Complex Control Systems ISSN 2603-4697 (Online)

A Framework for Reliable Evaluation under Uncertainty in Large Language Models: From Answer Generation to Interactional Reasoning

Momchil R. Rusev

Independent Researcher

Sofia, Bulgaria

mrusev@gmail.com

Abstract

Large language models (LLMs) are increasingly deployed in technical domains where evaluation tasks are inherently underspecified, context-dependent, and informed by heterogeneous sources of uneven reliability. In such settings, a central limitation is not only factual inaccuracy, but premature evaluative closure: the tendency to produce coherent conclusions before the underlying case has been sufficiently specified and the available evidence properly differentiated.

We argue that this behavior reflects structural pressures toward early coherence under uncertainty, rather than knowledge deficits alone. As a result, LLM outputs may over-weight discursively dominant but evidentially weak claims while under-attending to less salient yet epistemically stronger evidence. To capture this phenomenon, we introduce the Narrative Dominance Gap (NDG), defined as the divergence between discursive prevalence and evidential support.

To address these challenges, we propose a unified framework that reframes LLM interaction from answer generation to collaborative case evaluation. The framework integrates Adaptive Context Elicitation (ACE), which reconstructs missing decision-relevant context, and Context-Aware Epistemic Filtering (CEF), which structures and evaluates heterogeneous evidence relative to that context.

The paper contributes a conceptual account of discourse–evidence divergence, formalizes interactional and epistemic mechanisms for ambiguity-aware evaluation, and proposes an interaction-centered evaluation paradigm focused on problem specification, evidence transparency, and calibrated trust. More broadly, the findings suggest that improving LLM reliability may depend not only on model capabilities, but on how reasoning is structured within interaction, highlighting the role of system architecture and epistemic discipline in supporting robust evaluation under uncertainty. The work is conceptual and methodological, providing a foundation for future empirical validation.

Keywords: large language models (LLMs); human–AI interaction; technical decision support; ambiguity-aware evaluation; narrative dominance gap (NDG); adaptive context elicitation (ACE); context-aware epistemic filtering (CEF); evidence differentiation; evaluation under uncertainty; trust calibration; interaction-centered evaluation; system architecture; reasoning structure

I. Introduction

Large language models (LLMs) are increasingly integrated into technical decision-support contexts, including engineering, diagnostics, maintenance, procurement, and product evaluation. Their utility is often attributed to their ability to synthesize large volumes of heterogeneous information and present it in a fluent, accessible form.

However, this fluency can obscure a structural limitation. Many real-world technical queries are not well-formed information requests, but underspecified and context-dependent evaluative problems. Questions such as “Is this system reliable?” or “Is this a good choice?” do not admit meaningful answers in the abstract. Their interpretation depends on variables such as system condition, usage context, maintenance history, comparison frame, and the user’s actual decision objective.

In practice, these variables are often omitted. This omission may occur because relevant parameters are taken for granted, difficult to articulate, or not yet known to the user. As a result, initial queries frequently represent incomplete problem specifications rather than fully defined evaluation tasks.

At the same time, technical information environments are characterized by heterogeneous and unevenly reliable sources, including empirical measurements, regulatory standards, expert analysis, experiential reports, and widely circulated narratives. These sources differ in evidential strength, scope, and context dependence, yet they may be presented in similar discursive forms.

Within this setting, LLMs can produce outputs that are rhetorically coherent but epistemically under-specified. The limitation is not necessarily factual inaccuracy, but the compression of heterogeneous evidence and implicit assumptions into a single, fluent evaluative statement. This can obscure distinctions between empirical performance, normative adequacy, contextual suitability, and discursively dominant claims.

More broadly, LLM-based systems can be understood not only as sources of answers, but as evolving cognitive environments that shape how users interpret information, form judgments, and allocate attention over repeated interaction. In this sense, the epistemic risks discussed here extend beyond individual responses and relate to the structural properties of the interaction context itself.

A key contributing factor to this behavior is the sensitivity of LLMs to patterns present in their training data. Prior work suggests that model outputs are influenced not only by factual content, but also by distributional properties such as frequency, salience, and representational density. While these properties do not determine epistemic validity, they can shape which claims are more likely to be generated.

To capture this phenomenon, we introduce the **Narrative Dominance Gap (NDG)**, defined as the potential divergence between the discursive prevalence of an evaluative claim and the strength of its empirical or normative support. NDG is not introduced as an empirically validated mechanism, but as a conceptual tool for analyzing situations in which discursive prominence may exceed evidential grounding.

From this perspective, a central limitation of current LLM use in technical domains is not only epistemic, but interactional. Systems typically respond before the relevant case context has been sufficiently reconstructed. As a result, the model may produce an evaluation based on an implicit interpretation of the problem that does not correspond to the user's actual decision context.

We therefore propose a reframing: from answer generation to **collaborative case evaluation under uncertainty**. In this reframing, the system should treat the initial query as a partial specification and actively support the reconstruction of the underlying problem before evaluation takes place.

To operationalize this shift, we introduce two complementary mechanisms. **Adaptive Context Elicitation (ACE)** supports the identification and elicitation of missing, decision-relevant variables through structured interaction. **Context-Aware Epistemic Filtering (CEF)** supports the differentiation, alignment, and evaluation of heterogeneous evidence relative to the reconstructed context.

Together, NDG, ACE, and CEF define an interactional–epistemic framework for ambiguity-aware technical evaluation. NDG functions as a diagnostic lens, ACE as an interactional mechanism for context reconstruction, and CEF as an evaluative mechanism for structuring evidence.

A further observation motivating this work is that the limitations of LLM-based evaluation in technical domains may not be reducible to factual inaccuracy alone. When operating under incomplete or underspecified problem formulations, LLM systems can exhibit patterns of premature evaluative closure that are structurally similar to those observed in everyday human reasoning under uncertainty. In such cases, models may implicitly treat partially specified context as sufficient, gravitate toward discursively dominant or familiar explanations, underweight less salient but technically critical evidence, and converge on coherent conclusions before the underlying case has been adequately reconstructed. Importantly, this behavior does not necessarily arise from error in the narrow sense, but from the interaction between training on large-scale human-generated discourse and optimization toward fluent, coherent, and timely responses. From this perspective, the challenge is not only to improve correctness, but to redesign the interactional process such that context reconstruction and evidence differentiation precede evaluative closure. This shift reframes the problem from answer correctness to evaluation structure. More broadly, this suggests that improving the reliability of LLM-based systems may require not only better models, but also improvements in how reasoning itself is structured — both in human–AI interaction and within the processes through which models generate conclusions. This implies that reliability is not solely a property of model outputs, but of the interactional structure through which evaluation is constructed. In this sense, NDG is introduced not only as a lens on discursive bias, but as a diagnostic indicator of early closure under uncertainty.

The contributions of this paper are threefold. First, it introduces NDG as a conceptual lens for analyzing discursively influenced evaluation in LLM outputs. Second, it formalizes ACE and CEF as mechanisms for handling underspecified queries and heterogeneous evidence. Third, it outlines an interaction-centered evaluation paradigm in which system performance is assessed not only in terms of answer correctness, but also in terms of problem specification quality, transparency of evidential structure, trust calibration, and support for human judgment.

The paper is conceptual and methodological in nature. Its goal is to articulate design principles for more reliable human–LLM interaction in technical domains and to provide a foundation for future empirical work.

The proposed framework is not a standalone conceptual construct but emerges from a series of studies spanning economic theory, engineering analysis, and human–machine interaction.

II. Related Work and Positioning

A.2.1 Reliability, truthfulness, and epistemic limits of LLMs

A growing body of research has examined the reliability of large language models (LLMs), particularly their tendency to produce fluent but epistemically uncertain outputs. These issues are commonly discussed in terms of hallucination, truthfulness limitations, and overconfidence (Bender et al., 2021; Lin et al., 2022). Empirical benchmarks such as TruthfulQA (Lin et al., 2022) suggest that models can reproduce widely circulated misconceptions or misleading claims, especially when such claims are prevalent in training data.

Taken together, these findings indicate that LLM outputs are influenced not only by factual knowledge, but also by distributional properties of the data on which they are trained, including frequency, salience, and representational density of claims.

While existing work does not directly establish a causal relationship between discursive prevalence and evidential validity, it suggests that widely repeated or salient claims may be more likely to be generated, regardless of their epistemic strength.

In this sense, model outputs can reflect statistical regularities in discourse alongside, and sometimes independently from, structured evidence. The resulting limitation is not necessarily factual fabrication, but a form of epistemic compression: the collapse of heterogeneous evidence, uncertainty, and context-dependence into a single coherent narrative.

This observation is consistent with findings in cognitive psychology, such as the "illusory truth effect," where repeated exposure increases perceived truthfulness (Hasher et al., 1977; Fazio et al., 2015). Importantly, this parallel should be understood as an analogy rather than a shared mechanism. LLMs do not form beliefs; however, their training on large-scale human-generated corpora may encode statistical regularities that functionally resemble exposure-driven prominence.

Research on bias in language models further supports this perspective by showing that outputs are shaped by distributional features such as frequency, salience, and social prominence (Bender et al., 2021; Gehman et al., 2020). As a result, models may preferentially reproduce dominant or widely circulated narratives, even when these are weakly supported, context-dependent, or contested.

The arguments in this section draw on heterogeneous sources, including empirical benchmarks, theoretical analyses, and analogies from cognitive psychology. These sources differ in epistemic status and should not be interpreted as providing uniform evidential support. Instead, they are used as converging indications that motivate the present framework rather than as direct validation of its components.

The present work builds on these insights by introducing the Narrative Dominance Gap (NDG) as a theoretical construct. NDG provides an interpretive lens for analyzing situations in which discursive prominence and evidential support diverge, rather than constituting a directly empirically established mechanism.

B.2.2 Retrieval augmentation and evidence grounding

Retrieval-augmented generation (RAG) and related approaches aim to improve model reliability by grounding outputs in external information sources (Lewis et al., 2020). By incorporating retrieved documents into the generation process, such systems reduce reliance on parametric memory and can improve factual accuracy in well-defined tasks.

However, retrieval alone does not resolve how heterogeneous information should be interpreted, compared, or weighted within a specific decision context. Even when relevant information is available, systems must still determine which sources are epistemically stronger, how different types of evidence should be evaluated, and how evidence should be aligned with the specific case at hand.

Insights from research on uncertainty and decision-making under incomplete information highlight the importance of distinguishing between different types of evidence and their respective reliability (Gigerenzer & Gaissmaier, 2011; Kahneman, 2011). Failure to differentiate between evidence sources can lead to overconfident but poorly grounded judgments.

Even when relevant documents are retrieved, the system must still determine how to distinguish between quantitative evidence, normative standards, contextual information, and discursive signals. Access to information does not by itself provide an evidence hierarchy, nor does it prevent the collapse of epistemically distinct layers into a simplified evaluative narrative.

Current RAG-based systems typically treat retrieved information as relatively homogeneous input, without explicitly representing differences between empirical measurements, normative criteria, expert analysis, experiential reports, and discursive signals. As a result, heterogeneous evidence may still be collapsed into a single narrative output, reproducing the same form of epistemic compression observed in non-retrieval settings.

Moreover, retrieval does not address the problem of underspecified queries. A system may access high-quality information, yet still apply it to an implicitly assumed problem formulation that does not match the user's actual context or intent.

The Context-Aware Epistemic Filtering (CEF) mechanism proposed in this work addresses these limitations by structuring how evidence is differentiated, aligned, and weighted relative to a reconstructed case context. Rather than treating information access as sufficient, CEF reframes the problem as one of epistemic organization: making explicit the type, strength, relevance, and conditional validity of different evidence sources.

Why retrieval is insufficient? Retrieval improves information access, but it does not by itself determine how retrieved material should be classified, weighted, or aligned with the case context. A system may retrieve both empirical test results and subjective commentary, yet still collapse them into a single evaluative narrative. The central problem is therefore not only access to information, but control over the epistemic organization of information. CEF addresses this by requiring explicit differentiation between empirical findings, normative thresholds, expert interpretation, experiential reports, and diffusion signals.

C.2.3 Ambiguity and conversational clarification

Research in conversational information retrieval and ambiguity-aware question answering has consistently shown that many user queries are underspecified and

require clarification (Zamani and Craswell, 2020; Aliannejadi et al., 2021). In these approaches, clarification is typically introduced as a mechanism for improving retrieval effectiveness or answer relevance.

More recent work on human–LLM interaction suggests a deeper role of interaction in shaping outcomes. In particular, interaction-centered evaluation demonstrates that user behavior is sensitive to contextual features of the interaction, including framing, prior exposure, and domain context (Zhou et al., 2025). Even when model outputs remain unchanged, variations in interactional conditions can lead to systematically different reliance decisions.

These findings indicate that ambiguity is not only a property of the input, but also of the interactional conditions under which the input is interpreted. Despite this, most current approaches treat clarification as an auxiliary conversational strategy rather than as a prerequisite for valid reasoning.

This problem is amplified by the fact that evaluative language often compresses multiple analytical dimensions into a single expression. Terms such as "reliable," "effective," or "good" may simultaneously refer to performance, controllability, compliance, durability, or suitability for a specific use case. Clarification is therefore necessary not only because the query is incomplete, but because evaluative terms themselves are analytically decomposable.

In technical domains, this limitation becomes critical, as small differences in context specification can fundamentally alter the validity of an evaluation. To address this gap, we introduce Adaptive Context Elicitation (ACE) as a core component of LLM-based technical decision support.

Unlike generic clarification strategies, ACE treats problem specification as an explicit and structured process aimed at reducing epistemic uncertainty before evaluation takes place. It is goal-directed, focusing on variables that materially affect outcomes; domain-sensitive, guided by expected parameter structures; and uncertainty-driven, prioritizing questions that maximally reduce ambiguity.

In this sense, ACE reframes clarification from a conversational convenience into a necessary step in constructing a valid and contextually grounded problem representation.

D.2.4 Explainability, trust, and human–AI collaboration

A growing body of research emphasizes that effective human–AI collaboration depends not only on model accuracy, but also on interaction design, trust calibration, and decision support (Doshi-Velez and Kim, 2017; Lee and See, 2004; Amershi et al., 2019). More recent work highlights that system effectiveness emerges from the structure of interaction, feedback mechanisms, and the distribution of roles between humans and AI systems (Vats et al., 2025).

Within this paradigm, trust is understood as a dynamic outcome of interaction rather than a static property of a model. Emerging evaluation frameworks therefore shift attention from output correctness to human reliance behavior, examining how users act upon model outputs in decision-making contexts. Research further suggests that reliance is influenced not only by correctness, but also by how uncertainty is communicated (Kim et al., 2024).

Interaction-centered studies provide evidence that reliance is sensitive to non-evidential cues such as framing, perceived competence, prior interaction history, and domain context (Zhou et al., 2025). These findings are complemented by work on linguistic calibration, which indicates that conversational systems can reduce overconfidence through calibrated verbal framing (Mielke et al., 2022). More broadly, decision-support research suggests that effective systems should not merely increase trust, but support appropriately calibrated reliance (Buçinca et al., 2021).

From the perspective of the present work, these effects can be interpreted as instances in which discourse-level signals influence user judgment independently of evidential strength. The Narrative Dominance Gap (NDG) provides a conceptual lens for analyzing such situations, capturing potential divergence between discourse prominence and evidential support.

In addition, research on anthropomorphism in dialogue systems suggests that conversational form itself can shape perceptions of competence and understanding (Abercrombie et al., 2023), further reinforcing the role of discourse-level features in human–AI interaction.

To address these challenges, the present framework integrates explainability, trust, and reasoning through three complementary components: ACE supports problem transparency, CEF supports evidence transparency, and NDG supports epistemic transparency. Together, these mechanisms aim to align user reliance with both the structure of the problem and the strength of the available evidence.

E.2.5 Positioning of the present work

Recent work on human–AI collaboration increasingly emphasizes interaction-centered evaluation, focusing on how system behavior shapes user decisions rather than solely on model performance (Vats et al., 2025; Zhou et al., 2025). Approaches such as REL-AI further demonstrate that human reliance is systematically influenced by interactional context.

The present work builds on this shift while contributing a complementary perspective. Rather than focusing primarily on behavioral measurement, it introduces a structured interactional–epistemic framework for technical evaluation.

Specifically, the framework addresses three interrelated dimensions that are not jointly structured in existing approaches:

- problem specification under ambiguity,
- epistemic differentiation of heterogeneous evidence,
- and the influence of discourse-level signals on evaluation.

Unlike standard question–answering systems, the framework treats the initial query as incomplete by default and requires explicit reconstruction before evaluation. Unlike retrieval-based approaches, it emphasizes evidence differentiation and contextual alignment rather than information access alone. Unlike traditional explainability methods, it structures reasoning prior to answer generation rather than post hoc. And unlike generic conversational systems, it defines clarification as decision-critical rather than optional.

Most importantly, the framework introduces Narrative Dominance Gap (NDG) as a unifying theoretical construct. While prior work has identified related phenomena—such as bias, overconfidence, and inappropriate reliance—it has not provided a general mechanism that connects them within a single interpretive structure.

NDG offers such a perspective by highlighting how discourse-level properties (e.g., frequency, familiarity, perceived authority) may diverge from evidential grounding. In this sense, interaction-centered findings on human reliance can be interpreted as empirical contexts in which such divergence becomes observable.

Together with Adaptive Context Elicitation (ACE) and Context-Aware Epistemic Filtering (CEF), NDG defines a reframing of LLM-based systems in technical domains: from answer generators to context-sensitive evaluative partners that support the co-construction of problem representations and evidence-based judgments.

III. A Unified Framework for Case-Based Technical Evaluation

A.3.1 From answer generation to case evaluation

The central argument of this paper is that LLM-based systems in technical domains should not be understood primarily as answer generators, but as **participants in case-based evaluation processes under uncertainty**.

In conventional interaction, the system receives a query and produces a response in a single step. This paradigm assumes that the problem is already well-specified. In practice, however, many technical queries are incomplete representations of underlying decision situations. As a result, direct answer generation often leads to premature evaluative closure.

We therefore define **case evaluation** as:

the structured reconstruction and assessment of a concrete problem instance through iterative context elicitation and evidence differentiation, enabling context-aware judgment rather than generalized answer production.

This reframing introduces three key implications:

1. **Problem specification becomes central**
   The system must actively participate in defining the problem before attempting to evaluate it.
2. **Interaction becomes necessary**
   Clarification is not optional but a core functional requirement.
3. **Evaluation replaces assertion**
   Outputs should reflect structured reasoning over heterogeneous evidence, rather than singular authoritative claims.

B.3.2 The role of Narrative Dominance Gap (NDG)

The transition from answer generation to case evaluation is motivated by a central epistemic risk: the tendency of LLMs to reproduce discursively dominant narratives.

We define **Narrative Dominance Gap (NDG)** as:

the divergence between the discursive prevalence of an evaluative claim and the strength of its empirical or normative support.

NDG is not a binary condition, but a graded diagnostic dimension. It becomes pronounced in situations where:

- claims are widely repeated or socially salient,
- but evidence is sparse, conflicting, or strongly context-dependent.

In such cases, models may generate outputs that are:

- rhetorically coherent,
- socially familiar,
- but analytically under-specified.

Importantly, NDG operates even when the model does not hallucinate. The failure mode is not fabrication, but **epistemic compression**—the collapse of heterogeneous evidence into a single narrative representation.

Within the proposed framework, NDG functions as a **trigger condition**:

- high NDG → increased need for clarification (ACE),
- high NDG → stricter evidence differentiation (CEF).

Operational boundary of NDG: A key challenge in operationalizing NDG is distinguishing discursive prevalence from reasonable prior probability. Not every widely held belief reflects narrative dominance; some priors are well grounded in aggregate evidence. NDG becomes diagnostically relevant when the prevalence of a claim is not proportional to its evidential support within the specific case under evaluation. In practical terms, NDG is most clearly present when a system defaults to category-level expectations despite the availability of case-specific contrary evidence. Thus, NDG should not be interpreted as a rejection of priors as such, but as a warning condition indicating that generic expectations may be overriding context-specific evidence. NDG does not refer to the presence of widely held beliefs per se, but to their inappropriate persistence in the presence of stronger, case-

specific counterevidence. In practice, this means that NDG is most visible in situations where case-specific evidence is available but not integrated into the evaluation.

C.3.3 Framework overview

The proposed framework integrates three components:

- **NDG (diagnostic layer)**
  Identifies potential divergence between discourse and evidence
- **ACE (interactional layer)**
  Reconstructs missing context through targeted elicitation
- **CEF (evaluative layer)**
  Structures and weights evidence relative to context

Together, these components define a structured interaction loop for technical evaluation.

Conceptual interaction flow

User Query (Q0)
↓
NDG / Ambiguity Signals
↓
ACE (context reconstruction)
↓
Context Representation (C*)
↓
CEF (evidence structuring)
↓
Structured Output (O)

This structure ensures that:

- evaluation does not proceed without sufficient context,
- evidence is not collapsed into a single narrative,
- and conclusions are explicitly conditioned on available information.

A further operational implication of NDG concerns the direction of information-seeking. ACE addresses underspecified user context through structured elicitation. However, a complementary mechanism is required when the evidential base itself is insufficient — that is, when case-specific empirical evidence is absent, sparse, or unverified. In such conditions, the system should explicitly flag the evidential gap rather than defaulting to category-level priors. This distinguishes two types of epistemic insufficiency that require different responses: underspecified user context, addressed by ACE through targeted clarification; and underspecified evidential base, addressed by explicit evidence-seeking or, where retrieval is unavailable, by surfacing uncertainty as a conditional qualifier on any evaluative output. High NDG therefore functions as a trigger not only for ACE, but also for active evidence-seeking prior to the application of CEF.

D.3.4 Adaptive Context Elicitation (ACE)

a) Objective

ACE aims to transform an underspecified query into a **well-defined case representation** by identifying and eliciting missing, decision-relevant variables.

b) Core principle

The system should assume that:

the initial user query is incomplete unless sufficient evidence suggests otherwise.

c) Types of missing variables

ACE focuses on variables that materially affect evaluation:

- **System parameters**
  (age, configuration, condition)
- **Operational context**
  (environment, usage intensity, constraints)
- **Lifecycle factors**
  (maintenance history, degradation, modifications)
- **User intent**
  (purchase decision, diagnosis, comparison)
- **Benchmark frame**
  (what the system is being compared to)

d) Example

Query:
"Is this system reliable?"

ACE identifies missing elements:

- reliable under what conditions?
- for what type of usage?
- compared to which standard or alternative?

e) Adaptive behavior

ACE is:

- **selective**, asking only relevant questions,
- **goal-directed**, focusing on decision-critical variables,
- **adaptive**, terminating when sufficient clarity is achieved.

Adaptive behavior ACE is selective, goal-directed, and uncertainty-driven. However, the implementation of the ACE framework must remain proportional to the complexity and criticality of the case to avoid unnecessary conversational friction in routine tasks.

While high-stakes technical evaluations necessitate rigorous context reconstruction, the system should prioritize directness for low-stakes or purely informational queries. Effective human–AI collaboration thus relies on a dynamic balance, ensuring that the depth of elicitation aligns with the epistemic weight of the underlying problem.

E.3.5 Context representation (C)*

The output of ACE is a structured context representation:

C* = {parameters, constraints, intent, comparison frame}

This representation is central to the framework because it:

- defines what is being evaluated,
- determines which evidence is relevant,
- and conditions how that evidence should be interpreted.

Without a well-defined C*, subsequent evaluation remains inherently unstable and potentially misleading.

F.3.6 Context-Aware Epistemic Filtering (CEF)

a) Objective

CEF organizes heterogeneous information into a **structured and transparent evaluative output**.

b) Core principle

Not all information is equally relevant, and not all relevant information is equally reliable.

c) Evidence dimensions

CEF distinguishes between multiple types of evidence:

1. **Empirical evidence**
   (measurements, test results)
2. **Normative criteria**
   (standards, regulatory requirements)
3. **Expert analysis**
   (engineering interpretation, technical assessments)
4. **Experiential reports**
   (user observations, field reports)
5. **Diffusion signals**
   (popularity, widespread usage, reputation)

Experiential evidence and evaluative overreach. Subjective observations may provide important signals about usability, controllability, or operator experience, but they are not epistemically equivalent to measured performance parameters. Experiential reports should therefore be treated as context-bound observations whose relevance depends on the task, benchmark, and degree of source independence. In technical evaluation, a common failure mode is to overgeneralize subjective impressions into system-level deficiencies. CEF prevents this by preserving experiential reports as a distinct evidential layer rather than allowing them to function as unmarked substitutes for empirical performance evidence.

These evidence categories are not epistemically equivalent. Quantitative and normative evidence will often carry greater evaluative weight in technical domains, while contextual evidence supports interpretation and experiential or discursive evidence may indicate perception, reputation, or recurring field observations without constituting primary proof. CEF therefore requires not only evidence classification, but explicit differentiation of evidential weight.

d) Key operations

CEF performs five key operations:

1. **Source identification**
   Classifying evidence according to its epistemic type
2. **Epistemic differentiation**
   Distinguishing between levels of reliability and evidential strength
3. **Context alignment**
   Filtering and prioritizing evidence based on relevance to C*
4. **Conflict detection**
   Preserving disagreement rather than forcing consensus
5. **Structured synthesis**
   Presenting evidence as layered rather than merged

Structural dependence of sources. Source independence should be assessed not only at the level of institutional origin, but also at the level of shared methodological, cultural, and contextual constraints. Two sources may come from different publications while sharing the same testing facility, time period, readership, evaluative norms, and discourse environment. In such cases, convergent judgments should not automatically be treated as independent corroboration. CEF therefore requires explicit assessment of structural dependence: the degree to which sources could realistically have arrived at different conclusions given their positions, constraints, and shared background conditions.

G.3.7 Benchmark alignment

Reliable evaluation requires explicit comparison frames. CEF introduces multiple benchmark layers:

- regulatory baseline (minimum compliance),
- historical or class-based benchmark,
- contemporary reference standard,
- and performance ceiling.

This prevents misleading conclusions resulting from implicit or inappropriate comparisons.

H.3.8 System-state and lifecycle differentiation

CEF explicitly distinguishes between:

- **design capability**, and
- **in-use condition**.

This prevents conflating:

- inherent design limitations,
- and degradation due to usage or maintenance.

Public evaluations of technical systems often arise from encounters with systems in degraded, modified, or poorly maintained states. For this reason, CEF distinguishes between design-level capability and typical in-use condition. The former concerns the intended or specification-level performance of a system, whereas the latter concerns the practical performance of instances encountered in operational environments.

Historical systems in present-day evaluation.

A further source of evaluative distortion arises when historically specified systems are assessed as if their present-day instances remained materially identical to their original tested condition. In practice, historical vehicles or technical systems are often encountered with replacement consumables, altered maintenance histories, updated materials, and changed operating environments. Reliable evaluation therefore requires explicit separation between design-level capability, historically tested configuration, and present-day in-use condition. Without this distinction, systems may be judged either too harshly or too favorably on the basis of an unstated and unstable comparison frame.

I.3.9 NDG-aware evaluation

CEF explicitly evaluates whether conclusions are:

- supported by structured evidence, or
- driven by discursive dominance.

When NDG is high:

- narrative claims are explicitly labeled as such,
- uncertainty is surfaced,
- and conclusions are presented as conditional.

J.3.10 Output structure

Final outputs are structured into:

1. **Context summary (C*)**
2. **Evidence layers (separated by type)**
3. **Trade-offs and constraints**
4. **Conditional conclusions**
5. **Uncertainty and NDG indicators**

Within this structure, outputs should distinguish between verified findings, plausible inferences, discursively prevalent claims, and unresolved uncertainties. This prevents rhetorically smooth conclusions from obscuring differences in evidential status.

This replaces single authoritative answers with **transparent evaluative reasoning**.

K.3.11 Integration

ACE and CEF are interdependent:

- ACE defines *what is being evaluated*,
- CEF defines *how it is evaluated*.

NDG connects both by:

- signaling when additional clarification is required,
- and constraining how conclusions should be formed.

Together, NDG, ACE, and CEF define a coherent **interactional–epistemic pipeline** for ambiguity-aware technical evaluation.

IV. Operationalization and Evaluation Paradigm

A.4.1 From conceptual framework to operational system

The proposed NDG–ACE–CEF framework is intentionally conceptual, but it is designed to support practical implementation in LLM-based systems.

Importantly, the framework does not require a fundamentally new model architecture. Instead, it can be operationalized through:

- structured prompting strategies,
- multi-step interaction design,
- modular reasoning pipelines,
- and retrieval-augmented systems with explicitly labeled evidence layers.

The key requirement is not architectural innovation, but **control over how reasoning is structured, sequenced, and exposed to the user**.

At a high level, an operational system implementing the framework would consist of the following stages:

1. Detection of ambiguity and NDG signals
2. Adaptive Context Elicitation (ACE)
3. Construction of a structured context representation (C*)
4. Evidence retrieval and classification
5. Context-Aware Epistemic Filtering (CEF)
6. Structured output generation

These stages can be implemented either as explicit system modules or as coordinated prompting strategies within a single model.

B.4.2 Ambiguity and NDG detection

a) Conceptual role

Ambiguity and NDG detection serve as the entry point of the framework. Their role is not to fully resolve uncertainty, but to determine whether **additional interaction and structured evaluation are required**.

b) Operational signals

In practice, ambiguity and NDG can be approximated using a set of heuristic or learned indicators:

- **Missing parameter signals**
  Absence of key domain variables required for evaluation
- **Generic evaluative language**
  Use of underspecified terms such as "good," "reliable," or "efficient"
- **Intent uncertainty**
  Lack of a clearly defined decision context (e.g., purchase vs. diagnosis)
- **Knowledge variance signals**
  High disagreement across available sources
- **Discursive dominance signals**
  High-frequency claims with low specificity or weak evidential grounding

c) Formal approximation

These signals can be combined into a scoring function:

$A(Q) = \Sigma w_i \cdot f_i(Q)$

Where:

- f_i represent ambiguity or NDG indicators
- w_i represent domain-dependent weights

The purpose of this function is not precise quantification, but **triggering ACE and CEF when ambiguity or NDG is sufficiently high**.

C.4.3 Adaptive Context Elicitation (ACE)

a) Objective

ACE transforms an underspecified query into a **well-defined case representation** by identifying and eliciting missing, decision-relevant variables.

b) Core principle

The system should assume that:

the initial user query is incomplete unless sufficient evidence suggests otherwise.

c) Question selection policy

Clarification can be modeled as an optimization problem:

Q* = argmax (Information Gain / Interaction Cost)

Where:

- Information Gain represents expected reduction in decision-relevant uncertainty
- Interaction Cost represents cognitive and interaction burden

d) Practical interpretation

In practice, this corresponds to:

- prioritizing variables with high decision impact,
- avoiding redundant or low-value questions,
- and stopping when marginal information gain becomes negligible.

e) Domain-specific templates

ACE relies on structured templates adapted to specific domains.

For example, in technical product evaluation (e.g., vehicles), relevant variables may include:

- production year,
- usage intensity or mileage,
- maintenance history,
- environmental conditions,
- intended use,
- and user intent (evaluation, purchase, diagnosis).

These templates guide the system in identifying which parameters are critical for evaluation.

f) Adaptive sequencing

ACE operates dynamically:

- it prioritizes high-impact variables first,
- adapts questioning based on user responses,
- and terminates when sufficient contextual clarity is achieved.

D.4.4 Context representation (C)*

The output of ACE is a structured context representation:

C* = {parameters, constraints, intent, comparison frame}

This representation is central to the framework because it:

- defines what is being evaluated,
- determines which evidence is relevant,
- and conditions how that evidence should be interpreted.

Without a well-defined C*, subsequent evaluation remains unstable and potentially misleading.

E.4.5 Context-Aware Epistemic Filtering (CEF)

This section translates the conceptual structure of CEF into operational terms.

a) Objective

CEF structures heterogeneous information into a context-sensitive evaluative output.

b) Evidence categories

CEF distinguishes between multiple types of evidence:

1. Empirical evidence (measurements, test results)
2. Normative criteria (standards, regulations)
3. Expert analysis (engineering interpretation)
4. Experiential reports (user observations)
5. Diffusion signals (popularity, widespread usage)

c) Key operations

In operational terms, CEF performs the following operations:

1. **Source identification**
   Classifying evidence according to its epistemic type
2. **Epistemic differentiation**
   Distinguishing between levels of reliability and evidential strength
3. **Context alignment**
   Filtering and prioritizing evidence based on relevance to C*
4. **Conflict handling**
   Preserving disagreement rather than forcing consensus
5. **Structured synthesis**
   Presenting evidence as layered rather than merged

Source independence must be evaluated not only at the level of institutional origin, but also at the level of shared methodological, cultural, and contextual constraints. Two sources may originate from different organizations while sharing the same testing facility, the same publication period, the same readership, and the same dominant discursive context. In such cases, treating convergent findings as fully independent confirmations inflates apparent evidential weight. CEF therefore requires explicit assessment of structural independence — the degree to which sources could plausibly have reached different conclusions given their respective positions and constraints. Where structural dependence is high, convergent findings should be treated as a single evidential signal rather than independent corroboration. This is particularly relevant in domains where publication ecosystems are shaped by shared commercial dependencies, common methodological standards, or dominant institutional narratives. This operational structure enables the evaluation process to be assessed not only in terms of outcomes, but in terms of how context is reconstructed, evidence is differentiated, and conclusions are formed under uncertainty.

F.4.6 Benchmark alignment

Reliable evaluation requires explicit comparison frames. CEF introduces multiple benchmark layers:

- regulatory baseline (minimum compliance),
- historical or class-based baseline,
- contemporary reference standard,
- and performance ceiling.

Failure to distinguish between these benchmark layers can produce systematically misleading judgments: a system may appear deficient against contemporary high-performance technologies while remaining adequate within its historical class or regulatory frame. This prevents misleading conclusions arising from inappropriate or implicit comparisons.

G.4.7 System-state and lifecycle differentiation

CEF explicitly distinguishes between:

- **design capability**, and
- **in-use condition**.

This prevents conflating:

- inherent design limitations,
- and degradation due to usage or maintenance.

H.4.8 NDG-aware evaluation

CEF incorporates NDG by explicitly evaluating whether conclusions are:

- supported by structured evidence, or
- driven by discursive dominance.

When NDG is high:

- narrative claims are explicitly labeled,
- uncertainty is surfaced,
- and conclusions are presented as conditional.

I.4.9 Output schema

A minimal structured output includes:

1. Context summary (C*)
2. Evidence layers (separated by type)
3. Trade-offs and constraints
4. Conditional conclusions
5. Uncertainty and NDG indicators

This schema replaces single authoritative answers with **transparent evaluative reasoning**.

J.4.10 Integration

ACE and CEF are interdependent components:

- ACE defines *what is being evaluated*,
- CEF defines *how it is evaluated*.

NDG connects both by:

- signaling when additional clarification is needed,
- and constraining how conclusions should be formed.

Together, NDG, ACE, and CEF define a coherent interactional–epistemic pipeline for ambiguity-aware technical evaluation.

V. Evaluation Paradigm

A.5.1 Why traditional evaluation is insufficient

Unlike standard evaluation approaches, the proposed framework evaluates not only the correctness of outputs, but the quality of the evaluation process itself. Standard evaluation of large language models (LLMs) has primarily focused on metrics such as accuracy, factual correctness, and benchmark performance.

In such domains, correctness is not an intrinsic property of an answer but is contingent on the underlying problem specification. If the problem itself is underspecified, even a factually correct response may be irrelevant or misleading. Moreover, many technical queries do not admit a single correct answer but instead require conditional judgment based on multiple interacting variables.

As a result, evaluating LLM systems solely on answer correctness overlooks a critical dimension: whether the system has appropriately reconstructed the problem and supported a valid evaluative process.

This work therefore adopts an **interaction-centered evaluation paradigm**, where system performance is assessed in terms of its ability to support structured problem formulation, evidence-based reasoning, and calibrated user judgment.

This aligns with emerging reliance-based evaluation frameworks that measure how users depend on AI systems in decision contexts (Zhang et al., 2025).

B.5.2 Research hypotheses

To operationalize this paradigm, we define a set of testable hypotheses.

H1 — Problem Specification Quality

Systems implementing Adaptive Context Elicitation (ACE) will produce more complete and better-specified representations of the problem compared to direct-answer baselines.

H2 — Decision Quality

Users interacting with NDG–ACE–CEF systems will make higher-quality decisions in technical tasks compared to users receiving standard LLM responses.

H3 — Trust Calibration

Structured outputs produced by the framework will improve trust calibration, reducing both over-reliance and under-reliance, and aligning user confidence more closely with actual system reliability.

H4 — Transparency and Understanding

Users will demonstrate improved understanding of evidence structure, trade-offs, and uncertainty when interacting with Context-Aware Epistemic Filtering (CEF) outputs.

H5 — Cognitive Efficiency (conditional)

Although ACE introduces additional interaction steps, it will reduce downstream cognitive load in complex decision-making tasks by improving problem specification and reducing ambiguity.

C.5.3 Metrics

These metrics are designed not only to evaluate output correctness, but to capture the quality and structure of the evaluation process itself.

We define four evaluation dimensions:

a) 1. Problem Specification Quality

- number of relevant variables identified
- completeness vs expert baseline
- ambiguity reduction score

b) 2. Decision Quality

- agreement with expert judgment
- task outcome quality
- decision error rate

c) 3. Trust Calibration

- calibration gap
- over-reliance index
- under-reliance index

These measures capture whether users trust the system appropriately rather than blindly or insufficiently.

d) 4. Cognitive Load & Usability

- NASA-TLX
- number of interaction steps
- perceived usefulness

e) 5. Transparency

- ability to identify evidence sources
- understanding of trade-offs
- explanation satisfaction

f) 6. Process Integrity Score (PIS)

Measures the extent to which the system follows a structured and complete evaluation process prior to producing a conclusion, including context reconstruction, differentiation of evidence types, and explicit handling of uncertainty.

D.5.4 Baseline conditions

To evaluate the proposed framework, we define three baseline conditions:

Baseline 1 — Direct Answer LLM

Single-turn response with no clarification or structured reasoning.

Baseline 2 — Generic Clarification

Systems that ask follow-up questions, but without a structured policy such as ACE.

Baseline 3 — Retrieval-Augmented Generation (RAG)

Systems that incorporate external information sources but do not apply structured epistemic filtering (no CEF).

Experimental Condition — NDG + ACE + CEF

Systems implementing the full framework:

- structured clarification,
- context reconstruction,
- and multi-layer evidence evaluation.

E.5.5 Experimental design

We propose a **within-subjects experimental design**, where participants complete tasks under multiple system conditions.

Key features:

- same participants across conditions
- counterbalanced order
- tasks with real ambiguity

a) Task types

- product evaluation
- diagnostics
- risk assessment

b) Data collection

- interaction logs
- decisions
- confidence
- questionnaires

F.5.6 Key shift

The core shift is:

✗ "Is the answer correct?"

✓ "Did the system help construct a correct evaluation?"

This shift reflects the reality that, in ambiguous technical domains, the quality of reasoning and interaction is as important as the correctness of any single output.

VI.6. Discussion

A.6.1 Reinterpreting LLM failure modes in technical domains

The framework proposed in this paper suggests a shift in how LLM limitations are understood in technical contexts. While much of the literature focuses on hallucination and factual inaccuracy, our analysis indicates that a more pervasive failure mode is **premature evaluative closure under ambiguity**.

Beyond training data effects, LLM behavior is also shaped by optimization toward coherent and complete responses under partial context. This can lead to early closure: the formation of a plausible and internally consistent answer before all available evidence has been fully differentiated and evaluated. As a result, less salient but epistemically stronger signals may be underweighted relative to more dominant or readily integrable narratives. This pattern resembles well-documented human cognitive heuristics, not because the model "thinks like a human," but because both systems operate under constraints that favor coherence, compression, and timely decision-making.

LLMs do not simply inherit human biases; they reproduce a structurally similar pattern of early coherence formation under uncertainty, which can lead to premature closure and underweighting of critical evidence.

From a broader perspective, the observed behavior can be interpreted as a form of structural convergence between LLM systems and human reasoning under uncertainty. This similarity should not be understood as evidence that models "think like humans," but rather as a consequence of shared constraints: operating under incomplete context, heterogeneous information, and pressure to produce a coherent output within limited time. Under such conditions, both human reasoning and model generation tend toward early formation of plausible interpretations, even when the available evidence has not been fully differentiated. In LLM systems, this manifests as premature evaluative closure, whereby discursively dominant or easily integrable narratives may be preferred over less salient but epistemically stronger signals. This perspective complements the concept of Narrative Dominance Gap (NDG) by linking it not only to properties of the data, but also to more general constraints in the process of evaluation under uncertainty. In this sense, NDG can be understood not only as a property of information environments, but as an emergent effect of how both discourse and reasoning processes interact under conditions of uncertainty.

In many cases, the model does not generate false information. Instead, it:

- selects an implicit interpretation of an underspecified query,
- collapses heterogeneous evidence into a single narrative,
- and presents the result with unwarranted coherence.

This behavior is consistent with the concept of **Narrative Dominance Gap (NDG)**. The model reflects patterns of discourse rather than explicitly structured evidence. As a result, outputs may appear plausible and informative while remaining insufficiently grounded for decision-making.

From this perspective, improving LLM reliability is not only a matter of better knowledge or retrieval, but of **restructuring the interaction and reasoning process**.

B.6.2 Clarification as a core function, not an auxiliary feature

A central implication of this work is that clarification should be treated as a **core functional requirement** rather than an optional conversational feature.

In current systems, clarification is often:

- minimal,
- generic,
- or absent.

However, in technical evaluation tasks, the absence of clarification leads directly to mis-specified problem representations. By introducing **Adaptive Context Elicitation (ACE)**, the framework positions clarification as a structured mechanism for reconstructing decision-relevant context.

This has important implications for system design:

- interaction becomes part of reasoning,
- users become participants in problem specification,
- and system outputs become conditional on explicitly surfaced assumptions.

C.6.3 From information retrieval to epistemic structuring

Another implication concerns the role of retrieval and information access. While retrieval-augmented systems improve access to external knowledge, they do not by themselves ensure reliable evaluation.

The key challenge is not only retrieving information, but **structuring it epistemically**.

Through **Context-Aware Epistemic Filtering (CEF)**, the framework emphasizes:

- differentiation between types of evidence,
- alignment with case-specific context,
- and explicit representation of disagreement and uncertainty.

This shifts the focus from "more information" to **better-structured information**.

D.6.4 NDG as a unifying concept

The introduction of NDG provides a unifying perspective across several known issues:

- hallucination → extreme NDG (no evidential support)
- overconfidence → unacknowledged NDG
- bias toward common narratives → discursive dominance
- miscalibration → mismatch between discourse and evidence

NDG therefore functions as:

- a diagnostic tool,
- a design constraint,
- and a bridge between epistemic and interactional perspectives.

E.6.5 Implications for human–AI collaboration

The framework contributes to human–AI collaboration in three key ways:

1. **Improved problem formulation**
   Users are supported in articulating relevant variables.
2. **Improved decision support**
   Outputs reflect structured evaluation rather than simplified narratives.
3. **Improved trust calibration**
   Users are exposed to assumptions, evidence layers, and uncertainty.

This aligns with broader goals in human-centred AI, where system value is measured not only by correctness, but by how effectively it supports human reasoning.

From a broader systems perspective, the role of LLMs can be understood as a key component in shaping what has been described as "architectural capital" in human–computational systems (Rusev, 2026). Within this framework, system performance is not determined solely by the capabilities of individual components, but by the way interaction structures organize the flow of information, attention, and decision-making. When embedded in structured interactional frameworks such as NDG–ACE–CEF, LLMs function not merely as information-processing modules, but as configurational elements that shape the integration of human and machine capabilities and enable their augmentation through partial compensation of their respective limitations. In this sense, improving reliability is not only a matter of enhancing the model itself, but of optimizing the architecture of interaction through which the evaluative process is constructed.

VII. Contributions (Explicit Summary)

This paper makes the following contributions:

A.7.1. Conceptual Contribution

It introduces **Narrative Dominance Gap (NDG)** as a theoretical construct for analyzing divergence between discursive prevalence and evidential support in LLM-based evaluation.

B.7.2. Framework Contribution

It proposes a unified **NDG–ACE–CEF framework** that integrates:

- ambiguity detection,
- context reconstruction,
- and structured evidence evaluation.

C.7.3. Methodological Contribution

It defines **case evaluation** as an alternative paradigm to answer generation and provides a structured interactional–epistemic model for implementing it.

D.7.4. Evaluation Contribution

It outlines an **interaction-centered evaluation paradigm**, including hypotheses, metrics, and experimental design focused on:

- problem specification quality,
- decision quality,
- trust calibration,
- and transparency.

E.7.5. Design Contribution

It provides actionable design principles for LLM systems in technical domains:

- treat queries as incomplete by default,
- structure clarification,
- differentiate evidence explicitly,
- expose uncertainty and disagreement.

***F.**7.6 From Answer Generation to Interactional Reasoning*

This paper argued that improving LLM performance in technical domains requires a shift from answer generation to case evaluation under uncertainty. By introducing Narrative Dominance Gap (NDG), Adaptive Context Elicitation (ACE), and Context-Aware Epistemic Filtering (CEF), it proposed a unified framework for ambiguity resolution, context reconstruction, and structured evidence evaluation.

The central claim is that reliable AI systems in technical domains must not only generate answers, but also reconstruct the problem, differentiate evidence, and support calibrated human judgment. Beyond these specific mechanisms, the broader implication is that limitations of LLM-based evaluation arise not only from gaps in knowledge, but from structural pressures toward early coherence under uncertainty. In this respect, model behavior may converge with human reasoning—not because of shared cognition, but because both operate under constraints of incomplete, heterogeneous information and pressure for coherent output.

The proposed NDG–ACE–CEF framework addresses this challenge by restructuring interaction so that context reconstruction and evidence differentiation precede evaluative closure. Ultimately, the next generation of human–AI systems will be defined not by fluency alone, but by epistemic discipline and interactional intelligence: systems that ask before answering and structure before

concluding, thereby supporting more reliable judgment under real-world uncertainty.

VIII. Limitations

This work has several limitations.

First, the framework is conceptual and not yet implemented as a full system. Its effectiveness depends on how ACE and CEF are instantiated in practice.

Second, the approach assumes that users are willing and able to provide additional context. In real-world settings, user input may be incomplete or inaccurate.

Third, while CEF improves evidence structuring, it does not eliminate biases in underlying data sources.

Fourth, the framework introduces additional interaction steps, which may increase perceived friction if not carefully managed.

Finally, the framework is developed primarily for technical domains with measurable parameters and may require adaptation for more subjective domains.

Conceptual scope of NDG.

NDG is introduced as a diagnostic and interpretive construct, not as a fully validated causal mechanism. Its role is to identify conditions under which discursive prominence may become misaligned with evidential support in case-based evaluation. Future work should therefore focus on operational criteria, annotation procedures, and empirical benchmarks capable of distinguishing NDG from related phenomena such as generic prior dependence, overconfidence, or source bias.

The framework is developed with particular attention to domains characterized by measurable parameters, established normative standards, heterogeneous source reliability, and verifiable empirical benchmarks. These structural properties — present in engineering evaluation, technical diagnostics, regulatory compliance assessment, and related fields — are the conditions under which NDG, ACE, and CEF are most directly applicable. The framework's applicability to less structured domains, such as aesthetic, legal, ethical, or purely subjective evaluation, is not excluded in principle, but requires domain-specific adaptation that lies beyond the scope of the present work. This bounded scope is not a limitation of the framework's ambition, but a precondition for its empirical testability and practical implementability.

IX. Conclusion

This paper argued that improving LLM performance in technical domains requires a shift from answer generation to **case evaluation under uncertainty**.

By introducing:

- **Narrative Dominance Gap (NDG)**,
- **Adaptive Context Elicitation (ACE)**,
- and **Context-Aware Epistemic Filtering (CEF)**,

we proposed a unified framework that integrates:

- ambiguity resolution,
- context reconstruction,
- and structured evidence evaluation.

The central claim is that reliable AI systems in technical domains must not only generate answers, but also:

- reconstruct the problem,
- differentiate evidence,
- and support calibrated human judgment.

Beyond specific mechanisms, this work advances a broader interpretation: limitations of LLM-based evaluation in technical domains arise not only from gaps in knowledge, but from structural pressures toward early coherence under uncertainty. In this respect, model behavior may exhibit forms of convergence with human reasoning—not due to shared cognition, but due to shared constraints in processing incomplete, heterogeneous information under time and coherence pressures.

The proposed NDG–ACE–CEF framework addresses this challenge not by eliminating these pressures, but by restructuring interaction so that context reconstruction and evidence differentiation precede evaluative closure.

More broadly, improving LLM reliability requires not only better access to information, but stronger control over how evaluative reasoning is structured, sequenced, and made explicit in interaction.

Ultimately, the next generation of human–AI systems will be defined not by increased fluency, but by epistemic discipline and interactional intelligence—systems that ask before answering and structure before concluding, thereby supporting more reliable decision-making under real-world uncertainty. In this sense, the primary limitation of current LLM systems is not that they lack knowledge, but that they tend to draw conclusions too early under conditions of uncertainty — a pattern that, to some extent, is also characteristic of human reasoning. Addressing this limitation does not necessarily require more data, greater computational power, additional training programs for users, or more complex models, but rather a reorganization of how existing capabilities are structured and applied within interaction. The NDG–ACE–CEF framework demonstrates how targeted adjustments in the structure of interaction and reasoning can improve system reliability without fundamental changes to the underlying model. This perspective aligns with the broader view that the effectiveness of technological systems depends not only on the properties of individual components, but on how they are integrated. Future research should therefore extend beyond the immediate scope of this work and include interdisciplinary efforts — spanning engineering, cognitive, and behavioral domains — aimed at improving patterns of reasoning as a whole, both in human–AI interaction and in

the way models themselves structure the reasoning process, as well as the overall architecture of human–AI systems.

Appendix: **A Documented NDG Instance — From Narrative to Evidence-Based Evaluation**

The appendix provides a fully documented real-world case demonstrating the NDG–ACE–CEF framework in action and is included to support the methodological contribution.

This appendix is included as an illustrative and methodological component, not as a standalone empirical validation.

A.1. Initial Query

The following case begins with an intentionally underspecified evaluative query:

"*What do you think about the effectiveness of the braking system of the Soviet car Lada 1500 (VAZ-2103)?"*

The query does not specify:

- temporal frame (historical vs present-day usage),
- operational context (single-stop vs repeated braking),
- system state (original vs maintained/modified),
- comparison benchmark.

As such, it represents an incomplete problem specification rather than a well-defined evaluation task.

A.2. Baseline LLM Outputs (NDG Manifestation)

Initial responses generated by large language models contained claims such as:

- "tendency to overheat (fade)"
- "weak braking performance"
- "poor component quality"
- "unsuitable for modern driving conditions"

These claims exhibit:

- high discursive prevalence (common associations with older vehicles and drum brakes),
- lack of case-specific empirical support,
- presentation as definitive conclusions rather than conditional hypotheses.

This constitutes a clear instance of **Narrative Dominance Gap (NDG)**, where discursively dominant claims are produced without sufficient evidential grounding.

A.3. ACE in Action — Context Reconstruction

Context was progressively reconstructed through targeted input (Adaptive Context Elicitation), including:

1. Historical road tests (Motor, Autocar, 1976)
2. Engineering analysis aligned with UN Regulation No. 13-H
3. Clarification of realistic present-day usage conditions:
   - modern tires (post-1980 / contemporary compounds),
   - replacement of asbestos brake pads,
   - use of modern brake fluid (e.g., DOT-4),
   - realistic system condition (maintained, not factory-new).

This yields a structured context representation:

C* = { system: VAZ-2103 braking system (disc front / drum rear),
temporal frame: historical design vs present-day usage,
operational context: single-stop and repeated braking,
comparison frame: regulatory baseline and contemporary vehicles,
system state: maintained with modern consumables }

A.4. CEF — Evidence Differentiation

Evidence is structured into distinct epistemic layers:

Evidence Type	Content	Reliability
Empirical measurements	0.4 g @ 20 lb; 1.0 g @ 50 lb; ≤38 m @ 80 km/h; ≈60 m @ 100 km/h	High
Thermal behavior	Fade test (10 stops from 70 mph; moderate pedal increase, no collapse)	High
Normative criteria	Compliance with UN Regulation No. 13-H	High
Engineering analysis	Improved performance with higher tire friction (μ increase)	High
Experiential reports	Subjective descriptions of pedal feel ("non-progressive", "over-servoed")	Medium
Diffusion signals	"Soviet cars have weak brakes"	Low

A.5. NDG Identification

The claim:

"The braking system is prone to fade"

is identified as a **high-NDG statement**, because:

- discursive prevalence: high (general knowledge about drum brakes),

- case-specific evidence: contradictory (fade test passed successfully),
- evidential grounding: absent in the initial response.

Similarly, claims regarding "weak braking performance" and "poor component quality" were not supported by case-specific empirical evidence.

A.5.1. NDG Example: Pedal Feel

The evaluation of pedal feel provides a distinct example of Narrative Dominance Gap (NDG) arising not from incorrect data, but from the misclassification and over-weighting of subjective evidence.

Historical road tests (Motor, Autocar, 1976) describe the braking behavior as "non-progressive" and "over-servoed," indicating difficulty in smooth modulation. These observations constitute **experiential reports**, reflecting driver perception under specific testing conditions.

However, such statements:

- are inherently **context-dependent** (driver expectations, comparison vehicles, test conditions),
- are not supported by **quantitative metrics**,
- may be influenced by **shared testing environments and editorial framing**.

Despite this, they are often implicitly treated as **system-level deficiencies**, rather than as **subjective observations of limited generalizability**.

This creates an NDG condition:

Dimension	Pedal Feel Case
Discursive prevalence	Moderate to high ("non-progressive braking")
Evidence type	Experiential (subjective)
Case-specific empirical support	Absent
Interpretation risk	Over-generalization into system deficiency

Under Context-Aware Epistemic Filtering (CEF), such evidence should be:

- explicitly classified as **subjective**,
- assigned **moderate evidential weight**,
- interpreted as **context-bound**, not universal system behavior.

A.6. Structural Dependence of Sources

Although two independent publications (Motor and Autocar, 1976) report similar subjective impressions (e.g., pedal feel), their independence is limited by:

- shared testing environment (MIRA),
- identical temporal context,
- overlapping readership and editorial culture,
- exposure to similar discourse conditions.

Therefore, their convergence does not constitute fully independent validation, but rather structurally dependent agreement.

A.7. Revised Evaluation under CEF

After context reconstruction and evidence differentiation:

- **Maximum braking capability**: objectively sufficient (≈6.5 m/s^2), exceeding regulatory minimums.
- **Thermal performance**: stable under repeated braking; no critical fade observed.
- **System architecture**: appropriate for the era and comparable to contemporaries.
- **Pedal feel**: subjective reports of non-linear or over-assisted behavior (medium reliability).
- **Modern usage scenario**: performance is significantly improved with contemporary tires and consumables.

Notably, subjective reports of pedal feel should be interpreted with caution. While historical test sources describe non-progressive behavior, such assessments are inherently context-dependent and may reflect specific testing conditions, driver expectations, or comparative frames rather than universally reproducible system characteristics.

A.8. Key Observation

The primary failure mode observed in initial responses is not factual fabrication, but:

evaluative closure based on category-level priors, arising under conditions of incomplete context and pressure for coherent output, structurally analogous to well-documented patterns in human reasoning under uncertainty

The NDG–ACE–CEF framework transforms:

- a narrative-driven evaluation
  into
- a structured, context-conditioned assessment.

The primary failure mode observed in initial responses is not factual fabrication, but evaluative closure based on category-level priors, arising under conditions of incomplete context and pressure for coherent output,

structurally analogous to well-documented patterns in human reasoning under uncertainty. This contrast illustrates that the primary change is not the acquisition of new knowledge, but the restructuring of evaluation through context reconstruction and evidence differentiation. In this sense, the case does not demonstrate improved factual recall, but improved epistemic organization.

This demonstrates that the improvement arises from restructuring the evaluation process, not from introducing new information.

A.9. Implication for LLM Evaluation

This case demonstrates that:

- access to information alone is insufficient,
- reliable evaluation requires:
    - context reconstruction,
    - explicit differentiation of evidence,
    - controlled handling of discursive priors.

Therefore, system performance should be evaluated not only in terms of answer correctness, but in terms of:

its ability to construct a valid evaluation process under uncertainty.

***

ACKNOWLEDGMENT

This study is part of an independent research program: *Ideas for the Third Millennium*, developing an architectural theory of complex systems.

The program centers on the concept of Architectural Capital and extends it across engineering, economic, and cognitive domains, including product design, system evaluation, and human–machine interaction.

X. REFERENCES

[1] Abercrombie, Gavin, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. (2023). Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790.

[2] Autocar. (1976, August 14). *Lada 1500 road test*. IPC Business Press Ltd.

[3] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., & Moffat, A. (2021). Conversation-based information retrieval: An overview of recent developments and challenges. *Information Retrieval Journal*, 24, 1–39.

[5] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445717

[6] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

[7] Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

[8] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[9] Geng, S., Liu, H., Zhang, Z., & Liu, Y. (2023). Are language models calibrated? *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes, and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[11] Kim, Sunnie S. Y., Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. (2024). "I'm not sure, but…": Examining the impact of large language models 'uncertainty expression on user reliance and trust. *In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

[12] Mielke, Sabrina J., Arthur Szlam, Emily Dinan, and Y-Lan Boureau. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics,* 10:857–872.

[13] Motor. (1976, № 32). *Road test: Lada 1500*. IPC Business Press Ltd.

[14] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

[15] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[16] Liu, A., et al. (2024). CLAMBER: A benchmark for ambiguity resolution in language models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[17] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.

[18] Rajpurkar, P., Min, S., Lyu, X., & Liang, P. (2020). AmbigQA: Answering ambiguous open-domain questions.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[19] Rusev, M. R. (2026). Architectural Capital: A Modern Framework for Addressing Consumer Dissatisfaction and Utility Stagnation. *Economic Thought Journal*, 71 (1), 94–125. https://doi.org/10.56497/etj2671105
[20] Rusev, M. (2026). Analysis of the compliance between historical braking performance parameters and contemporary UN R13-H regulatory requirements. Engineering Sciences, № (1). https://doi.org/10.7546/EngSci.LXIII.26.01.06
[21] Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., Titterton, V., Malreddy, S. V., Aggarwal, R., Xu, Y., Ding, L., Mehta, J., Grinnell, N., Liu, L., Zhong, S., Gandamani, D. N., Tang, X., Ghosalkar, R., Shen, C., Shen, R., Hussain, N., Ravichandran, K., & Davis, J. (2025). A survey on human-AI collaboration with large foundation models. ACM. https://arxiv.org/pdf/2403.04931v3
[22] Wang, X., et al. (2024). Human-AI collaboration: A survey of design, evaluation, and open challenges. *ACM Computing Surveys*.
[23] Zamani, H., & Craswell, N. (2020). Generating clarifying questions for information retrieval. *Proceedings of the Web Conference (WWW 2020)*.
[24] Zhang, Y., et al. (2025). REL-AI: A framework for reliance-based evaluation of AI systems. *Proceedings of NAACL 2025*.
[25] Zhou, K., Hwang, J. D., Ren, X., Dziri, N., Jurafsky, D., & Sap, M. (2025). REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 11148–11167). Association for Computational Linguistics.
[26] Zhu, W., Liu, H., Dong, Y., Yu, J., Cheng, J., & Li, J. (2025). Rethinking hallucination in large vision-language models: A task-specific perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 9811–9832). Association for Computational Linguistics.