

BLACKGRID LABS

Infrastructure Beneath the World

On the Limits of Statistical Language Systems

A practical boundary for language models: useful assistance, with independently checked evidence and rules for consequential conclusions.

BGL/DI-R1 | PUBLIC RESEARCH

Published 2026-08-02 | Revised 2026-10-11

blackgridlabs.com | Part of The Veydros Group

In essence

Why we are talking about this

A fluent answer can still be wrong. Institutions need to distinguish useful generated text from a conclusion they can defend with evidence.

Purpose

Define where language-model assistance fits and what checks a consequential decision needs.

What we found and focus on

- The Transformer is an architecture; it does not itself guarantee factual accuracy. [1]
- GPT-4's technical report describes next-token pretraining and subsequent alignment. [2]
- TruthfulQA documents false answers on a defined benchmark of earlier models. [3]
- Those results do not prove that every present or future model has the same limitations.
- A generated explanation is not automatically a valid derivation.
- Retrieval, tools and human review can help, but need evaluation on the actual task.
- A deterministic verifier can also be wrong if its rules or inputs are wrong.
- BGL's proposed boundary puts accountable checks between model output and consequential decisions.

Conclusion and next steps

Specify the task, evidence requirements and verifier. Evaluate the complete workflow on held-out cases, including misleading sources and correct-looking but unsupported answers.

Full revised research

The question this brief addresses

Language models can help draft, extract and organize information. The question here is what must happen before an institution treats an output as a consequential conclusion. Readability and confidence are not sufficient evidence.

This edition states a proposed workflow boundary. It does not disclose the private initiative's methods, partners or program, and it does not claim results from a sealed BGL evaluation.

What the cited literature shows

Attention Is All You Need describes the Transformer architecture. The GPT-4 technical report describes next-token pretraining followed by alignment, while also reporting capabilities and limitations. Neither supplies a general proof that a generated answer is correct. [1][2]

TruthfulQA measured responses to questions designed to expose common false beliefs in a specified set of earlier models. That finding is evidence about a benchmark and those models, not a universal impossibility theorem. New systems and workflows need their own evaluation. [3]

Explanations and evidence

An explanation can sound like a derivation without establishing that each step follows from an admitted premise. For a consequential conclusion, identify the source facts, applicable rules and checks that connect them to the result.

A citation should support the associated claim. A real link attached to an unsupported sentence is still an unsupported sentence. Check whether the source says what the answer claims, whether it is current enough and whether contradictory evidence has been omitted.

A proposed operating boundary

Use model output as a candidate where that helps the task. Validate extracted facts against their sources and test proposed calculations or code with an appropriate checker. Keep a named owner responsible for the final decision.

Independent checking can include deterministic rules, formal verification or human review, depending on the task. Research on power-system neural networks shows that formal guarantees can be developed within specified models and domains. That supports carefully bounded verification rather than a blanket exclusion of learned tools. [4]

How to evaluate the whole workflow

Freeze the task, dataset and scoring rules before testing. Include unsupported answers, contradictory sources, stale documents and cases that require refusal. Score the final accepted conclusions, not just the model's raw responses.

Test the verifier too. It may accept a valid calculation built from a false premise or reject a correct answer for the wrong reason. Record human overrides and review whether they improve outcomes. Repeatability is useful evidence about consistency, not a substitute for correctness.

Conclusion and next steps

BGL's proposed position is that consequential conclusions need accountable evidence and validation. Language models can contribute where the full workflow meets that standard.

Publish a bounded evaluation protocol and independent results before making broader performance claims. Keep the initiative's restricted research private.

Sources

Numbered citations refer to these sources. Proposed methods and next steps are BGL research proposals, not results established by the cited publications. Links checked for this revision on October 11, 2026.

[1] Vaswani et al.: Attention Is All You Need (2017)

https://arxiv.org/abs/1706.03762

[2] OpenAI: GPT-4 Technical Report (2024 revision)

https://arxiv.org/abs/2303.08774v6

[3] Lin, Hilton and Evans: TruthfulQA, Measuring How Models Mimic Human Falsehoods (2022)

https://arxiv.org/abs/2109.07958v2

[4] Venzke and Chatzivasileiadis: Verification of Neural Network Behaviour, Formal Guarantees for Power System Applications (2020)

https://arxiv.org/abs/1910.01624v4