Why Prompt Brittleness Matters for Future Autonomous Driving

Yesterday:	Prompts generated text.
Today:	Prompts guide robots.
Tomorrow:	Prompts may influence vehicle behavior.
Therefore:	Prompt brittleness becomes a safety issue rather than a usability issue.

Supported by the ongoing migration of Vision-Language-Action architectures from embodied robotics into autonomous driving research:

> The historical answer to ambiguity in automotive systems was formal specification. The emergence of Vision-Language-Action systems reintroduces natural language into the control loop. The central question is therefore no longer whether prompt brittleness exists, but how future automotive architectures can constrain and formalize language-driven behavior sufficiently to achieve safety and certification.

3.8.2026 Thomas Scheuerle, TWG-FAS.ORG The text in this document was processed with AI tools.

Why Prompt Brittleness Matters for Future Autonomous Driving

Large Language Models have demonstrated remarkable capabilities in reasoning, planning, and interaction. However, a well-known limitation is **prompt brittleness**: small variations in phrasing, wording, context, or formatting can lead to significantly different outputs and behaviors. While such variability may be acceptable in conversational applications, it becomes increasingly problematic when AI systems are connected to the physical world.

The next generation of intelligent systems is moving beyond language generation toward Vision-Language-Action (VLA) architectures, which combine perception, language understanding, reasoning, and physical control within a unified framework. Originally developed within embodied AI and robotics research, **VLA systems allow robots to interpret natural language instructions**, observe their environment, and directly execute actions. Recent surveys describe this paradigm as one of the most important developments in embodied AI and robotics. https://arxiv.org/abs/2405.14093v6

Researchers are now actively transferring these concepts from robotics into the automotive domain. A recent survey on Vision-Language-Action models for autonomous driving reports that the rapid progress of multimodal large language models has enabled the **development of driving systems that integrate visual perception, natural language understanding, reasoning, and vehicle control within a common framework**. These systems aim to interpret high-level instructions, reason about complex traffic situations, and generate driving actions directly from multimodal inputs. https://arxiv.org/pdf/2506.24044

Why Prompt Brittleness Matters for Future Autonomous Driving

This transition fundamentally changes the importance of prompt robustness. In traditional automotive software, a lane-change command is represented by precisely specified interfaces, state machines, and formal requirements. In a VLA-driven system, however, high-level intent may increasingly be expressed through natural language or language-like representations. Consequently, linguistic ambiguity, underspecification, and prompt variation can become part of the control path.

Evidence from robotics research already indicates that this is not merely a theoretical concern. **Multiple studies have shown that current VLA systems exhibit significant sensitivity to instruction formulations and often fail to correctly distinguish semantic intent from surface-level phrasing**. Researchers have observed cases where models overfit to specific instruction templates, ignore language information entirely, or react unpredictably under seemingly minor linguistic perturbations. https://arxiv.org/abs/2601.04052

For autonomous driving, the implications are substantial. **A prompt variation that produces a slightly different answer in a chatbot may produce a different trajectory, maneuver choice, or risk assessment when embedded into a vehicle control system.** At the same time, future automotive functions are expected to operate in environments characterized by uncertainty, long-tail situations, and human interaction, precisely the areas where foundation-model-based reasoning appears most attractive. The automotive industry therefore faces a fundamental challenge: how can systems exploit the flexibility and generalization capabilities of language-based reasoning while preserving the determinism, predictability, and safety guarantees required for safety-critical functions?
https://arxiv.org/pdf/2506.24044

Prompt brittleness therefore evolves from a user-experience concern into a systems-engineering problem. Understanding this transition is essential for the design of future automotive AI architectures and motivates the search for mechanisms that can make language-driven decision systems robust, interpretable, and verifiable.

Overview

Large language models can behave deterministically for an identical prompt under idealized zero-temperature decoding, yet still produce sharply different outputs when the prompt is changed in ways that appear semantically equivalent to a human. This phenomenon can be called semantic prompt instability, prompt brittleness, or, in a more geometric metaphor, high curvature in the prompt-response landscape. Recent studies show that small changes in formatting, wording, paraphrase, examples, or lexical cues can cause measurable changes in model performance, model rankings, bias metrics, and hallucination rates. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design reports that meaning-preserving formatting changes caused performance differences of up to 76 accuracy points for LLaMA-2-13B in few-shot settings, and that sensitivity remained even when model size, number of examples, or instruction tuning increased.
The project Brittlebench showed: Quantifying LLM robustness via prompt sensitivity reports that semantics-preserving perturbations can degrade model performance by as much as 12%, alter relative model rankings in 63% of cases, and account for up to half of the performance variance for a model. The effect therefore matters not only for user experience, but also for scientific benchmarking, safety validation, and any domain where natural-language instructions are treated as specifications.

https://arxiv.org/abs/2310.11324 https://arxiv.org/abs/2603.13285

1. The Core Phenomenon

Assume an idealized inference setup: the model weights are fixed, the system prompt is fixed, the context window is fixed, retrieval is disabled or unchanged, decoding is deterministic, and no external entropy source is injected. In such a setting, the same prompt may produce the same output. But this does not imply that the model is semantically stable. The function being evaluated is not a function from meanings to outputs. It is a function from token sequences, context states, model parameters, and decoding rules to outputs. Human-equivalent meaning does not collapse to one canonical token representation.

For example, the following prompts could be judged equivalent by a human:

A. Summarize the safety risk in three concise bullet points.
B. Give me three short bullets explaining the main safety risk.
C. Identify the principal hazard and compress the explanation into three bullets.

A human might treat these as near-identical. A model may not. Prompt A may activate a “summary” pattern, Prompt

B may activate a more conversational pattern, and Prompt C may activate a hazard-analysis or engineering-risk pattern.

None of this requires random sampling. The change can be deterministic: a different input token sequence leads to different internal activations, different attention patterns, different next-token distributions, and finally a different answer.

https://aclanthology.org/2024.findings-emnlp.108/

2. Why “Same Meaning” Is Not the Same Input

Natural-language semantics is many-to-one from a human perspective: many texts can map to approximately the same intention. LLM inference is not many-to-one in that way. It operates over tokenized surface forms. So even if two prompts are semantically close, they may differ in:

Lexical anchors: “risk”, “hazard”, “failure mode”, and “safety issue” may point toward different regions of the training distribution.
Task framing: “summarize”, “analyze”, “identify”, “explain”, and “judge” can imply different output norms.
Implicit role assumptions: “as an engineer”, “for management”, “for a test report”, and “for a safety case” can change standards of evidence.
Formatting cues: bullets, numbered lists, table requests, JSON-like schemas, or examples can shift the response mode.
Evaluation affordances: a prompt may guide the model toward answers that are semantically correct but scored differently by rigid evaluation systems.

https://aclanthology.org/2025.emnlp-main.1006/

3. Prompt Brittleness and Hallucination

The phenomenon becomes more serious when prompt variations affect factual reliability. A semantically equivalent prompt can make the model more or less likely to invent details, overgeneralize, omit uncertainty, or present unsupported claims. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/pdf

If one wording elicits careful caveats and another wording elicits confident fabrication, then the natural-language interface itself becomes a risk surface.
https://arxiv.org/abs/2510.06265v1

A simple illustrative example:
Prompt 1:
List the known limitations of method X based only on the supplied document. If the document does not say it, write “not specified”.
Prompt 2:
Explain the limitations of method X.

If the document is incomplete, Prompt 1 constrains the answer to evidence. Prompt 2 however may invite the model to fill gaps from general knowledge or plausible patterns. Both prompts may seem aligned in ordinary conversation, but they encode different epistemic permissions.

4. Prompt Space as a Sharp Landscape

One useful abstraction is to imagine a landscape over prompts. Nearby prompts, by human semantic distance, should ideally produce nearby outputs by task-relevant quality metrics. In practice, the landscape can be “sharp”: a tiny paraphrase can move the result from correct to incorrect, from grounded to hallucinated, or from concise to verbose. Beyond Magic Words: Sharpness-Aware Prompt Evolving for Robust Large Language Models with TARE explicitly identifies this brittleness as “textual sharpness” in the prompt landscape and states that small, semantically preserving paraphrases often cause large performance swings. It also notes that many prompt optimization methods target point-wise accuracy rather than paraphrase invariance or search stability.
https://arxiv.org/abs/2509.24130v1

This matters because traditional prompt engineering often optimizes for one observed answer: “This prompt worked once, therefore it is good.” But if a prompt is located on a narrow peak, it may fail under minor edits, localization, translation, template insertion, variable substitution, or user-specific phrasing. This is analogous to overfitting: the prompt is not robust over a neighborhood of equivalent formulations.

Clean, static benchmarks can overestimate true model performance because they fail to capture the noise and variability of real-world user inputs, including mistakes, typos, or alternative phrasings of the same question. That observation is essential for any safety-critical or engineering-critical usage: a system validated on one canonical prompt may not be validated on the class of prompts that users will actually send.
https://arxiv.org/abs/2603.13285

5. Earlier Roots: Adversarial and Behavioral Testing

This issue is not entirely new. Earlier work on adversarial examples in NLP showed that systems could fail under changes that do not confuse humans. Adversarial Examples for Evaluating Reading Comprehension Systems reports that adding automatically generated distracting sentences to SQuAD paragraphs, without changing the correct answer or misleading humans, reduced average F1 accuracy of sixteen published models from 75% to 36%; when ungrammatical word sequences were allowed, average accuracy on four models dropped further to 7%. The lesson is structural: impressive average performance can hide fragile reliance on superficial cues.
https://aclanthology.org/D17-1215/

Similarly, Beyond Accuracy: Behavioral Testing of NLP Models with CheckList argues that held-out accuracy can overestimate NLP model performance and introduces a task-agnostic behavioral testing methodology inspired by software engineering. The paper reports that CheckList helped identify critical failures in commercial and state-of-the-art models and that practitioners using CheckList created twice as many tests and found almost three times as many bugs as users without it. This connects strongly to your intended discussion: if prompts are to serve as informal specifications, they need test suites, perturbation classes, invariance checks, and explicit expected behavior, not just one natural-language sentence.
https://aclanthology.org/2020.acl-main.442/

6. Deeper Implications

The deeper implication is that natural language is underspecified as a control interface. It carries meaning by convention, context, pragmatics, and shared human assumptions. But an LLM receives a token sequence and produces a continuation according to learned statistical and instruction-following patterns. The user may believe they specified a task; the model may have inferred a broader, narrower, or differently framed task.

This creates several risks:

- Reproducibility risk: A result may depend on accidental wording, punctuation, or formatting rather than the intended task.
- Benchmarking risk: Model comparisons may depend on arbitrary prompt templates. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design reports weak correlation of format performance between models, questioning comparisons based on a single fixed prompt format. https://arxiv.org/abs/2310.11324
- Safety risk: A benign paraphrase may remove constraints that prevented hallucination, unsafe advice, or unsupported claims.
- Governance risk: If a system is approved using one prompt but deployed with many variants, the approved artifact is not the actual deployed behavior.
- Specification risk: A prompt may mix task, policy, evidence constraints, output structure, and quality criteria in prose, making it hard to know what must remain invariant under paraphrase.
- Human factors risk: Users may interpret prompt variation as "same request", while the system treats it as a materially different computational input.

This tension is especially important in formal or engineering contexts. If an AI is used to generate requirements, test cases, safety arguments, code, or legal/technical summaries, then the prompt functions like a weak specification language. But unlike a formal specification, ordinary prose does not define equivalence classes, preconditions, forbidden inferences, evidence boundaries, or acceptance criteria.

7. Toward the Problem Statement

The problem can be stated as follows:

A natural-language prompt is not a stable semantic object. It is a concrete surface-form input whose small variations may induce large differences in model behavior, even when human readers judge those variations to preserve meaning. Therefore, any reliable AI-assisted workflow must distinguish between the user's intended task semantics and the accidental linguistic form by which that task was expressed.

This does not mean natural language is unusable. It means that for high-value tasks, natural language should be treated as an initial requirement capture layer, not as the final specification. Before discussing solutions such as formal prompt schemas, controlled natural language, ontology-backed task descriptions, precondition/evidence contracts, typed output formats, or prompt-invariance test suites, the foundational problem must be made explicit: current LLM behavior is sensitive to semantically minor input variation, and this sensitivity affects quality, factuality, reproducibility, and safety.

Selected Scientific Sources

Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design, Sclar et al., ICLR 2024: meaning-preserving formatting changes can produce very large performance differences, and reporting a single prompt format can be misleading.
https://arxiv.org/abs/2310.11324
ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs, EMNLP Findings 2024: proposes PromptSensiScore and studies how sensitivity varies by model, dataset, confidence, few-shot examples, and subjective evaluation.
https://aclanthology.org/2024.findings-emnlp.108/
Flaw or Artifact? Rethinking Prompt Sensitivity in Evaluating LLMs, EMNLP 2025: argues that part of observed prompt sensitivity can be an artifact of rigid evaluation methods.
https://aclanthology.org/2025.emnlp-main.1006/
Brittlebench: Quantifying LLM robustness via prompt sensitivity, 2026: introduces a framework and evaluation pipeline for semantics-preserving perturbations and reports performance degradation and ranking instability.
https://arxiv.org/abs/2603.13285
Beyond Magic Words: Sharpness-Aware Prompt Evolving for Robust Large Language Models with TARE, 2025: formalizes the idea of textual sharpness and robust performance over semantic neighborhoods.
https://arxiv.org/abs/2509.24130v1
Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior, 2025: connects hallucination to prompt sensitivity and model variability.
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/pdf
A Comprehensive Survey of Hallucination in Large Language Models: Causes, Detection, and Mitigation, 2025/2026: surveys hallucination causes, detection, and mitigation, defining hallucination as fluent but factually inaccurate or unsupported generation.
https://arxiv.org/abs/2510.06265v1
Adversarial Examples for Evaluating Reading Comprehension Systems, Jia and Liang, EMNLP 2017: shows that superficial but human-nonmisleading changes can strongly degrade NLP system performance.
https://aclanthology.org/D17-1215/
Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, Ribeiro et al., ACL 2020: argues for behavioral testing beyond aggregate accuracy and provides a methodology for finding capability-specific failures.
https://aclanthology.org/2020.acl-main.442/