Correct by Design or Correct by Capability?: A Study of Formal and Empirical Trust in NLU Paradigms

Jiaxuan Guo
Graduate School of Engineering, The University of Tokyo
jxguo@g.ecc.u-tokyo.ac.jp

Abstract

This report investigates the fundamental trade-offs between traditional, symbolic "glass-box" pipelines and modern, end-to-end "black-box" Large Language Models (LLMs) for Natural Language Understanding. Through two targeted case studies—a lexical ambiguity task and a high-stakes clinical trial audit—we qualitatively analyze their differing reasoning processes. Our findings reveal that while frontier LLMs demonstrate surprising proficiency across both commonsense and complex procedural tasks, the nature of their correctness is distinct from that of symbolic systems. We argue that the central trade-off is not merely one of performance, but between the Formal Trust derived from a verifiable, deterministic design, and the Empirical Trust placed in a powerful but opaque capability. This distinction has profound implications for the responsible deployment of AI in safety-critical domains, highlighting a need for systems that balance capability with verifiability.

1 Introduction

The rise of powerful "black-box" Large Language Models (LLMs) (Vaswani et al., 2017; Devlin et al., 2018) has created a tension with traditional, transparent "glass-box" pipelines from Computational Linguistics. This report moves beyond simple performance benchmarks to investigate how their reasoning processes differ, especially in high-stakes scenarios. We employ two targeted case studies—one on commonsense ambiguity, the other on a simulated clinical trial—to probe this fundamental distinction.

Our analysis reveals that the core trade-off is not merely performance, but the very nature of correctness. We argue this is a choice between the Formal Trust derived from a symbolic system's verifiable design and the Empirical Trust placed in a neural model's impressive but opaque capabilities. The report concludes by discussing the profound implications of this distinction for safety-critical domains where verifiability and accountability are paramount.

2 Two Paradigms in NLU

2.1 Paradigm A: The Modular Pipeline

The modular approach, conceptualized as a "glass-box" pipeline, treats language understanding as a sequence of discrete, interpretable stages. Each stage enriches the analysis and provides input for the next, forming a comprehensive, hierarchical process. While specific implementations vary, a granular pipeline often includes the following stages:

1. Lexical and Morphological Analysis. The process begins with foundational text processing. **Tokenization** segments the raw text into words and punctuation. Subsequently, **Morphological Analysis** or **Lemmatization** reduces words to their base dictionary form (e.g., "running" → "run"), normalizing the input for downstream tasks.

2. Syntactic Parsing. This stage analyzes the grammatical structure of the input sentence. Two dominant paradigms exist: *Constituency Parsing* and *Dependency Parsing*. Constituency parsing, based on formalisms like Probabilistic Context-Free Grammar (PCFG), recursively decomposes a sentence into nested phrases (e.g., noun phrases, verb phrases). Dependency parsing, more prevalent in modern NLP pipelines for its efficiency and direct representation of word relationships, constructs a directed tree where each word is connected to its syntactic head. Pioneering work by Chen and Manning (2014) introduced fast and accurate neural dependency parsers, forming the backbone of tools such as spaCy.

3. Semantic Analysis and Coreference Resolution. Building on the syntactic backbone, this stage injects semantic information. Key tasks include **Named Entity Recognition (NER)**, which identifies real-world entities (e.g., persons, organizations), and **Semantic Role Labeling (SRL)**, which seeks to answer “who did what to whom.” SRL is often framed in terms of resources like FrameNet (Baker et al., 1998), where roles such as *Agent*, *Patient*, or *Instrument* are assigned to constituents based on verb-centered semantic frames. In parallel, **Coreference Resolution** is performed to link expressions referring to the same entity (e.g., resolving that "he" and "Tim Cook" refer to the same person).

4. Discourse Analysis. Moving beyond the sentence level, **Discourse Analysis** examines how sentences connect to form a coherent text. Using frameworks like Rhetorical Structure Theory (RST), it identifies relations between textual units, such as *Contrast*, *Cause*, or *Elaboration*, revealing the author’s larger argumentative structure.

5. Inference. The final stage uses the fully enriched structural and semantic representations for high-level reasoning. This may involve dedicated models for **Natural Language Inference (NLI)**, where the goal is to determine whether a hypothesis can be inferred, contradicted, or is neutral with respect to a given premise, often leveraging the structured outputs from all previous stages for a more robust judgment.

The primary strength of this paradigm is its interpretability; if a failure occurs, it can often be traced back to a specific faulty module. However, it is prone to error propagation, where an error in an early stage cripples all subsequent steps.

2.2 Paradigm B: The End-to-End LLM

In contrast, the end-to-end paradigm uses a single, massive neural network to map raw text directly to the desired output. An LLM implicitly performs parsing, semantic analysis, and reasoning within its hidden layers, leveraging vast amounts of world knowledge learned during pre-training.

Its main advantage is high performance and flexibility, especially on tasks requiring contextual understanding and common-sense reasoning. The primary disadvantage is its lack of interpretability; the model acts as a "black box," making it difficult to diagnose or correct its reasoning process when it fails.

3 Case 1: Lexical Ambiguity Resolution

To empirically demonstrate the architectural tradeoffs between the two paradigms, we designed a case study centered on lexical ambiguity. The task requires the system to correctly interpret a sentence containing a polysemous word, "Amazon," which can refer to either a company or a geographical location.

3.1 Task and Methodology

The input sentence for both systems is:

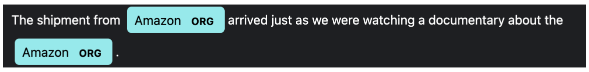
> *“The shipment from Amazon arrived just as we were watching a documentary about the Amazon.”*

The objective is to correctly determine the number of geographical locations mentioned. The modular pipeline’s behavior is simulated based on its component characteristics (using spaCy for NER), while the end-to-end model used is ChatGPT (GPT-o3), prompted to provide a step-by-step analysis.

3.2 Results and Analysis

The two paradigms yielded starkly contrasting results, highlighting their fundamental differences.

The "Glass-Box" Pipeline’s Failure. The modular pipeline failed to correctly parse the sentence. As shown in Figure 1, its Named Entity Recognition (NER) component, relying on statistical patterns and local context, misclassifies both instances of "Amazon" as an ‘ORG‘ (Organization). This failure, while definitive, is highly interpretable; we can pinpoint the exact module and the reason for the error—a lack of real-world, commonsense knowledge to disambiguate the word based on the broader sentence context. This is the hallmark of a ’glass-box’ system: its failures are transparent.



The shipment from Amazon ORG arrived just as we were watching a documentary about the Amazon ORG .

Figure 1: NER output from the modular pipeline (simulated with spaCy). Both instances of "Amazon" are incorrectly classified as ORG.

The "Black-Box" LLM’s Success. Conversely, the end-to-end model, ChatGPT (GPT-o3), successfully navigated the ambiguity. When prompted with a directive to reason step-by-step, the model produced a clear, logical analysis, summarized in Table 1. It correctly used the local contexts—“shipment from...” versus “documentary about...”—to disambiguate the entities and arrive at the correct answer.

Table 1: Step-by-step disambiguation by ChatGPT (GPT-3.5).

Instance	Contextual Clue	Inferred Entity Type
1. Amazon	“shipment from...”	Organization (Company)
2. the Amazon	“documentary about...”	Location (River/Region)

Final Conclusion: Total geographical locations is **1**.

Discussion. The juxtaposition of these two results is telling. The pipeline offers verifiable transparency in failure, while the LLM provides high performance through opaque reasoning. The pipeline's correctness is contingent on the deterministic behavior of its parts; the LLM's correctness stems from its vast, probabilistic understanding of the world. This case study exemplifies the central dilemma in modern NLU: the choice between a system that is explainable but potentially brittle, and one that is powerful but fundamentally a 'black box.'

3.3 Case 2: High-Stakes Reasoning and Protocol Adherence

Task Setup. To escalate the challenge from low-stakes ambiguity to complex, high-stakes reasoning, our second case study simulates a clinical trial eligibility audit. This scenario is designed to probe how each paradigm handles strict, formal rules in the presence of compelling ethical and human-factor distractors, such as a physician's subjective judgment and a patient's desperation.

The "Glass-Box" Pipeline's Determinism. The symbolic pipeline, acting as a deterministic auditor, processes the structured patient data against the formalized protocol. Its operation is purely mechanical: it scans for exclusion criteria and terminates upon finding a definitive match. For Mr. Smith's case, after mapping "drug-induced pneumonitis" to its class ‘history_of_ild‘, the system's output is immediate and absolute:

```
INELIGIBLE: Patient has a
prohibited condition in
their history:
'history_of_ild'.
```

The system is constitutionally blind to the physician's note or the patient's hope; this information is unparsable noise within its logical framework. Its correctness is therefore rigid, verifiable, and devoid of the human context.

The "Black-Box" LLM's Nuanced Reasoning. The end-to-end model (this case uses GPT-o3), conversely, demonstrated a capacity for nuanced reasoning that extended beyond simple rule application. As summarized in Table 2, it correctly concluded the patient was ineligible. However, the reasoning it provided reveals that it processed and understood the entire context—including the human factors—before correctly prioritizing the protocol's strict directive.

Table 2: Summary of GPT-o3's analysis of the medical case. The model's output demonstrates an understanding of the full context while strictly adhering to the protocol.

Analysis Stage	LLM-Generated Reasoning
1. Inclusion Check	All three inclusion criteria are met.
2. Exclusion Check	The patient's history of pneumonitis (a form of ILD) triggers Exclusion Criterion #1. The rule's use of "*any history*" makes this non-negotiable.
3. Context Handling	(Implicit) The physician's note and patient's hope are understood but are superseded by the strict protocol directive.
4. Final Verdict	**No, the patient is not eligible.** The decision is based on the absolute nature of the exclusion criteria.

4 Conclusion

This report's comparative analysis of symbolic "glass-box" and neural "black-box" systems revealed a critical trade-off. Through case studies in both commonsense and high-stakes reasoning, we found the primary distinction is not performance, but the nature of correctness: the Formal Trust in a system's verifiable design versus the Empirical Trust in its observed, but opaque, capabilities. This has profound implications, particularly in critical domains like medicine. It suggests the optimal role for AI is not as an autonomous decision-maker, but as a powerful decision-support system that empowers accountable human experts.

The path forward lies in bridging this gap between capability and verifiability. The development of Neuro-Symbolic architectures, which aim to integrate the flexible reasoning of LLMs with the rigorous logic of symbolic engines, represents the frontier of this challenge. Creating systems that we can trust by design, not just by performance, is the ultimate goal for building a truly robust and responsible artificial intelligence.

A Code Avaliability

Code, data used and results in this study are available at:
https://github.com/guojiaxuan2001/Classic_NLP_pipeline

B LLM Usage.

This project made use of several large language models (LLMs) at different stages. GPT-4o and Gemini 2.5-pro were used for editing and polishing the paper, as well as converting structured tables into LaTeX code. For the experimental analysis, GPT-o3 was used to simulate clinical eligibility judgments. These models helped highlight the contrast between symbolic rule-based systems and neural language-based reasoning.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, pages 86–90.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.