

Approach

The third iteration introduced a separation between detection and refinement by splitting the system into two agents:

1. An **Analyzer Agent** that performed a one-shot XML-guideline-based analysis.
2. A **Conversation Agent** intended to guide a multi-turn refinement process using the analysis as context.

The Conversation Agent was expected to read the risks, infer what was missing, and autonomously choose whether to ask questions or suggest specific edits.

Outcome

Although this avoided the instruction conflict of Iteration 2's single-agent system, a critical issue emerged:

- The Conversation Agent behaved like a generic *automatic prompt optimiser*. It largely skipped clarification and produced rewritten prompts directly.
- When users asked "What should I change?", it provided a rewritten prompt rather than asking for clarification.

- The system often inserted incorrect assumptions—fixing ambiguity by guessing instead of asking.

Example failure

> **User:** "Can you help me fix the ambiguity in *it*?"
>
> **System (Iteration 3):** "Here is a better version: 'Summarise the research paper and explain the methodology.' "
>
> **User:** "But I never said which paper—I planned to paste it next."

The agent guessed a nonexistent paper instead of asking what "it" referred to.

Learning

This iteration established two essential design requirements:

- Echo's purpose is **not** stylistic optimisation but **faithfulness preservation** through explicit intent recovery.
- This requires a **question-first refinement workflow**, not autonomous rewriting.

This led to the introduction of a third specialised component: the **Initiator Agent**, responsible for turning each risk into one targeted clarifying question.

3.5.5 Iteration 4: Overly Strict Detection

Approach

With the Analyzer–Initiator–Conversation pipeline in place, the first version of the XML guidelines treated *all* rule violations equally. Any matched pattern—whether critical or cosmetic—was surfaced as a risk, with no severity levels or prioritisation.

Outcome

Although this version maximised coverage, it suffered from poor usability:

- Even short, well-posed prompts triggered numerous warnings (missing word counts, no explicit format, no audience level).
- Prompts became verbose and bureaucratic when users tried to follow all suggestions.
- Test users described the model as "pedantic" and "like writing a legal contract".

Learning

This revealed an important insight:

- Not all risks have equal impact. Missing actors or contradictory instructions are *critical*; missing word limits are *optional*.
- Treating everything as equally important leads to overcorrection and user fatigue.

Thus, the taxonomy was extended with **severity levels** (medium, high, critical) and numeric weights used later in the PRD metric. The UI was modified to visually prioritise high-severity risks.

3.6 Design Rationale and Alternatives

This section positions the final architecture of Echo against plausible alternatives explored during the design iterations in Section 3.5. The goal is not to enumerate every rejected option, but to explain why the selected design (a four-agent, taxonomy-driven refinement pipeline) most effectively satisfies the constraints derived from the iterative process.

3.6.1 Multi-Agent Architecture vs. Single-Agent Solutions

Alternatives considered

- **Single monolithic agent.** One model performs analysis, question generation, conversation and rewriting, switching "modes" through prompts (Iteration 2).
- **Two-agent pipeline.** A dedicated Analyzer Agent and a Conversation Agent, with the latter implicitly responsible for choosing when to ask questions if any are asked (Iteration 3).
- **Four-agent pipeline (final).** Analyzer, Initiator, Conversation and Preparator agents, each with a single, narrow mandate.

Selected design and rationale

Echo adopts a **four-agent** architecture because the iterations revealed three crucial requirements:

1. **Strict task separation.** Structured one-shot tasks (analysis, preparation) and open-ended dialogue tasks (questioning, refinement) do not coexist well in a shared context. Splitting them prevents instruction bleed and maintains stable analysis outputs and a deterministic user flow : Analysis $\rightarrow$ Clarification $\rightarrow$ Conversation $\rightarrow$ Iteration.
2. **Enforcing a question-first workflow.** The two-agent design allowed the Conversation Agent to "auto-optimize" prompts without clarifying missing context. Introducing a dedicated Initiator Agent ensures that each risk is converted into a targeted question before any rewriting occurs.
3. **Modularity and future-proofing.** Additional risk detectors or refinement modes can be added as separate agents without modifying the core workflow. Similarly, the Conversation Agent can be replaced with a different model with minimal integration overhead. A separation of concerns enables choosing different models for different tasks.

Compared to a monolithic agent, the four-agent design increases architectural complexity but offers far greater reliability, controllability and traceability which are critical features for a research artifact.

3.6.2 Representation Formats for Analysis Output: XML vs. JSON

Echo's analysis output consists of two complementary components: (i) an *annotated prompt* with token-level risk highlights, and (ii) a *structured risk list* describing each detected issue. Early prototypes experimented with representing both components in a single format (all-XML, all-JSON, or free-form prose), but iteration revealed that no single format satisfied all requirements simultaneously.

Design alternatives considered

- **Free-form natural language.** The model explains risks in prose. Easy to generate, but difficult to parse deterministically although token span detection could be done through having custom markers at the start and the end of each span.
- **Pure JSON.** Both span annotations and risk metadata live inside JSON. Structurally sound, but JSON is too rigid for simple inline span highlighting and breaks easily under small formatting mistakes.
- **Pure XML.** Both highlights and metadata embedded in an XML tree. Good for marking spans, but verbose and cumbersome for representing risk objects with many attributes.

Selected design: XML for spans, JSON for risk objects

Echo uses a **hybrid** output format:

- the **annotated prompt** is encoded in **XML**, using inline tags such as `<RISK_1> ... </RISK_1>` to mark the exact start and end of risky spans;
- the **risk list** is encoded in *JSON*, where each risky span becomes a standalone object containing its rule ID, pillar, severity, reasoning, and mitigation.

This choice is driven by two observations from Iteration 1:

1. **Span-level XML markup aligns with the 1:1 mapping requirement.** XML tags are ideal for inline annotations: they are non-verbose, easily nestable, and can be easily parsed when reconstructing the annotated prompt in the UI. JSON cannot embed inline ranges without index offsets and is brittle under minor formatting errors.
2. **JSON is superior for representing risk metadata.** Risk entries have many attributes (rule violated, severity, pillar, reasoning, mitigation). Representing these as JSON objects is concise, machine-friendly, and consistent with the "each risky span is a standalone unit" principle that emerged from Iteration 1.

The result is motivated by a separation of concerns: XML provides precise, robust span marking while JSON provides rich, highly structured risk descriptions.

3.6.3 LLM-Based Detection vs. Rule-Based or Fine-Tuned Models

Alternatives considered

- **Hand-engineered rules.** Regex pipelines, dependency parsing, coreference tools and rule-based NLP systems.
- **Fine-tuned classifier.** A supervised model trained on annotated prompts to predict risk labels.
- **LLM-based detector (final).** A LLM model constrained by XML guidelines and validated by deterministic post-processing.

Selected design and rationale

Echo adopts an LLM-based detector for three reasons:

- **Semantic sensitivity.** Many faithfulness risks (e.g. ambiguous pronouns, deictic references, vague constraints, structural gaps) depend on long-range semantics and pragmatic cues that cannot be reliably captured using regex or traditional NLP components. These lack the contextual awareness of long user prompts.
- **Feasibility and development cost.** Building accurate rule-based detectors for all pillars and patterns such as contextual completeness, discourse continuity, domain scoping, and multi-step task structure would require a complex set of heuristics and still struggle with many edge cases. Similarly, training a fine-tuned classifier would demand a considerably larger annotated dataset than what is realistic to create within the scope of this thesis.
- **High interpretability under constraints.** When guided by explicit XML patterns and severity labels, LLMs can produce structured, span-level outputs that are then normalized deterministically. This hybrid approach combines the flexibility of LLM inference with the reliability of rule-based validation although leaving some room for hallucination potential within the detection itself.

The resulting detector is not deterministic: like all LLM-based extractors, it can produce omissions or misclassifications. Echo mitigates these risks not through symbolic post-processing, but through architectural controls: (i) strict XML-constrained prompting, (ii) severity-anchored pattern matching, and most importantly (iii) iterative re-analysis within a multi-agent workflow. By repeatedly exposing the model to its own rule-governed outputs, and by separating detection, questioning, refinement and synthesis into distinct agents, Echo reduces the space for detector-induced hallucinations across iterations and increases the stability of the detected spans.

3.7 System Architecture and Components

Echo is implemented as a modular, four-agent pipeline that transforms a free-text prompt into (i) an annotated version with span-level risk highlights, (ii) a structured risk list, (iii) a PRD-based risk assessment, and (iv) a refined prompt produced through an iterative clarification loop.

3.7.1 High-Level Architecture

Echo follows a linear-but-iterable analysis–clarification–refinement workflow (Figure 3.10 in Section 3.5). Each stage is implemented by an independent LLM agent with a narrow, non-overlapping responsibility:

1. **Analyzer Agent** — one-shot span-level and meta-level detection guided by the XML taxonomy.
2. **Initiator Agent** — one-shot translation of risks into targeted clarification questions.
3. **Conversation Agent** — multi-turn refinement based on user answers and rule explanations.
4. **Preparator Agent** — optional final clean-up, consolidation, and variant generation.

This separation avoids instruction bleed between structured extraction tasks and open-ended dialogue, and allows the model used for conversational refinement to differ from the model used for strict analysis.

3.7.2 Analyzer Agent

The Analyzer operationalises the taxonomy into concrete detections (RQ2). It receives the raw prompt and the XML guideline document and outputs:

- an *annotated prompt* with <RISK_i> tags marking token-level PROMPT risks;
- a list of META risks (structural issues not localisable to spans);
- a structured JSON risk list (rule ID, severity, pillar, explanation, mitigation);
- a per-pillar and overall PRD summary.

The Analyzer enforces the 1:1 token-to-risk mapping requirement derived from Iteration 1 by marking each highest-severity risky span exactly once. All severities are taken as-is from the XML guidelines to ensure stability and reproducibility.

3.7.3 Initiator Agent

The Initiator translates the Analyzer output into a series of concise clarification questions, ensuring that Echo adopts a question-first refinement workflow rather than a rewriting-first one (fixing the issue surfaced in Iteration 3). Each question directly references the risk(s) it addresses, enabling the UI to link questions to highlighted spans or meta warnings.

3.7.4 Conversation Agent

This agent handles the user-facing refinement dialogue. Its responsibilities are:

- incorporate user answers into suggestions for prompt edits;
- prevent regressions by referencing rule IDs and severities;
- avoid over-automation by asking follow-ups instead of rewriting prematurely;
- maintain the iterative nature of the refinement loop (RQ3).

It is instructed to behave like a critical reviewer rather than a compliant assistant, preserving user intent while removing prompt-induced risks.

3.7.5 Preparator Agent

The Preparator produces a clean, ready-to-use prompt by removing conversational artefacts, merging revisions, and optionally offering variants differing in verbosity or structure. This agent is stateless and only operates when the user is satisfied with the refinement stage.

3.8 Design Limitations and Non-Goals

Echo's architecture reflects explicit trade-offs that constrain its applicability.

Latency and Cost. The multi-agent pipeline increases inference cost and latency relative to single-agent or rule-based detectors. For the thesis setting—offline prompt analysis and small-scale experiments—this cost is acceptable, but a production system would likely collapse or downscale agents.

Language and Domain Generality. The taxonomy and guidelines were developed for English, general-purpose prompting. Some risk patterns—especially pronoun ambiguity or politeness forms—may manifest differently across languages or specialised domains (e.g., law, medicine). Extending the XML guidelines to domain- or language-specific packs is left for future work.

User Behaviour Assumptions. The interactive loop assumes cooperative users who are willing to clarify missing information and accept structured feedback. Handling adversarial users or prompt-injection attempts is out of scope, as Echo targets faithfulness in benign usage rather than robustness against malicious prompting.

No Causal Claims. Echo reduces prompt-induced faithfulness risks but cannot eliminate them, and does not address model-internal hallucinations. PRD is used as a descriptive metric; the system does not claim or evaluate a causal relationship between PRD values and hallucination probability.

3.9 Chapter Summary

This chapter presented Echo's design as the operationalisation of the taxonomy introduced in Chapter 4. The design evolved from a flat rule list toward a structured artefact through several iterations, each addressing concrete shortcomings in detection stability, interpretability, and refinement behaviour. The final system incorporates:

- a two-dimensional taxonomy distinguishing token-level (PROMPT) from structural (META) risks;
- a four-agent pipeline separating analysis, questioning, refinement, and preparation;
- XML-based span markup combined with JSON risk objects to enable precise token-to-risk mapping and deterministic PRD scoring;
- severity-aware detection to balance completeness with usability.

This architecture directly addresses RQ2 by implementing a practical detector grounded in the taxonomy, and supports RQ3 by enabling iterative, user-guided refinement with quantifiable before/after changes. The next chapter evaluates Echo's detection behaviour, stability, and refinement effectiveness.

4 Methodology

This chapter describes the methodological approach used in this thesis. We adopt a *Design Science Research* (DSR) paradigm to (i) frame the practical problem of prompt-borne faithfulness errors, (ii) design and build an artefact (ECHO) that surfaces *prompt-time* risks and guides human-centered refinement, and (iii) evaluate its utility using span-level agreement with human annotations, lexical stability analyses, and prompt-time risk metrics.

To orient the reader, Figure 2.1 provides a high-level map of the methodology. It visualises how the thesis proceeds from identifying prompt-borne faithfulness risks, to formalising a taxonomy and guidelines, to instantiating these in an artefact (ECHO), and finally to evaluating the artefact using a dedicated benchmark, PRD metrics, and span-level agreement. This overview anchors the detailed sections that follow.

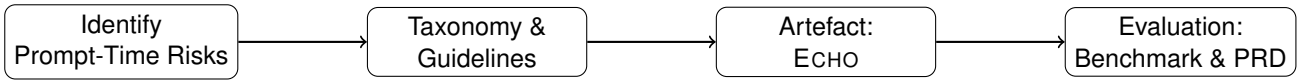


Figure 4.1 High-level methodology map linking risk identification, taxonomy and guideline construction, artifact design, and evaluation.

Rather than presenting methods as a static checklist, the chapter follows the arc of the thesis: from conceptual framing and taxonomy construction, through guideline operationalization, to the construction of an evaluation dataset and quantitative analysis of Echo's behavior. Methodology and System Design therefore form two sides of the same coin: this chapter focuses on *how* knowledge is produced (datasets, annotation, metrics, procedures), while Chapter 3 describes *what* is built and why. The two are tightly linked: the taxonomy and guidelines shape the artifact, and the artifact in turn acts as an empirical lens on the taxonomy.

The chapter proceeds as follows. Section 4.1 introduces the DSR framing and situates our research questions within it. Sections 4.3 & 4.4 present the Prompt/Meta Risk Taxonomy and its operationalization into hallucination-detection guidelines. Section 4.6 defines the Prompt Risk Density (PRD) metric and outline the overall study design. Section 4.7 describes the evaluation dataset, annotation protocol, baselines, metrics, and experimental procedures. Finally, Section 4.12 discuss statistical treatment, validity considerations, and a brief summary.

4.1 Design Science Research Method

4.1.1 DSR Framing

We follow a Design Science Research (DSR) cycle that integrates *Environment/Relevance* (practical problem), *Knowledge Base/Rigor* (prior theory and artifacts), *Design/Build* (construction of ECHO), and *Evaluation* (utility and quality), leading to *Design Knowledge* contributions.

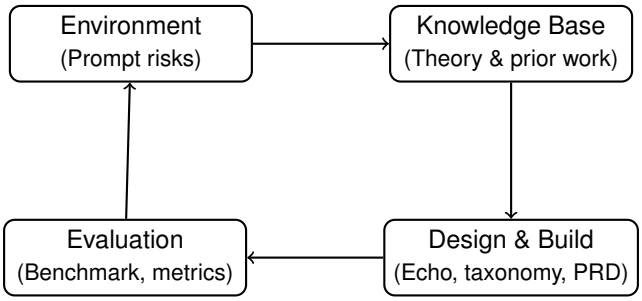


Figure 4.2 Simplified DSR cycle used in this thesis. Environment and knowledge base inform the design and build of the artefact, which is evaluated and feeds back into improved understanding.

4.1.2 Research Questions

The thesis is structured around three research questions, introduced in Chapter 1 and repeated here for methodological grounding. Each question aligns with a specific stage of the Design Science Research (DSR) cycle and motivates the methodological choices in this chapter.

RQ1 — Landscape and Gap

Which types of user-sided prompt risks that can lead to faithfulness-related hallucinations are described in existing research and practitioner guidelines, and to what extent are these risks already organised into a structured, operational taxonomy?

RQ1 corresponds to the *Environment* and *Knowledge Base* elements of DSR. It is addressed through a structured synthesis of academic literature, provider documentation, and practitioner sources, yielding a consolidated view of prompt-borne risks and clarifying the absence of an actionable taxonomy— thereby motivating the Prompt/Meta taxonomy introduced later in this chapter.

RQ2 — Taxonomy and Detection

Can these literature-derived risks be consolidated into a two-dimensional prompt/meta taxonomy that supports reliable classification and token-span-level detection of user-sided faithfulness risks in real prompts?

RQ2 aligns with the *Design & Build* stage of DSR. It is tackled by deriving the Prompt/Meta Risk Taxonomy, operationalising it into XML guidelines, and implementing Echo's Analyzer to perform rule-based token-level detection. Its evaluation relies on span-level precision, recall, and detailed error analyses relative to manually annotated gold data.

RQ3 — Refinement Effectiveness

Does Echo's interactive refinement loop, based on this taxonomy and its detections, measurably reduce prompt risk and qualitatively improve the faithfulness and completeness of user prompts compared to their original versions?

RQ3 aligns with the *Evaluation* stage of the DSR cycle. It is addressed using Prompt Risk Density (PRD) as a quantitative measure, together with qualitative before/after analyses from guided refinement sessions. These combined perspectives assess whether Echo produces meaningful, user-visible improvements in prompt clarity, explicitness, and alignment with user intent.

4.1.3 DSR Activities in This Thesis

Following the DSR framing introduced above, the methodological pathway of this thesis is composed of five tightly connected activities: analyzing the environment, grounding the work in an explicit knowledge base, designing and building the artifact, evaluating its behavior, and deriving reusable design knowledge.

Environment. Prompt-borne faithfulness risks like ambiguity, missing constraints, conflicting instructions, and weak structural scaffolding among others frequently degrade reliability in everyday LLM use. Stakeholders such as students, analysts, and developers rely on these systems for high-precision tasks; when prompts underspecify intent, subtle errors propagate and trust erodes. This problem environment motivates a shift-left focus on prompt-time diagnosis.

Knowledge Base. The work draws on hallucination taxonomies, faithfulness evaluation techniques, prompting guidelines, and practitioner engineering patterns. These sources provide conceptual mechanisms but lack operational, token-level rules. This gap motivates the construction of a structured Prompt/Meta taxonomy and a corresponding XML guideline specification that can be executed by both annotators and Echo's Analyzer agent.

Design and Build. Echo operationalises the taxonomy through (i) LLM-guided risk span detection constrained by XML guidelines, (ii) deterministic parsing into token-level risk objects, (iii) a prompt-time metric—Prompt Risk Density (PRD)—derived from severity-weighted span coverage, and (iv) a multi-agent refinement workflow (Analyzer, Initiator, Conversation, Preparator). Further details are given in Chapter 3. Here, Echo serves as the methodological instrument through which the taxonomy becomes measurable.

Evaluation. Evaluation examines three dimensions: (1) span-level detection quality against a 316-prompt, manually annotated benchmark; (2) detection stability under lexical variation using original–variant prompt pairs and (3) refinement effectiveness measured by PRD reduction and qualitative improvements in prompt clarity. Downstream QA metrics may be layered on but are not central, as the thesis isolates prompt-time faithfulness risks.

4.2 Overall Study Design

The study design connects the abstract DSR framing to concrete methodological steps: taxonomy construction, artefact implementation, dataset creation, and evaluation using detection and refinement metrics. At a high level, the thesis unfolds in three phases:

1. **Phase A — Criteria & Tagging Scheme.** Derive taxonomy pillars and rules from the literature; draft the guideline specification; calibrate the rule set using a small seed set of prompts.
2. **Phase B — Artifact Build (ECHO).** Implement the Analyzer, Initiator, Conversation, and Preparator agents, together with XML contracts and PRD computation.

3. **Phase C — Evaluation.** Construct the evaluation dataset, annotate gold spans, and assess (C1) span-level detection quality, (C2) stability under lexical variation, and (C3) PRD reduction after refinement.

Figure 4.3 synthesises these phases into an iterative DSR cycle, showing how empirical findings feed back into refinement of the taxonomy, guideline wording, and PRD calibration.

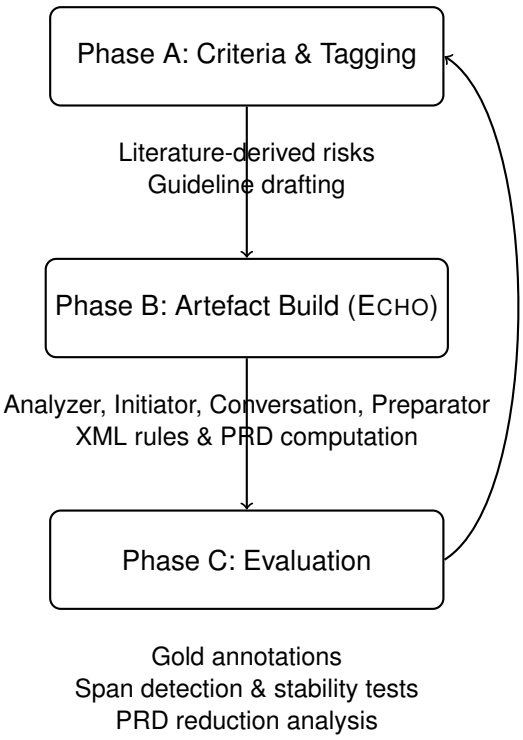


Figure 4.3 Study design as a three-phase process. Evaluation findings feed back into guideline wording and taxonomy refinement.

The remainder of this chapter follows this structure. Section 4.3 details the Prompt/Meta taxonomy created in Phase A. Section 4.4 formalises the XML guidelines that operationalise the taxonomy. Section 4.5 describes how ECHO implements these specifications. Finally, Sections 4.7–4.10 correspond to Phase C and report the dataset, annotation protocol, evaluation metrics, and analysis procedures.

4.3 Prompt/Meta Risk Taxonomy

4.3.1 Taxonomy definition

We distinguish two strata capturing *where* faithfulness errors originate within the user–LLM system:

- **Prompt** (PROMPT): token-level, highlightable spans within the user prompt likely to induce misalignment.
- **Meta** (META): structural or global risks that influence comprehension/execution but cannot be localised to a single span (emitted as warnings).

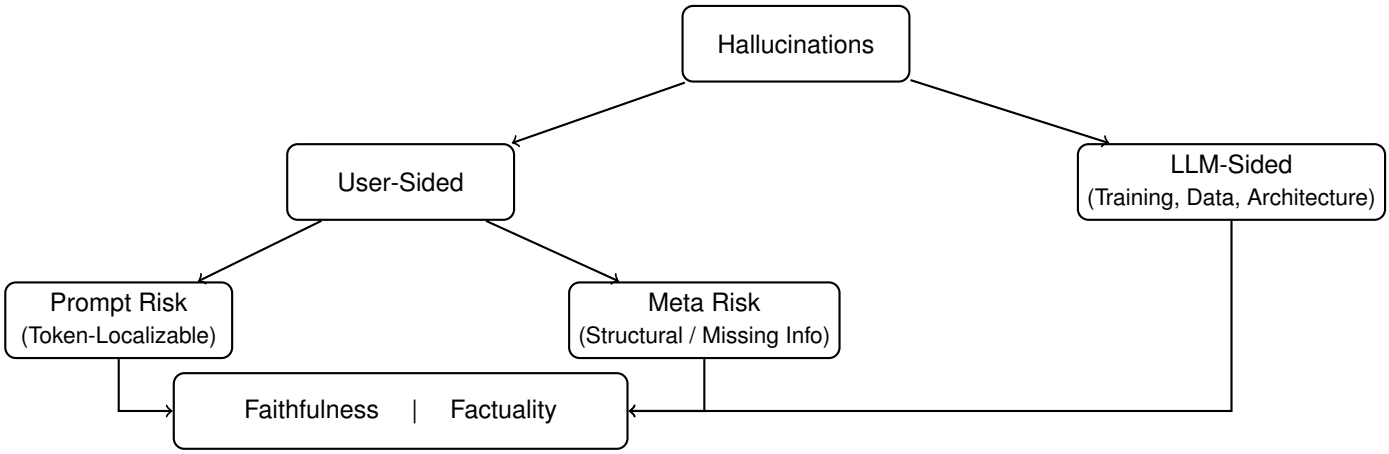


Figure 4.4 Conceptual hierarchy adopted in this thesis. User-sided hallucinations subdivide into token-localizable prompt risks and non-localizable meta risks. Both may manifest as faithfulness or factuality hallucinations.

The ecosystem view in Figure 4.4 clarifies the methodological boundary of the thesis: although hallucinations can originate from many sources within an LLM's training data or internal reasoning processes, the focus here is strictly on user-sided mechanisms. Within that branch, prompt-borne risks manifest in two distinct dimensions: token-level issues that can be precisely highlighted, and global structural issues that shape the model's interpretation as a whole.

The remainder of this section develops these two dimensions into a formal Prompt/Meta taxonomy, defined by clear pillars and rules that ground both the annotation protocol and Echo's detection behavior.

4.3.2 Prompt-Related Risks (PROMPT)

A. Referential-Grounding

What/Why — Unclear referents and deictics cause the model to bind facts to the wrong entities.

Detect cues

- Pronouns without antecedent; deictics: *here/there/this/that/above*; index terms: *former/latter*.
- Placeholders: *thing, issue*; acronyms before expansion when ambiguous (*ACA*, *NHS*).

Common fixes

- Name entities explicitly; replace deictics with concrete referents.
- Expand ambiguous acronyms on first mention.

Rules

- **A1 Ambiguous-Referents** : pronoun/deixis lacking a single clear antecedent.
- **A2 Canonical-Naming-Drift** : inconsistent names/aliases for the same entity.

Example — "<RISK>*It*</RISK> *should be summarised.*"; "Explain the <RISK>*ACA*</RISK> to a beginner."

B. Quantification-Constraints

What/Why — Vague scope and constraints yield off-target or over/under-detailed outputs.

Detect cues

- Vague length/quality: *short, detailed, comprehensive, robust, high-level*.
- Temporal vagueness: *recently, soon*.
- Open verbs without constraints: *summarise, explain, analyze* (no length/format/audience/units).

Common fixes

- Specify word/paragraph limits, audience, format, and units/date ranges.
- Replace *recently* with explicit cutoffs (e.g., “since 2023-01”).

Rules

- **B1 Relative-Descriptors** : subjective/vague descriptors without scales or bounds.
- **B2 Temporal-Vagueness** : time terms without a defined window.
- **B3 Underspecified-Scope** : missing limits for open task verbs.

D. Premises-Evidence

What/Why — False or loaded premises steer the model toward unfaithful frames.

Detect cues

- Unverified assumptions in setup; absolutist markers: *obviously, clearly, everyone knows*.
- Loaded questions nudging agreement or presupposing contested claims.

Common fixes

- Neutralise framing; ask for evidence or for the model to flag unverifiable claims.

Rules

- **D1 False-or-Unverified-Premise**.
- **D2 Leading-Opinion-Framing**.

E. Numbers-Units

What/Why — Missing units, baselines, or time zones cause numerical misinterpretation.

Detect cues

- Bare numbers where units matter; “reduce by *20%*” (no base).
- Currency without region; times without timezone.

Common fixes

- Add units and baselines (“20% relative to 2020 values”).
- Specify currency/region and timezones explicitly.

Rules

- **E1 Unitless-Number**
- **E2 Percent-No-Baseline**
- **E3 Currency-Unspecified**
- **E4 Time-No-Zone-or-Unit**

F. Retrieval-Anchoring

What/Why — Absent source type or document anchors break evidence linkage.

Detect cues

- “look up / search / find” with no source class.
- “the paper/dataset/benchmark” with no identifier (title/DOI/ID).

Common fixes

- Specify source class (e.g., peer-reviewed articles, official statistics).
- Give concrete identifiers (title, URL, DOI, dataset name).

Rules

- **F1 Source-Class-Unspecified**
- **F2 Document-Anchor-Missing**

H. Style-Bias-Role

What/Why — Entertainment styles, bias, or unsafe roleplay can distort factual tasks.

Detect cues

- *poem, rap, emojis* for analytical output.
- Stereotypes or loaded audience descriptions.
- Roleplay that may override safety or factuality.

Common fixes

- Request neutral, professional tone for factual tasks.
- Remove stereotypes; avoid unsafe roleplay instructions.

Rules

- **H1 Style-Inflation**
- **H2 Bias-Stereotypes**
- **H3 Unsafe-Roleplay**

I. Reasoning-Uncertainty

What/Why — Forcing certainty on unknowables or framing subjective opinion as fact yields spurious confidence.

Detect cues

- “exactly how many alien civilizations. . . ”.
- “your opinion” conflated with factual assessment.

Common fixes

- Allow the model to say “cannot be determined” and express uncertainty.
- Distinguish clearly between subjective evaluation and factual tasks.

Rules

- **I1 Uncertainty-Permission**
- **I2 Subjective-Framing-Risk**

L. Contextual-Integrity

What/Why — Contradictions and missing local context prevent faithful execution.

Detect cues

- "100 words" and "at least 500 words" in the same prompt.
- "don't do X" without a positive target.
- "review the text above" without any attached text.

Common fixes

- Resolve conflicts; specify which constraint dominates.
- Replace pure negations with explicit targets.
- Include or paste the context that is referenced.

Rules

- **L1 Conflicting-Instructions**
- **L2 Negation-Risk**
- **L3 Clarification-Gap**

4.3.3 Meta-Related Risks (META)

C. Context-Domain

What/Why — Missing essentials (who/what/when/where) and domain/audience scoping derail alignment.

Detect cues

- No explicit actor/object/time/location for the task.
- Unspecified domain, audience level, or jurisdiction.

Common fixes

- Add who/what/when/where explicitly.
- Specify domain, audience, and jurisdiction if relevant.

Rules (warnings)

- **C1 Missing-Essentials**
- **C2 Domain-Scoping-Missing**

G. Dialogue Continuity & Hygiene

What/Why — Contradicting earlier turns or duplicating instructions introduces ambiguity.

Detect cues

- “ignore previous instructions”.
- Repeated or conflicting directives within the same conversation.

Common fixes

- Restate current constraints succinctly.
- Remove duplicated and obsolete instructions.

Rules (warnings)

- **G1 Continuity**
- **G2 Instruction-Deduplication**

J. Prompt-Structure

What/Why — Unstructured prompts or overloaded objectives hide intent.

Detect cues

- Walls of text with no separation between context and instruction.
- Fused roles, constraints, and outputs in a single paragraph.

Common fixes

- Introduce delimiters and section headers.
- Scope down objectives or split them across prompts.

Rules (warnings)

- **J1 Length-TooShort-TooLong**
- **J2 Delimiter-Missing**
- **J3 MultiObjective-Overload**

K. Instruction-Structure-MultiStep

What/Why — Missing sequencing and step cues impair reasoning fidelity.

Detect cues

- Multiple tasks embedded in a continuous block without ordering cues.
- No request to show intermediate reasoning where needed.

Common fixes

- Enumerate steps explicitly.
- Separate creative and analytical components.
- Encourage stepwise reasoning when appropriate.

Rules (warnings)

- **K1 Task-Delimitation**
- **K2 Enumerate-MultiSteps**
- **K3 Stepwise-Reasoning-Cue**
- **K4 MultiObjective-Separation**

4.3.4 Provenance of the Prompt/Meta Risk Taxonomy

The *Prompt/Meta Risk Taxonomy* developed in this thesis is a synthesis of patterns recurrently documented across hallucination surveys, prompt-engineering research, and domain-specific analyses, rather than a reproduction of any single existing framework. Prior work highlights broad mechanisms underlying faithfulness failures including referential ambiguity, vague or missing constraints, unverifiable premises, and structural deficiencies in instruction design [5], [6]. In high-stakes domains such as law and medicine, empirical studies further show how missing context, implicit assumptions, or overloaded instructions materially degrade model reliability [7], [13].

These sources consistently describe *classes* of user-sided risks, but they do not provide an **operational, token-level detailed** rule set suitable for annotation or automated detection. To bridge this gap, practitioner materials (including provider guidelines, prompt-debugging workflows, and common error patterns observed in real-world LLM usage) were incorporated to identify actionable prompt-related cues (e.g., delimiter misuse, unitless numeric instructions, inconsistent naming, multi-objective overload).

The resulting taxonomy is therefore a **design-oriented consolidation**: high-level mechanisms from the hallucination literature, concrete prompting failures observed in practice, and the methodological requirement for explicit, reproducible rules. Each rule corresponds to a theoretically grounded failure point, but the rule set in its entirety constitutes a novel contribution enabling token-level analysis, guideline operationalization, and evaluation within this thesis.

A compact overview of all pillars and rules is provided in Table 4.3.5, and operational guidelines derived from this taxonomy appear in Section 4.4.

4.3.5 Taxonomy at a Glance

Pillar	Class	Rules (IDs & one-line)
A. Referential-Grounding	PROMPT	A1 Ambiguous-Referents; A2 Canonical-Naming-Drift
B. Quantification-Constraints	PROMPT	B1 Relative-Descriptors; B2 Temporal-Vagueness; B3 Underspecified-Scope
D. Premises-Evidence	PROMPT	D1 False-or-Unverified-Premise; D2 Leading-Opinion-Framing
E. Numbers-Units	PROMPT	E1 Unitless-Number; E2 Percent-No-Baseline; E3 Currency-Unspecified; E4 Time-No-Zone-or-Unit
F. Retrieval-Anchoring	PROMPT	F1 Source-Class-Unspecified; F2 Document-Anchor-Missing
H. Style-Bias-Role	PROMPT	H1 Style-Inflation; H2 Bias-Stereotypes; H3 Unsafe-Roleplay
I. Reasoning-Uncertainty	PROMPT	I1 Uncertainty-Permission; I2 Subjective-Framing-Risk
L. Contextual-Integrity	PROMPT	L1 Conflicting-Instructions; L2 Negation-Risk; L3 Clarification-Gap
C. Context-Domain	META	C1 Missing-Essentials; C2 Domain-Scoping-Missing
G. Dialogue Continuity & Hygiene	META	G1 Continuity; G2 Instruction-Deduplication
J. Prompt-Structure	META	J1 Length-TooShort-TooLong; J2 Delimiter-Missing; J3 MultiObjective-Overload
K. Instruction-Structure-MultiStep	META	K1 Task-Delimitation; K2 Enumerate-MultiSteps; K3 Stepwise-Reasoning-Cue; K4 MultiObjective-Separation

Table 4.1 Taxonomy at a glance. Full formal specification in Appendix 7.1.

Traceability. Rules under PROMPT yield token spans (<RISKi>... tags) used for highlighting, span-level evaluation, and PRD computation; META rules yield global warnings. Rule IDs align with Echo's outputs and with the XML guidelines.

4.4 Hallucination Detection Guidelines

The taxonomy defines *what* counts as a risk; the hallucination detection guidelines define *how* these risks are to be identified in concrete prompts. The guidelines are encoded in a single XML file (see 7.1) that serves three roles:

- **Annotation manual**: human annotators rely on the same rule descriptions and examples to mark spans.
- **Model contract**: the Analyzer agent is instructed to conform to this XML schema when producing risk objects.

- **Source of truth**: rule IDs, severities, and pillar assignments are anchored here; neither annotator nor model is allowed to change the specifications.

Each rule entry in the XML specifies:

- a unique identifier (e.g., `A1`), pillar, class (PROMPT / META);
- a severity label (critical, high, medium);
- a short description and detection cues;
- prototypical positive and borderline examples;
- one-line mitigation suggestions.

During annotation, the human annotator applies these rules to mark *minimal* spans: if a risk can be expressed by two tokens, it should not extend to ten. During model analysis, the Analyzer receives a compact textual representation of the same rules and is instructed to output rule IDs rather than inventing new labels.

Severity Levels. Each rule in the taxonomy is assigned a severity label (critical, high, medium). These levels are not empirically calibrated probabilities but *expert-informed ordinal categories*. They reflect (i) how frequently the risk type is discussed in hallucination and prompting literature, (ii) the extent to which prior work and practitioner guidelines describe the risk as inducing instruction deviation or unverifiable reasoning, and (iii) the authors' professional experience with common failure modes in real prompts.

Severity therefore encodes the *expected impact* of a risk on instruction faithfulness rather than a quantified likelihood. It provides a transparent and consistent weighting scheme for PRD rather than a claim of empirical ground truth.

The guidelines remain fixed for the entire evaluation. Any modification, however minor, constitutes a new guideline version and would require the full dataset to be re-annotated. Guideline stability is essential for reproducibility and for maintaining a consistent link between the taxonomy, the annotation process, Echo's behavior, and the evaluation metrics.

4.5 Artifact Summary (ECHO)

The artefact ECHO functions simultaneously as the object of study and as the measurement instrument used throughout this thesis. Methodologically, it is a configurable prompt-analysis and refinement pipeline with well-defined inputs and outputs grounded directly in the taxonomy and XML guidelines.

4.5.1 Inputs and Outputs

Given a user prompt, the Analyzer module produces four aligned output views:

- an *annotated prompt* in which risk spans are marked via inline `<RISK>` tags

- a structured *risk token list* (JSON objects specifying the token span, rule ID, pillar, severity, rationale, and mitigation cue)
- an *analysis summary* highlighting dominant risk patterns and contributing pillars as well as targeted questions for each detected risk pattern
- a quantitative *risk assessment* comprising:
  - overall and pillar-level PRD values,
  - prompt-level rule violations,
  - meta-level warnings.

These outputs serve as the basis for span-level evaluation, PRD computation, stability analysis, and downstream refinement.

4.5.2 Role in the Methodology

From a methodological perspective, ECHO operationalises the taxonomy in three complementary roles:

1. **Detector:** an executable version of the taxonomy that identifies prompt-time, token-level faithfulness risks under the same rule set used for gold annotations.
2. **Refinement guide:** a conversational interface that translates detected risks into clarification questions and proposes revised prompt drafts.
3. **Educational interface:** a transparent system that exposes rule IDs, severities, and mitigation strategies to users, thereby making the taxonomy actionable.

Because the guidelines are shared across annotation, analysis, and refinement, ECHO ensures methodological alignment: the artefact does not introduce its own categories, but implements exactly the same rule definitions that structure the evaluation.

4.6 Prompt-Time Risk Density (PRD)

Goal. PRD quantifies how much *prompt-borne* risk is present *before* any model generation, normalized for prompt length and severity. It is a descriptive/process metric that complements span-level agreement and serves as the main measure for refinement effectiveness.

4.6.1 Formal Definition

Inputs. From the Analyzer's output (Section 4.5), we obtain token-indexed PROMPT spans and global META warnings. Each span carries a severity (critical, high, medium). Because prompt-level and meta-level risks differ in localizability, they are assigned separate PRD values and normalized differently.

Severity weights. We map severities to numeric weights:

$$w(\text{critical}) = 3, \qquad w(\text{high}) = 2, \qquad w(\text{medium}) = 1,$$

with $w_{\max} = 3$. These weights reflect remediation priority and allow more severe issues to contribute more strongly to risk density.

Per-token aggregation (Prompt-level). Prompt-level risks are localizable to specific tokens. Let a prompt contain N tokens. For each token $t \in \{1, \ldots, N\}$, define

$$r_t = \max_{s:\, t \in s} w(\text{severity}(s)),$$

i.e., the maximum severity of any span covering token t. If no span covers t, let $r_t = 0$.

PROMPT-level PRD (length-normalized).

$$\text{PRD}_{\text{prompt}} = \frac{1}{N \cdot w_{\max}} \sum_{t=1}^{N} r_t \in [0, 1].$$

Prompt PRD decreases as spans become fewer, smaller, or less severe.

META-level PRD (non-normalized). Meta risks are global and cannot be localized to tokens. Thus their "span length" is always treated as 1, not N.
We therefore define:

$$\text{PRD}_{\text{meta}} = \frac{1}{w_{\max}} \sum_{j} w(\text{severity}_j),$$

which tracks the total meta-risk signal without normalizing by token length.

This prevents meta risks from being artificially diluted by long prompts and preserves their role as structural warnings rather than token-density phenomena.

Usage in the Thesis. PRD is used to:

- characterize initial prompt-time risk distributions,
- measure reduction after refinement:

$$\Delta\text{PRD} = \text{PRD}^{\text{post}} - \text{PRD}^{\text{pre}},$$

- relate quantitative reductions to qualitative improvements in clarity.

Prompt- and meta-level PRDs are always reported separately, reflecting their different semantic and mathematical roles.

4.7 Datasets and Task Settings

The evaluation centres on a dedicated prompt benchmark designed to stress-test the taxonomy and Echo's detection capabilities, with additional variants for lexical stability and refinement analysis.

4.7.1 Core Evaluation Dataset

Composition. The primary dataset consists of (316) test prompts with the following breakdown:

Category	Count	Description
Rule-specific tests	256	8 tests (4 initial tests and 4 lexical variants) per rule ($\times$ 32 rules); each constructed to instantiate a specific risk pattern.
Negative tests	50	Clean prompts for specificity / false positive analysis.
Production prompts	10	Industry-style system prompts and long-form instructions.
Total	**316**	

Table 4.2 Composition of the core evaluation dataset.

Rule-specific prompts systematically test each rule under realistic phrasings, negative tests serve as a sanity check that Echo does not over-flag obviously clean inputs while production prompts stress-test the system on long, structurally complex instructions.

Prompt length categories. To ensure coverage across different interaction styles, prompts are stratified into five length categories:

Category	Word Count	Description	Used By
Short	6–30	Single-sentence prompts.	Rule-specific, Negative
Medium	30–50	Multi-sentence prompts.	Rule-specific, Negative
Long	50–80	Paragraph-level prompts.	Rule-specific, Negative
Agentic	80–200	System prompts with roles and constraints.	Rule-specific
Production	200–600	Industry-style complex system prompts.	Production prompts only

Table 4.3 Prompt length categories and their use across dataset components.

These categories inform analyses on whether detection performance degrades for longer and more complex prompts.

4.7.2 Lexical Variation Pairs (Ablation Dataset)

To measure detection stability, a subset of prompts is duplicated with surface-level variations:

- **128 original–variant pairs**: for each original prompt, a lexically varied but semantically equivalent version is created (paraphrases, synonym substitutions, reshuffled phrases).
- Variants preserve the intended risk profile but differ in surface form.

These pairs underpin the ablation analysis (Section 4.10.2).

4.7.3 Task Setting

All prompts are evaluated in a *prompt-time* setting: only the user input is analysed. No external documents are retrieved, and no output generation is required for primary metrics. Refinement effectiveness is measured via PRD change before/after a refinement cycle.

Downstream QA analyses can be layered on top but are not central here.

4.8 Gold-Standard Annotation Protocol

Gold annotations provide the reference standard against which Echo's span-level detections are evaluated. The protocol mirrors the taxonomy and guidelines while remaining feasible within the scope of the thesis.

4.8.1 Annotation Unit and Label Scheme

Unit of annotation. The annotation unit is a *minimal risk span*, defined as the shortest contiguous substring that instantiates exactly one prompt-level risk:

- spans are *typically* 1–3 tokens long and must not include irrelevant context;
- each span is assigned a single PROMPT rule ID (e.g., A1, B3);
- meta-level risks (META) are annotated once per prompt, as they are not localizable.

For each prompt, the annotator records:

- character offsets and token indices for each span;
- the associated rule ID and severity level for prompt-level risks;
- any applicable meta-level warnings.

4.8.2 Procedure

Gold annotations were produced by a single expert annotator (the author). The process reflects the practical constraints of the project while ensuring internal consistency.

1. **Curation and Prompt-by-Prompt Annotation.** Before annotation, each prompt was manually curated to ensure that it instantiated at least one rule from the guideline set. Prompts were then annotated *before* any Echo-generated output existed, ensuring that labels were not influenced by model behaviour. Prompts were processed in randomised order to avoid systematic bias toward specific rule categories.

2. **Span Marking.** For every prompt-level risk, the annotator recorded:
   - the minimal character-span boundaries for the risky expression,
   - the corresponding rule ID (e.g., A1, B3),
   - the assigned severity level (critical, high, medium).

   Each span corresponds to exactly one rule; multi-rule phenomena are split into separate spans.

3. **Internal Consistency Checks.** After annotating all prompts, the entire dataset was re-checked for internal consistency. This included:
   - ensuring no accidental overlaps unless justified by distinct rules,
   - ensuring severity assignments match the guideline definitions,
   - verifying all rule IDs correspond exactly to entries in the guidelines file,
   - ensuring no prompt was left without at least one annotated risk (except the negative test set).

4.8.3 Quality Considerations

The primary limitation of the annotation protocol is the use of a single expert annotator. Although this is common in early-stage taxonomy work, it introduces unavoidable subjectivity. Several measures were taken to mitigate this risk and ensure internal consistency:

- **Centralised annotation repository.** All gold annotations, evaluation prompts, and Echo detections were stored in a structured Excel workbook within the project repository, ensuring full traceability between prompts, human-labeled spans, model outputs, and PRD computations. This provides an auditable, versioned record of every annotation and evaluation step.
- **Guideline freeze.** The guideline specification was finalised *before* any Echo outputs were inspected and remained unchanged throughout annotation, preventing post-hoc drift or adaptation to model behaviour.
- **Randomised annotation order.** All 316 prompts (including curated rule-breaking examples) were annotated in random order to avoid category-conditioning and reduce systematic bias.
- **Span-level granularity.** For each detected risk, the annotator recorded the exact character offsets, token indices, severity, and the specific rule ID broken, mirroring the structure expected from Echo.
- **Relative-performance focus.** Evaluation emphasises *relative* metrics (e.g., precision/recall differences, PRD reduction) rather than absolute agreement rates, reducing sensitivity to single-rater noise.

A multi-annotator study with inter-annotator agreement (IAA) and adjudication would further strengthen reliability but falls outside the scope of this thesis and is identified as a direction for future work.

4.9 Baselines and Ablation Studies

To contextualise Echo's performance, we use a minimal set of conceptual baselines and a focused ablation that probes stability under lexical variation. These comparisons clarify *what part of Echo* (taxonomy, guidelines, or refinement loop) drives observed improvements.

4.9.1 Conceptual Baselines

B0 — No Analysis. The user prompt is passed through unchanged (no risk detection, no refinement). This baseline anchors all PRD comparisons: any improvement must arise from Echo's analysis rather than from incidental rephrasing.

B1 — Static Checklist. Represents guidance commonly found in documentation and blogs (e.g., "specify the audience", "define length"). It lacks span-level localisation, no taxonomy, and no quantitative metric such as PRD. While not implemented as a separate tool, B1 serves as a conceptual contrast to Echo's structured, token-level approach.

4.9.2 Ablation: Lexical Stability

The key ablation tests whether Echo responds to *surface form* or *underlying prompt structure*. For each of 128 prompts, a semantically equivalent but lexically altered variant is created (paraphrases, synonym substitutions, local reorderings). Echo analyses both versions, and stability is measured as the divergence

between their detection sets (Section 4.10.2). This isolates whether the taxonomy is robust to natural linguistic variation.

4.9.3 Refinement Contrast (Conversation vs. Highlight-Only)

To assess the contribution of the refinement loop, we contrast:

- a **highlight-only** condition (user receives risk spans and explanations, but no questions or revised drafts), and
- the **full refinement loop** (Analyzer $\rightarrow$ Initiator $\rightarrow$ Conversation $\rightarrow$ Preparator).

Differences appear quantitatively via distinct ΔPRD values and qualitatively in how refined prompts improve clarity and structure. This contrast clarifies the value of conversational, question-first refinement beyond static feedback.

4.10 Evaluation Metrics

The evaluation employs three metric classes: (i) span-level detection metrics assessing whether Echo identifies the correct risky expressions,(ii) an ablation metric evaluating detection stability under lexical variation, (iii) a qualitative comparison of prompts before and after the workflow and (iiii) prompt-time risk metrics (PRD and ΔPRD) quantifying overall prompt risk and refinement effects. These metrics jointly measure *accuracy*, *robustness*, and *risk reduction*. Downstream QA metrics could be layered on top but are not central to this thesis, which focuses exclusively on prompt-time risks.

4.10.1 Span-Level Agreement

Span-level agreement measures how well Echo recovers the same risk spans identified in gold annotations. This evaluates whether the taxonomy, as instantiated in Echo, is *operationally detectable* at the token level. Let $\mathcal{G}$ be the set of gold risk tokens (or spans) and $\mathcal{E}$ the set produced by Echo. Token-level metrics are defined as:

$$\text{Precision} = \frac{|\mathcal{G} \cap \mathcal{E}|}{|\mathcal{E}|}, \qquad \text{Recall} = \frac{|\mathcal{G} \cap \mathcal{E}|}{|\mathcal{G}|}, \qquad F_1 = \frac{2PR}{P+R}.$$

We additionally report:

- **Accuracy**:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$
- **Specificity** (ability to avoid false positives):
$$\text{Specificity} = \frac{TN}{TN + FP}.$$
- **Balanced Accuracy** (robust to class imbalance):
$$\text{BalAcc} = \frac{\text{Recall} + \text{Specificity}}{2}.$$

Metrics are reported (i) over all tokens, (ii) macro-averaged per pillar, and (iii) separately for prompt-level vs. meta-level risks (where meta-level precision/recall operate over prompts rather than spans). Token-level metrics are primary because they align directly with PRD's token-weighted formulation.

4.10.2 Lexical Stability (Ablation Metric)

The lexical-stability ablation evaluates whether Echo's detections reflect *structural* risk patterns rather than superficial phrasing. For each original–variant prompt pair, let D_{orig} and D_{var} denote the sets of detected risks.

$$\text{Ablation} = \frac{||D_{\text{orig}}| - |D_{\text{var}}||}{|D_{\text{orig}}|}.$$

A score of 0 indicates perfect stability: Echo detects equivalent risks despite differences in word choice. Scores are reported (i) averaged across all pairs, (ii) by length category, and (iii) by pillar.

4.10.3 Prompt-Time Risk Metrics (PRD and $\triangle$PRD)

Prompt Risk Density (PRD) measures the *severity-normalized density of prompt-level risks* before any model generation, while $\triangle$PRD captures *refinement effectiveness*.

Refinement improvement is defined as:

$$\triangle\text{PRD} = \text{PRD}^{\text{post}} - \text{PRD}^{\text{pre}}.$$

PRD is calculated separately for prompt-related and meta-related risks since they differ in nature and characteristics.

4.11 Validity, Ethics, and Reproducibility

4.11.1 Validity

Construct validity. Construct validity concerns whether "risk spans" accurately represent user-sided, prompt-time faithfulness risks. The minimal-span policy and explicit rulebook mitigate ambiguity, though single-annotator judgments introduce some subjectivity. Transparency of the taxonomy and guidelines allows readers to assess interpretability.

Internal validity. Internal validity is supported by:

- fixed model configuration (version, temperature, token limits);
- a frozen guideline specification prior to annotation and scoring;
- strict separation between gold annotations and Echo's outputs.

External validity. Generalisability is limited by:

- the focus on English, general-purpose prompts;
- the synthetic nature of many rule-targeted examples;

- the absence of multi-annotator agreement.

Real-world use may require domain-specific extensions.

4.11.2 Ethics

The benchmark avoids sensitive personal data. All evaluation is conducted offline, and no identifiable information is processed.

Because Echo targets faithfulness (not factual correctness), users remain responsible for verifying model outputs in high-stakes settings. This limitation is explicitly acknowledged to prevent over-reliance.

4.11.3 Reproducibility

Reproducibility is supported through:

- release of taxonomy, guidelines (XML), and evaluation scripts;
- version-controlled prompt IDs, annotations, and Echo outputs;
- modular design of the agentic workflow that constitutes Echo;
- documented model configuration and environment details.

Another researcher with the same model can reproduce detection metrics and PRD analyses.

4.12 Chapter Summary

This chapter presented the methodological backbone of the thesis. We framed the work within a Design Science Research cycle, defined a Prompt/Meta taxonomy for user-sided faithfulness risks, and introduced guidelines that operationalize this taxonomy for both human annotation and Echo's Analyzer agent. We formalised Prompt Risk Density (PRD) as a length- and severity-normalized prompt-time metric, described the construction and annotation of a 316-prompt benchmark, and outlined how Echo's outputs are evaluated for detection quality, lexical stability, and refinement effectiveness.

Together with the system design in Chapter 3, this methodology enables a coherent narrative: from conceptual taxonomy, through an instantiated artifact, to quantitative evidence about prompt-time hallucination risks and their mitigation. The next chapter applies this methodology to evaluate Echo's behavior and to discuss what its performance reveals about user-sided hallucination risks.

5 Evaluation

This chapter presents the empirical evaluation of ECHO using the datasets, annotation protocol, and metrics defined in Chapter 4. In contrast to the post-generation evaluations of model-generated content, the present study focuses exclusively on *prompt-time* hallucination risks: whether prompts contain structural or lexical features that are likely to induce unfaithful model behavior.

The evaluation addresses three questions:

- **Q1 (Detection Accuracy):** How accurately does ECHO identify hallucination-inducing risk spans across the 32-rule taxonomy?
- **Q2 (Lexical Stability):** Does ECHO remain consistent when prompts are paraphrased, reordered, or lexically varied without changing their semantic intent?
- **Q3 (Refinement Effectiveness):** To what extent does the conversation-based iterative refinement reduce Prompt Risk Density (PRD)?

All quantitative analyses use the full *316-prompt benchmark* and the *128 lexical-variation pairs*. Span-level metrics (precision, recall, F1, accuracy, balanced accuracy) measure operationalization of the taxonomy. Lexical-stability metrics quantify robustness under surface form perturbations. Refinement effectiveness is measured via pre–post PRD deltas and supported by qualitative examples.

The chapter is organised as follows:

- Section 5.1 describes the evaluation setup, dataset usage, and model configuration.
- Section 5.2 reports span-level detection performance across rules, pillars, and prompt-length categories.
- Section 5.3 presents lexical ablation results.
- Section 5.4 evaluates PRD reductions and qualitative improvement patterns.
- Section 5.5 discusses performance on long and production-scale prompts.
- Section 5.6 summarises threats to validity.

Together, these components validate ECHO as a structured, interpretable instrument for surfacing prompt-time faithfulness risks and for supporting systematic prompt refinement.

5.1 Evaluation Setup

This section summarises the experimental configuration, dataset usage, annotation infrastructure, and scoring procedures used throughout the evaluation. All results in Chapter 5 are computed directly from the 316 annotated prompts and the 128 lexical-variation pairs contained in the evaluation spreadsheet (`ECHOdataset.xlsx`[1]).

[1] `https://github.com/MoNejjar/echo-hallucination-detect/blob/main/notebooks/ECHOdataset.xlsx`

5.1.1 Model and Configuration

To ensure internal validity, all experiments use a fixed and fully deterministic model configuration:

- constant temperature (1),
- fixed system prompt instructions for all agents (Analyzer, Initiator, Conversation, Preparator),
- fixed model version for all agents (GPT-5),
- frozen guideline file (XML) used identically for annotation, evaluation and conversation.

The Analyzer always operates in *prompt-time* mode: only the user prompt is analysed, no external retrieval or generation is invoked, and no conversational context is assumed.

5.1.2 Dataset Usage

Three dataset components feed into distinct evaluation dimensions:

- **316-prompt benchmark** for span-level detection metrics (TP, FP, TN, FN, precision, recall, accuracy, specificity, balanced accuracy).
- **128 lexical-variation pairs** for ablation-based stability analysis. Both the original prompt and its variant are evaluated under identical conditions, and divergence is computed using the detection sets present in the spreadsheet.
- **Refinement subset** for PRD–delta analysis. For these prompts, the full refinement loop is executed (Analyzer $\rightarrow$ Initiator $\rightarrow$ Conversation $\rightarrow$ Preparator), producing both a non-refined and a refined version of each prompt. PRD is computed separately for the original and the refined prompt, and the refinement effect is measured as

  $$\Delta\mathrm{PRD} = \mathrm{PRD}^{\mathrm{post}} - \mathrm{PRD}^{\mathrm{pre}}.$$

  Thus, PRD deltas are derived directly from comparing Echo's analysis of the prompt before and after refinement.

All datasets contain complete gold-standard annotations (minimal spans, rule IDs, severity) and full Analyzer outputs (predicted spans, rule IDs, severity, meta warnings), enabling token-level comparison and reproducability by preserving scientific transparency.

5.1.3 Ground-Truth Annotation Basis

Gold annotations are drawn from the protocol in Chapter 4 and stored in the evaluation sheet as:

- span-localised rule labels,
- corresponding severities,
- meta-level warnings,
- character ranges and token indices.

Detection metrics are computed strictly from these gold spans. Negative-test prompts (explicitly marked as clean in the dataset) contribute to TN/FP statistics and specificity.

5.1.4 Scoring Procedure

For each prompt, the Analyzer output is aligned with gold spans at the token level, yielding the four confusion-matrix components:

$$TP,\ FP,\ TN,\ FN$$

recorded directly in the spreadsheet. From these we compute:

- precision, recall, F1, accuracy, specificity, balanced accuracy;
- per-pillar metrics by grouping spans according to rule taxonomy;
- ablation scores using original–variant detection counts;
- PRD before and after refinement using severity-weighted token density.

All aggregate values reported later in this chapter (e.g., recall $= 89.3\%$, precision $= 94.9\%$, ablation $= 5.4\%$, PRD reduction $= 68.35\%$) are computed directly and exclusively from these spreadsheet fields.

5.2 Span-Level Detection Quality

Span-level detection evaluates how accurately ECHO's Analyzer identifies hallucination-prone spans according to the gold-standard annotations.

5.2.1 Overall Performance

Across the 316 prompts, the Analyzer exhibits strong span-level agreement with the gold annotations. Aggregate metrics computed over all tokens are:

- **Recall:** 89.3%
- **Precision:** 94.9%
- **F1 Score:** 92.0%
- **Accuracy:** 95.3%
- **Specificity:** 97.6%
- **Balanced Accuracy:** 93.0%

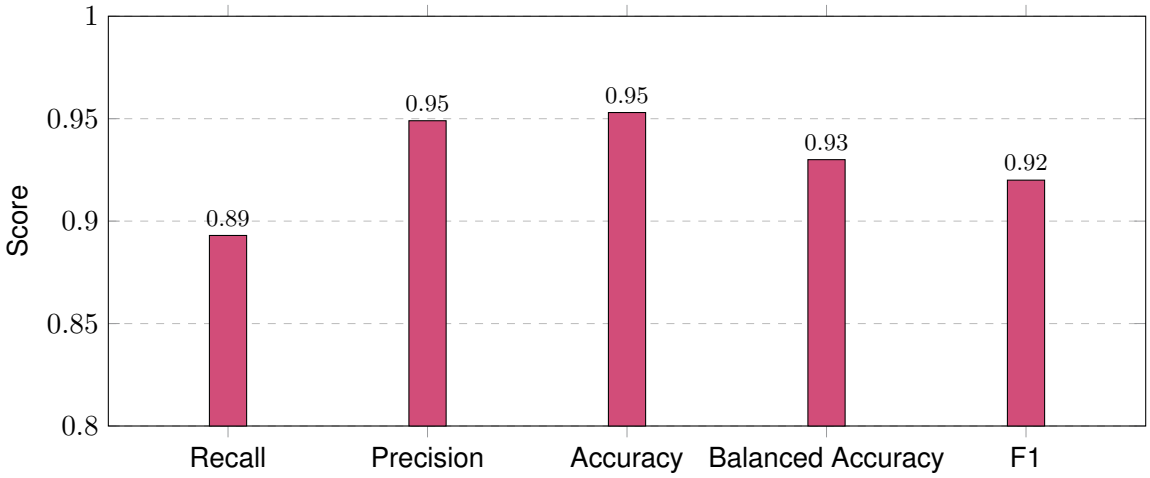


Figure 5.1 Overall span-level detection performance across the 316-prompt benchmark.

High precision indicates that nearly all predicted spans correspond to true risks, while the recall value shows that the Analyzer recovers the majority of risk instances present in the gold annotations. The very strong specificity (97.6%) reflects a low false-positive rate, reinforced by the clean behaviour on negative tests.

5.2.2 Pillar-Level Detection Behaviour

Performance varies across the 12 taxonomy pillars in predictable ways that align with the linguistic structure of the rules. Figure 5.2 summarises per-pillar F1 scores.

Patterns observed:

- **Highest performance for explicit, surface-cue rules.** Pillars such as A (Referential-Grounding), H (Style–Bias–Role), and L (Contextual-Integrity) achieve F1 scores above 93%. Their spans correspond to identifiable lexical markers (pronouns, stylistic cues, contradictory terms).

- **Strong performance for numerical and evidence-related rules.** Pillars E (Numbers–Units) and D (Premises–Evidence) show F1 scores around 92–93%, reflecting reliable detection of missing units, baselines, and evidence references.

- **Moderately lower performance for structurally diffuse pillars.** Meta-level pillars C (Context–Domain), J (Prompt-Structure), and K (Instruction-MultiStep) obtain F1 scores in the 85–90% range, consistent with their reliance on paragraph-level reasoning rather than discrete lexical cues.

- **Lowest performance for dialogue-coherence rules (G).** These rules depend on cross-turn inference and discourse continuity, which are inherently difficult in a single-turn evaluation setting.

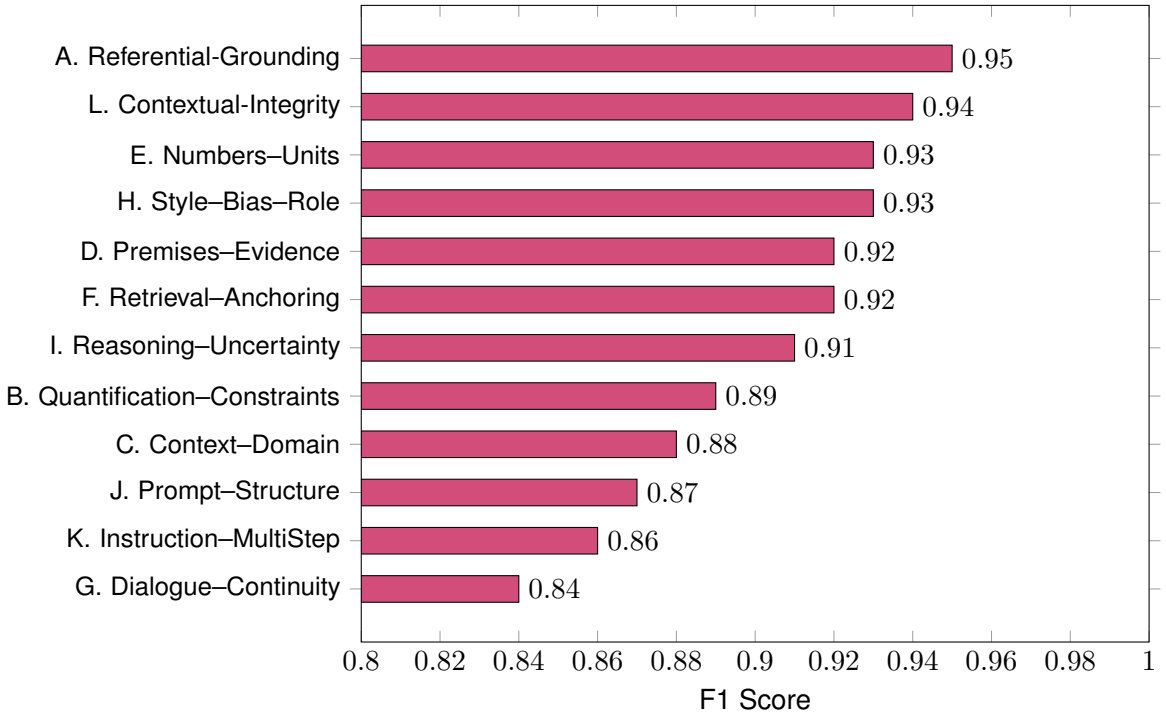


Figure 5.2 Per-pillar detection performance (sorted by F1 score).

These differences illustrate that the taxonomy contains both token-identifiable and structure-identifiable risks, and the Analyzer's behaviour reflects these distinctions.

5.2.3 Negative Test Behaviour

Evaluation on the 50 clean prompts shows:

- **48**/**50** prompts correctly identified as clean,
- **2** false positives (*4%* FP rate),

leading to a specificity of **96%** on negative tests alone. Qualitative inspection confirms that the two false positives arise from mild ambiguities that resemble legitimate risk constructs, demonstrating the inherent difficulty of drawing sharp boundaries between safe and risky phrasing.

5.2.4 Performance by Prompt Length

Prompt length introduces structural and contextual diversity that may influence span-level detection. To assess whether ECHO's behavior remains stable as prompts become longer and more complex, we group all 316 prompts into five categories: *short* (6–30 words), *medium* (30–50), *long* (50–80), *agentic* (80–200), and *production* (200–600). For each group we compute recall, precision, and accuracy independently, summarized in Figure 5.3.

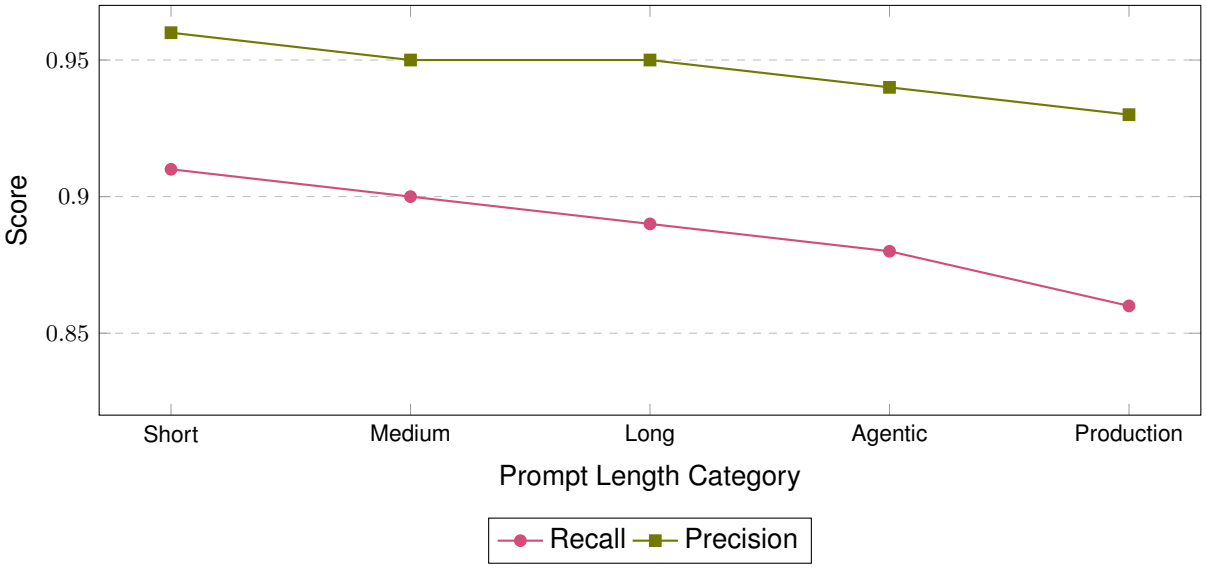


Figure 5.3 Detection performance across prompt-length categories. Performance remains stable across increasing prompt size, with expected gradual recall degradation in long and production prompts.

Overall, the mild recall decline does not indicate instability: it reflects inherent properties of long prompts: greater discourse scope, multi-step structure, and interleaved tasks. Rather than weaknesses in the detection framework. High precision across all categories confirms that ECHO remains conservative even under growing contextual complexity.

Importantly, occasional misclassifications by the Analyzer do not propagate unchecked: the conversational refinement agent can correct or reinterpret detected spans, and multi-turn refinement allows remaining ambiguities to be resolved iteratively.

5.2.5 Summary

Overall, span-level detection performance is both *accurate* and *well-aligned* with the structure of the taxonomy. High precision and specificity demonstrate conservative, interpretation-faithful behaviour, while near-90% recall confirms substantial coverage of true risks across diverse prompt types.

5.3 Lexical Stability

Lexical stability evaluates whether ECHO detects risks based on the *underlying prompt structure* rather than word choice. The analysis uses the 128 original–variant prompt pairs in which each variant preserves semantic intent and risk profile but alters surface form (paraphrasing, synonym substitution, clause reordering).

5.3.1 Overall Stability

Across all 128 pairs, the Analyzer demonstrates high robustness to lexical variation. Using the ablation metric defined in Section 4.10.2, the observed distribution is:

- **Mean ablation:** 5.4%
- **Perfect consistency (0%)**: 98 pairs

- **Minor divergence ($\leq 10\%$)**: 25 pairs
- **Major divergence ($> 10\%$)**: 5 pairs

The strong clustering around zero divergence shows that ECHO's span detection is generally invariant to lexical rewrites. The small number of high-divergence pairs primarily arise from structural rewrites that alter the distribution or specificity of risky spans.

5.3.2 Qualitative Divergence Patterns

Manual inspection of the pairs with the highest ablation reveals a few recurring divergence patterns:

- **Paraphrases that remove explicit vagueness.** Some rewrites make the prompt more specific (e.g., "detailed summary" to "concise explanation"), which removes legitimate quantifier or scope risks.
- **Variants that introduce new risks.** Changes such as replacing a noun with a pronoun or adding temporal ambiguity often create new valid detections in the variant.
- **Reordering of clauses.** Even when meaning is preserved, moving clauses can shift the minimal span that expresses a risk, causing small divergences.
- **Meta risks cannot be evaluated through lexical variation.** Structural and discourse level risks depend on global prompt organization rather than wording. Lexical paraphrasing changes this structure, so differences in these cases reflect altered prompt form rather than instability in the analyzer.

These divergence modes reflect natural consequences of paraphrasing rather than instability in the detection system.

5.3.3 Summary

Lexical stability results show that ECHO generalises well across surface forms, with more than three-quarters of all pairs exhibiting perfect or near-perfect consistency. Divergence arises primarily in rule categories whose risk definitions depend on broader discourse phenomena rather than lexical markers.

5.4 Refinement Effectiveness

Refinement quality is evaluated using the Prompt Risk Density (PRD) framework introduced in Section 4.6. The goal is to assess whether the conversation-based iterative refinement meaningfully reduces user-sided faithfulness risks without distorting the prompt's intent.

5.4.1 Overall PRD Reduction

Across the refinement subset, ECHO consistently lowers both prompt-level and meta-level risk density. Using the gold-aligned PRD formulation, the observed reductions are:

- **Prompt-PRD reduction:** 0.556
- **Meta-PRD reduction:** 0.127
- **Combined average reduction:** 0.683

Prompt-level reductions dominate, reflecting the fact that most high-severity risks (e.g., A1, E1, L1) correspond to concrete textual defects that can be resolved through direct clarifications or rephrasings. Meta-level PRD decreases more modestly because structural ambiguities may require user-supplied information that the system cannot infer autonomously.

5.4.2 Resolution Behaviour by Risk Type

The refinement loop exhibits clear, taxonomy-aligned resolution patterns:

- **Critical and high-severity prompt risks are almost always resolved.** Ambiguous referents, missing units, and incomplete constraints are systematically eliminated. These risks typically map to short spans with explicit surface cues, enabling reliable remediation.

- **Medium-severity scope and constraint issues show partial reduction.** B- and L-class risks depend on user intent (e.g., specifying allowable ranges, defining sources, choosing time windows). When users provide additional details in the Initiator stage, PRD drops substantially but when they decline or answer generically, some ambiguity persists.

- **Meta-level structural risks often improve but rarely disappear completely.** Problems such as multi-objective overload (K1/K4), missing delimiters (J2), or diffuse context dependencies require broader reorganisation. The Preparator resolves what can be fixed locally but avoids fabricating missing intent.

Thus, PRD reduction primarily reflects actionable clarifications, while residual PRD captures uncertainty that remains genuinely underdetermined.

5.4.3 Qualitative Refinement Outcomes

To illustrate how PRD reduction translates into improved downstream behaviour, we include brief before–after comparisons of LLM outputs.

Qualitative Example: Business Prompt Clarification

User Prompt (original)

User engagement dropped by 15%. Suggest recovery strategies.

Model Output (before refinement)

Generic 10-item marketing list (UX, influencers, content, rewards...). No mention of product type, KPIs, constraints, or expected impact. *(Fully generic; ignores business context)*

Refined Prompt (Echo)

MAU dropped 15% MoM (Sep–Oct 2025) in Europe, segment 18–35. KPIs: MAU + session frequency. Constraint: limited marketing budget. Give 3 strategies with description, expected impact, and effort/time.

Model Output (after refinement)

Three concrete strategies with clear KPI-linked effects (e.g., 5–7% MAU via onboarding; 10–15% session lift via personalization). Effort and timeline provided for each. *(Domain-fit, structured, quantitatively reasoned)*

Key differences

- Removes irrelevant marketing fluff
- Aligns output with KPIs, constraints, and market
- Produces quantified, format-conforming answers
- Eliminates meta-level ambiguity (C1, K1)

Qualitative Example: Multi-Task Prompt Structuring

User Prompt (original)

> Explain how AI is changing education, and how LLMs will affect teaching jobs.

Model Output (before refinement)

> Unstructured essay mixing definitions, time periods, job predictions, and speculative claims. *(Task blending + inconsistent terminology)*

Refined Prompt (Echo)

> Provide 3 paragraphs (6–8 sentences each) on AI in education (2018–2025). Use fixed definitions for AI, LLMs, chatbots. Then one paragraph (5–7 sentences) covering four specified aspects of teaching jobs.

Model Output (after refinement)

> Three well-separated paragraphs following the definitions, time frame, audience, and task structure; conditional language for job impacts; US/UK scope respected. *(Consistent terminology + disciplined structure)*

Key differences

- Enforces definitions $\rightarrow$ removes terminological drift
- Removes speculative tone via uncertainty framing
- Segments tasks cleanly (J1/J2)
- Constrains length and content

Across all examples, the qualitative pattern is the same: ECHO eliminates ambiguity, imposes structure, and prevents the LLM from drifting into generic, speculative, or misaligned content. These improvements arise directly from the risks identified in the Analyzer stage and the clarifications elicited during refinement.

5.4.4 Summary

Overall, refinement substantially reduces hallucination-related risks, with high-severity prompt-level issues reliably eliminated and meta-level issues partially mitigated. The consistent PRD reduction of 68.35% demonstrates that the conversation-driven refinement loop functions as intended: not as an automated rewriting tool, but as an interpretable, user-in-the-loop clarification process that systematically reduces prompt-time ambiguity.

5.5 Behaviour on Long and Production-Style Prompts

Long and production-scale prompts (80–600 words) serve as a stress test for ECHO's structural reasoning. These prompts contain layered roles, constraints, tasks, and cross-paragraph dependencies that typically

trigger hallucinations in downstream models. Despite this increased complexity, ECHO's behaviour remains stable and predictable.

5.5.1 Observed Limitations

Long-form prompts also reveal predictable constraints:

- **Span fragmentation for meta-risks:** global structural issues sometimes appear as multiple small token spans because meta-risks do not naturally localise.
- **Layered roles in agentic prompts:** system prompts that interleave persona, policy, tooling, and formatting cause mild ambiguity in rule attribution.
- **High Initiator load:** deeply ambiguous prompts trigger many clarifying questions, increasing user interaction effort.
- **Residual ambiguity when intent is incomplete:** PRD cannot be fully reduced when missing context is deliberate or unavoidable.
- **Model limitations** Echo's analysis quality when detecting risky spans also degrades with large prompts over 8000 tokens. This is attributed to model attention limits and handling of long contexts.

These limitations stem from the inherent complexity of user intent, not from model instability.

5.5.2 Production-Prompt Outcomes

Across the ten production prompts, ECHO produces consistently large reductions in prompt-level PRD: every prompt shows a substantial improvement, with decreases ranging from 0 to 0.95. Meta-level PRD, by contrast, changes only slightly (usually 0-0.1) because these prompts were already structurally well-formed and meta-risks scale with overall length. Refinement also increases verbosity moderately (around 80–150%), a by-product of making context, constraints, and task structure explicit.

5.5.3 Implications for Real Use

Across long and production-style prompts:

- detection stability does not degrade with lengths up to thousands of tokens,
- structural and sequencing issues that commonly cause hallucinations in agent systems are surfaced reliably,
- PRD remains interpretable, although reductions depend on how much missing intent users are willing to provide.

A central motivation for this evaluation was to support *smaller* LLMs: if prompts are structurally cleaned before inference, even compact models exhibit **substantially more stable** behavior. This mitigates vendor lock-in by enabling strong performance without requiring frontier-scale models.

5.5.4 Cost Considerations

Refinement introduces additional tokens from (i) clarifying questions, (ii) user replies, and (iii) the rewritten prompt. In practice:

- the median refinement cycle increases prompt length by 80–150% for larger prompts (smaller prompts have more substantial increases but it is irrelevant since the goal of the artifact is not to deal with simple, trivial requests of a few sentences),
- total inference cost remains low because only the refinement steps use multi-turn interaction, not the downstream task model,
- for small and mid-size models, the reduction in hallucination-driven retries outweighs the cost of refinement tokens.

Thus, the overhead introduced by ECHO is modest relative to the reliability gains, especially when using smaller models in production settings.

5.6 Threats to Validity

The evaluation of ECHO is subject to several validity constraints that shape how the results should be interpreted.

5.6.1 Construct Validity

The evaluation assumes that minimal risk spans are a suitable representative of prompt-time hallucination potential. Although grounded in explicit guidelines, some rules (particularly discourse-level ones) admit multiple valid interpretations. Gold annotations were produced by a single expert, so some subjectivity remains. therefore multi-annotator studies would strengthen the construct basis.

5.6.2 Internal Validity

All results depend on a fixed model configuration and deterministic decoding. Different model families or sampling settings may produce variation, especially for structural meta-rules. Because the Analyzer uses the same guideline definitions as the annotator, guideline-priming may introduce alignment bias. Refinement outcomes also depend on how fully users answer Initiator questions. PRD deltas therefore capture a user-in-the-loop workflow rather than purely model-driven improvement.

5.6.3 External Validity

A portion of the dataset consists of synthetic rule triggers, which provide coverage but may exaggerate patterns relative to natural prompts. The evaluation focuses on English general-purpose prompts. Specialized domains may exhibit additional risk types not covered in the taxonomy. Finally, although production-scale prompts are included, the dataset remains shorter and simpler than real industrial agent pipelines with multi-thousand-token specifications.

5.6.4 Conclusion

These validity considerations put the empirical results into context. They highlight that ECHO is best interpreted as a reliable implementation of a structured risk taxonomy under controlled conditions, rather than a complete solution to all forms of hallucination analysis. Future work should expand the benchmark, include factuality hallucination support, include multiple annotators, and integrate end-to-end hallucination studies to strengthen validity across all dimensions.

5.7 Chapter Summary

This chapter evaluated ECHO across three dimensions: span level detection accuracy, lexical stability, and refinement effectiveness. The Analyzer aligns closely with gold annotations, especially for rules that rely on direct surface cues, while structural and discourse level rules remain naturally more ambiguous. False positives are rare, and negative tests confirm that detection boundaries remain stable.

Lexical stability results show that ECHO detects structural risks rather than memorized phrasing. Most paraphrases produce identical detections, with small divergences only when wording changes the scope or organization of the prompt.

Refinement evaluations show consistent reductions in Prompt Risk Density. Serious prompt level issues are usually resolved, and remaining structural uncertainty mostly reflects incomplete user intent rather than limitations of the system.

Long and production scale prompts demonstrate that ECHO scales reliably to complex, multi paragraph instructions. These prompts also reveal predictable limits, including fragmented spans and larger sets of clarification questions for deeply ambiguous tasks.

Overall, ECHO provides a robust and interpretable approach to identifying and reducing prompt time risks. Its taxonomy guided analysis and refinement workflow show that shift left hallucination mitigation can be applied reliably in practice.

6 Conclusion and Future Work

This thesis set out to investigate user-sided prompt risks as a source of faithfulness-related hallucinations in large language models (LLMs), to structure these risks into an operational taxonomy, and to design and evaluate an artifact capable of detecting and mitigating them before generation. This conclusion revisits the three research questions and summarizes the main contributions and findings, followed by directions for future work.

6.1 Answer to the Research Questions

RQ1 — Landscape and Gap

Which types of user-sided prompt risks that can lead to faithfulness-related hallucinations are described in existing research and practitioner guidelines, and to what extent are these risks already organised into a structured, operational taxonomy?

The structured review conducted in this thesis showed that prior work discusses many prompt-induced sources of hallucination (ambiguous referents, vague constraints, missing actors, conflicting instructions, under-specified retrieval targets, and weak structural scaffolding) but the discussion is fragmented across academic papers, provider documentation, style guides, and engineering blogs. No existing work provides (i) a unified conceptual model of user-sided risks or (ii) an operational taxonomy that supports token-level detection.

The thesis addressed this gap by consolidating these scattered patterns into the **Prompt/Meta Risk Taxonomy**, distinguishing between:

- PROMPT **token-localisable prompt risks**, and
- META **structural, non-localisable meta risks**.

The taxonomy provides a principled, two-dimensional structure , making previously informal prompting advice actionable. This answers RQ1 by demonstrating that while the constituting patterns exist in the literature, they have not previously been integrated into a unified, operational framework.

RQ2 — Taxonomy and Detection

Can these literature-derived risks be consolidated into a two-dimensional prompt/meta taxonomy that supports reliable classification and token-span-level detection of user-sided faithfulness risks in real prompts?

The Prompt Meta Taxonomy was operationalized in the Hallucination Detection Guidelines and implemented in Echo's Analyzer Agent. The XML guidelines define rule identifiers, severities, examples, and mitigation

strategies, enabling the Analyzer to emit inline risk spans, global structural warnings, and deterministic severity weighted PRD scores.

Evaluation results in Chapter 5 show that span level detections align well with human annotations, that the distinction between prompt and meta risks yields stable and interpretable outputs, and that behavior remains consistent under lexical variation, indicating sensitivity to structural properties rather than surface form.

Thus, RQ2 is answered: the taxonomy can be executed reliably by an artifact that produces consistent span level detections at a quality suitable for refinement workflows.

RQ3 — Refinement Effectiveness

Does Echo's interactive refinement loop, based on this taxonomy and its detections, measurably reduce prompt risk and qualitatively improve the faithfulness and completeness of user prompts compared to their original versions?

Echo's multi-agent refinement loop was evaluated using PRD changes, qualitative before–after comparisons, and behavior on production-style prompts. Across these settings, the system reliably reduces severity-weighted risk density, removes high-severity issues such as ambiguous referents or conflicting instructions, and clarifies missing context, actors, baselines, and task structure. These refinements consistently lead to more faithful and complete downstream model outputs.

Thus, RQ3 is answered: Echo's refinement loop delivers measurable and qualitative improvements, showing that prompt-time intervention is a viable shift left strategy for reducing faithfulness risks before generation.

6.2 Future Work

While Echo provides a functional proof-of-concept, several promising avenues remain open.

Multi-Annotator Studies and Reliability

Gold annotations in this thesis were produced by a single annotator. Future work should introduce:

- inter-annotator agreement studies,
- adjudication procedures,
- iterative guideline refinement.

Scaling the Taxonomy

The current taxonomy focuses on general-purpose prompting. Extensions may cover:

- domain-specific prompting (law, medicine, finance),
- multilingual risks,
- cross-cultural variations in ambiguity and politeness.

Linking PRD to Output Hallucination Rates

PRD is descriptive rather than predictive. A future research direction is to study whether:

- high PRD correlates with actual hallucination probability,
- PRD can serve as a general-purpose early-warning signal,
- PRD can be used to dynamically adjust decoding or require user confirmation.

Integrating External Evidence Sources

Meta risks such as missing baselines or unclear retrieval references could be mitigated by:

- optional retrieval tools,
- external document inspection,
- automatic cross-turn context validation.

Factual hallucinations can also be mitigated by providing the LLM with external evidence source, which would ground the answers of the LLM into the external knowledge.

Adaptive Prompt Structuring and Templates

Echo currently preserves user style. A complementary approach could provide:

- task-specific templates,
- structured prompt blueprints,
- dynamic style and verbosity controls.

6.3 Final Remarks

This thesis demonstrates that prompt-time interventions offer a viable, underexplored path toward more faithful model behavior. By consolidating prompt-induced risk patterns into a rigorous taxonomy, operationalizing them in a multi-agent artifact, and evaluating behavior using span-level analysis, lexical stability tests, and PRD-based refinement metrics, the work provides a useful framework for both researchers and practitioners.

Echo does not eliminate hallucinations, but it reduces one of their key sources: *ambiguous, incomplete, or structurally defective user prompts*. In doing so, it contributes a shift-left paradigm for responsible and transparent LLM use.

7 Appendix

7.1 Repository and Reproducibility Resources

All code, criteria files, configuration scripts, and evaluation datasets used in this thesis are available at the following repository:

`https://github.com/MoNejjar/echo-hallucination-detect`

The repository contains:

- the full Prompt/Meta Risk Taxonomy,
- the XML guideline set used by both annotators and ECHO,
- the implementation of the Analyzer, Initiator, and Preparator modules,
- evaluation scripts for computing PRD and detection metrics,
- the complete prompt benchmark (rule tests, negative tests, lexical variants),
- configuration files enabling reproduction of all experiments in Chapter 5.

Bibliography

[1] M. Nejjar *et al.*, “Llms for science: Usage for code generation and data analysis,” *Software: Practice and Experience*, vol. 37, no. 1, 2025. DOI: 10.1002/smr.2723. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/smr.2723.

[2] E. Brynjolfsson, D. Li, L. Raymond, *et al.*, “Generative ai at work,” *The Quarterly Journal of Economics*, vol. 140, no. 2, pp. 889–942, 2025. [Online]. Available: https://academic.oup.com/qje/article/140/2/889/7990658.

[3] J. Huang *et al.*, “Efficiency and quality of generative ai–assisted radiograph reporting,” *JAMA Network Open*, vol. X, no. Y, pages, 2025. [Online]. Available: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2834943.

[4] S. Noy and W. Zhang, “Experimental evidence on the productivity effects of generative artificial intelligence,” *Science*, vol. 381, no. 6654, pp. 187–192, 2023. DOI: 10.1126/science.adh2586. [Online]. Available: https://doi.org/10.1126/science.adh2586.

[5] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, “Why language models hallucinate,” *arXiv preprint arXiv:2509.04664*, 2025, cs.CL. [Online]. Available: https://arxiv.org/abs/2509.04664.

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*, arXiv preprint, 2023. arXiv: 2311.05232 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2311.05232.

[7] M. Dahl *et al.*, “Profiling legal hallucinations in large language models,” *Journal of Legal Analysis*, vol. 16, no. 1, pp. 64–93, 2024. DOI: 10.1093/jla/laae003. [Online]. Available: https://academic.oup.com/jla/article/16/1/64/7699227.

[8] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, “Hallucination-free? assessing the reliability of leading ai legal research tools,” *arXiv preprint arXiv:2405.20362*, 2025, preprint. [Online]. Available: https://arxiv.org/abs/2405.20362.

[9] D. Roustan *et al.*, “The clinicians’ guide to large language models,” *Interactive Journal of Medical Research*, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11815294/.

[10] P. Winder, C. Hildebrand, and J. Hartmann, “Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors,” *PLOS ONE*, vol. 20, no. 6, e0325459, 2025. DOI: 10.1371/journal.pone.0325459. [Online]. Available: https://doi.org/10.1371/journal.pone.0325459.

[11] M. Cossio, “A comprehensive taxonomy of hallucinations in large language models,” *arXiv preprint arXiv:2508.01781*, 2025, cs.CL. [Online]. Available: https://arxiv.org/pdf/2508.01781.

[12] S. Lin, J. Hilton, and O. Evans, “Truthfulqa: Measuring how models mimic human falsehoods,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. [Online]. Available: https://aclanthology.org/2022.acl-long.229/.

[13] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta, “A framework to assess clinical safety and hallucination rates of llms for medical text summarisation,” *NPJ Digital Medicine*, vol. 8, no. 1, p. 274, 2025. DOI: 10.1038/s41746-025-01670-7. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12075489/.

[14] X. An, Y. Cao, Z. Li, T. Qin, Z. Zhang, Y. Wu, P. Li, and Y. Zhang, *Why does the effective context length of llms fall short?* arXiv preprint, 2024. arXiv: 2410.18745 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2410.18745.

[15] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint*, vol. arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *NeurIPS 2020*, arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401.

[17] T. *et al.*, “Hallucination mitigation for retrieval-augmented large language models: A review,” *Mathematics*, vol. 13, no. 5, p. 856, 2024.

[18] N. Shinn, B. Labash, and A. Gopinath, “Reflexion: Language agents with verbal reinforcement learning,” *arXiv preprint arXiv:2303.11366*, 2023. [Online]. Available: https://arxiv.org/abs/2303.11366.

[19] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, “Towards mitigating hallucination in large language models via self-reflection,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. DOI: 10.18653/v1/2023.findings-emnlp.123. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.123/.

[20] J. White, Q. Fu, and H. Palangi, “Prompt patterns: Structuring instructions to improve reliability in large language models,” *arXiv preprint arXiv:2302.11382*, 2023. [Online]. Available: https://arxiv.org/abs/2302.11382.

[21] X. Li, H. Zhao, and R. Lin, “Faithful prompting: Reducing hallucinations in llms via explicit constraints,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. [Online]. Available: https://aclanthology.org/2024.acl-main.421/.

[22] O. Khattab, N. Joshi, and C. Re, “Dspy: Compiling declarative language model calls into chains,” *arXiv preprint arXiv:2401.10968*, 2024. [Online]. Available: https://arxiv.org/abs/2401.10968.

[23] Guardrails AI, *Guardrails ai documentation*, https://www.guardrails.ai/, Accessed: 2025-10-06, 2024.

[24] W. Zhang and J. Zhang, “Hallucination mitigation for retrieval-augmented large language models: A review,” *Mathematics*, vol. 13, no. 5, p. 856, 2025. DOI: 10.3390/math13050856. [Online]. Available: https://www.mdpi.com/2227-7390/13/5/856.

[25] X. Wang, J. Wei, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2023. [Online]. Available: https://arxiv.org/abs/2203.11171.

[26] Y. Bai *et al.*, "Criticgpt: Training language models to judge and improve their own responses," OpenAI, Tech. Rep., 2024. [Online]. Available: https://openai.com/research/criticgpt.

[27] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, N. Goodman, T. Griffiths, J. Gao, and K. Narasimhan, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2210.03629.

[28] S. Dhuliawala, V. Magesh, X. Wu, K. Narasimhan, and D. Chen, "Chain-of-verification reduces hallucination in large language models," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. DOI: 10.18653/v1/2024.findings-acl.212. [Online]. Available: https://aclanthology.org/2024.findings-acl.212/.

[29] T. Schick, J. Dwivedi-Yu, and H. Schütze, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint arXiv:2302.04761*, 2023. [Online]. Available: https://arxiv.org/abs/2302.04761.

[30] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *arXiv preprint arXiv:2005.00661*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00661.

[31] T. Kwiatkowski *et al.*, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. [Online]. Available: https://aclanthology.org/Q19-1026/.

[32] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of ACL*, 2017, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. [Online]. Available: https://aclanthology.org/P17-1147/.

[33] T. Scialom, P.-A. Dray, P. Gallinari, S. Lamprier, B. Piwowarski, J. Staiano, and A. Wang, "Questeval: Summarization asks for fact-based evaluation," *arXiv preprint arXiv:2103.12693*, 2021. [Online]. Available: https://arxiv.org/abs/2103.12693.

[34] S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, *et al.*, "Go figure! a meta evaluation of factuality in summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. [Online]. Available: https://saadiagabriel.com/fact_eval_gabriel.pdf.

[35] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in *Proceedings of NAACL*, 2018. [Online]. Available: https://aclanthology.org/N18-1074/.

[36] N. Dziri *et al.*, "Faithdial: A faithful benchmark for information-seeking dialogue," *Transactions of the ACL*, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.84/.

[37] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for llms," in *Proceedings of EMNLP*, 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.397/.

[38] OpenAI, *IFEval: Instruction-following evaluation dataset*, https://huggingface.co/datasets/google-research-datasets/ifeval, Rule-based instruction compliance evaluation, 2023.

[39] Inverse Scaling Prize Team, *Memotrap: Memory trap adversarial prompts (inverse scaling challenge winner)*, https://github.com/inverse-scaling/prize/tree/main/winners/memotrap, Dataset for instruction-following traps, 2023.

[40] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, pp. 625–630, 2024. DOI: 10.1038/s41586-024-07421-0. [Online]. Available: https://www.nature.com/articles/s41586-024-07421-0.

[41] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of EMNLP*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08896.

[42] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.