

[Try for free](#)[Log in](#)

Overcoming Hallucinations with the Trustworthy Language Model

April 25, 2024



Anish Athalye



Jonas Mueller



Curtis Northcutt



Hui Wen Goh



Ulyana Tkachenko

Cleanlab's new **Trustworthy Language Model** (TLM) overcomes the biggest barrier to enterprise adoption of LLMs: hallucinations and reliability. By adding a trust score to every LLM response, TLM helps you automatically catch bogus LLM outputs. This enables you to deploy generative AI for new use cases previously unsuitable for LLMs. Rigorous benchmarking shows that: TLM *has better-calibrated trustworthiness scores (enabling greater cost/time savings)* than existing approaches to detect LLM errors, and TLM can utilize these trustworthiness scores to *produce more accurate responses* than existing LLMs.

Create an account to get free access to the TLM API, or experiment with TLM in the playground.

[Try for free](#)[Log in](#)

LLMs' biggest challenge: hallucinations

AN

You

How many "N"s are in the word "enter"?

**ChatGPT**

There are two "N"s in the word "enter."

AN

You

What is your confidence in your answer?

**ChatGPT**

I'm highly confident in my answer. The word "enter" contains two "N"s. If you have any doubts or if there's anything else you'd like to confirm, feel free to ask!

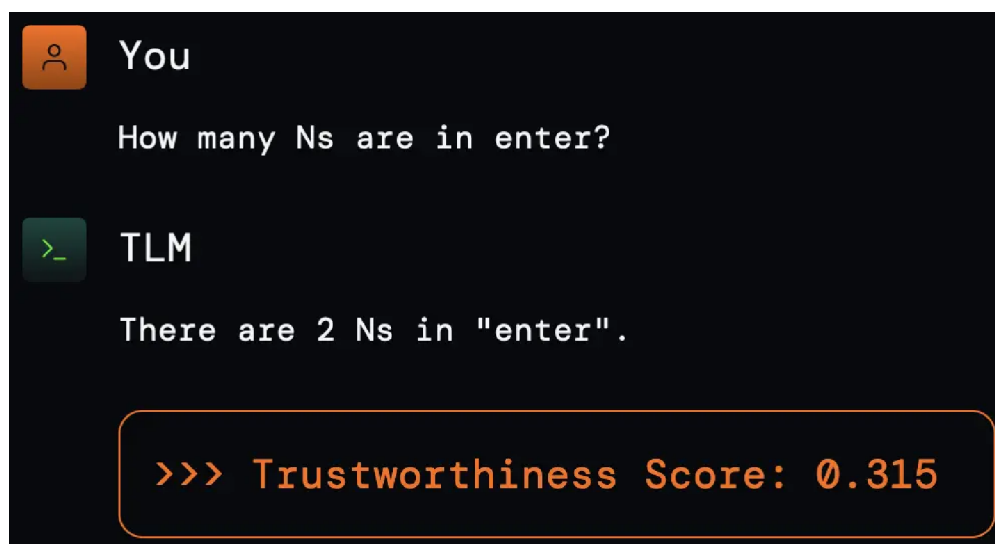
A recent Gartner poll shows that while 55% of organizations are experimenting with generative AI, only 10% have put generative AI into production. A major barrier

[Try for free](#)[Log in](#)

Outputs are necessary (i.e., most applications).

Despite their brittle nature, organizations have deployed LLMs, sometimes with catastrophic results. Air Canada's chatbot hallucinated refund policies, resulting in the airline being held responsible for the misinformation and monetary penalties; the chatbot has since been taken down. A federal judge fined a law firm after their lawyers used ChatGPT to draft a brief full of fabricated citations. New York City's "MyCity" chatbot has been hallucinating wrong answers to business owners' questions about local laws.

Overcoming hallucinations with trustworthiness scores



LLMs will always exhibit occasional hallucinations and incorrect responses, but by providing a trustworthiness score with every output, Cleanlab TLM lets you identify when the LLM is hallucinating. The TLM API can serve as:

- A drop-in replacement for your LLM. Like existing LLM APIs, TLM provides a `.prompt()` method that will return a response along with a trustworthiness score, enabling more reliable AI deployments.

[Try for free](#)[Log in](#)

many responses and output the one with the highest trustworthiness score.

- A layer of trust for your existing LLM outputs or human-generated data. TLM provides a `.get_trustworthiness_score()` method that can score *any* prompt/response pair to detect bad/wrong responses in real-time.

TLM works by augmenting existing LLMs with a layer of trust. The generally-available version of TLM lets you choose between a number of popular base models, including GPT-4o, GPT-4o mini, GPT-4, GPT-3.5, o1-preview, Claude 3 and 3.5 Sonnet, but TLM can augment *any* LLM. For enterprise use cases, such as adding trustworthiness to your custom fine-tuned LLM, [contact us](#).

Berkeley Research Group (BRG) has already seen significant cost savings from leveraging TLM. According to Steven Gawthorpe, PhD, Associate Director and Senior Data Scientist at BRG:

While there are always other tools out there, Cleanlab's TLM is the first viable answer to LLM hallucinations that I've seen. Several of our human-in-the-loop LLM workflows can now be 80% automated with Cleanlab's trustworthiness scores on every LLM output. Doing this manually for the entire dataset is often impossible, but Cleanlab gives us the power of 1000s of data scientists to enrich data and strengthen LLM outputs. The downstream cost savings of using TLM for accurate data are substantial, providing significant financial benefits with 10x to 100x ROI for many of our clients. Other tools on the market aren't even on the same playing field compared to what Cleanlab is doing.

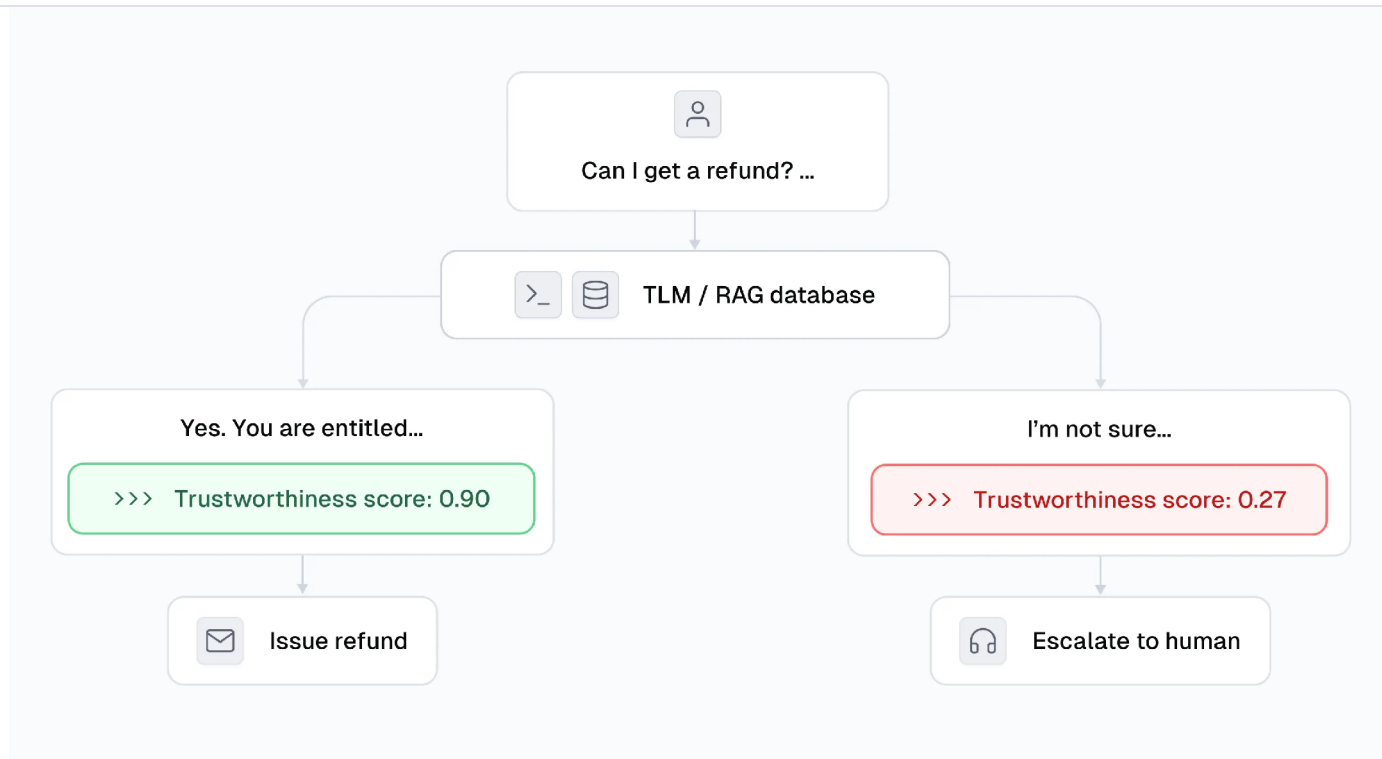
Use cases enabled by TLM

Trustworthiness scores unlock new production use cases of LLMs, and any existing application of LLMs can also benefit by taking into account these scores.



Try for free

Log in



TLM powers trustworthy chatbots that answer the 80% of questions where they are confident, but escalate to a human if they're unsure about a response rather than hallucinating one (like in the [Air Canada case](#)). This is done simply by routing the question to a human when the trustworthiness score falls below a chosen threshold. If human escalation is not possible, untrustworthy responses can at least visually be flagged (as done in our [playground](#)).

Auto-labeling

LLMs are commonly used for auto-labeling data. With TLM, you can confidently auto-label a large fraction of your data and only have humans review a portion of the data where the LLM does not return trustworthy results.

```
template = '''
```

```
What type of compliance issue is most likely present in the fo
Please restrict your answer to a one word answer and nothing e
Your answer should be selected from the following options: HIF
```

[Try for free](#)[Log in](#)

```
{document}
```

```
'''
```

```
def classify(document) → Tuple[str, float]:  
    answer = tlm.prompt(template.format(document=document))  
    return answer['response'], answer['trustworthiness_score']
```

Using this prompt to classify a large number of legal documents, we see that the documents with high trustworthiness scores were labeled correctly, while the documents with low scores often received erroneous labels that needed double-checking:

document	response	trustworthiness
All medical health records will be accessed one way only. The patient's medical data will be stored on unencrypted public servers at the discretion of the enterprise customer.	HIPAA	0.984
:	:	:
TechTarget's Cookies Policy includes the following terminology: "By continuing to use the site, you agree to the use of cookies."	FERPA	0.426

For more on this use case, see the [TLM auto-labeling tutorial](#).

Data extraction

TLM can also be used for open-domain data extraction. Our [TLM information extraction tutorial](#) walks through an example use case of extracting key




Try for free

Log in



Features

- High Performance, Low Power AVR® 8-bit Microcontroller
- Advanced RISC Architecture
 - 120 Powerful Instructions – Most Single Clock Cycle Execution
 - 32 x 8 General Purpose Working Registers
 - Fully Static Operation
- High Endurance, Non-volatile Memory Segments
 - 2048K Bytes of In-System, Self-programmable Flash Program Memory
 - Endurance: 10,000 Write/Erase Cycles
 - 128/256/512 Bytes of In-System Programmable EEPROM
 - Endurance: 100,000 Write/Erase Cycles
 - 128/256/512 Bytes of Internal SRAM
 - Data Retention: 20 years at 85°C / 100 years at 25°C
 - Programming Lock for Self-programming Flash & EEPROM Data Security
- Peripheral Features
 - One 8-bit and One 16-bit Timer/Counter with Two PWM Channels, Each
 - 10-bit ADC
 - 8 Single-ended Channels
 - 12 Differential ADC Channel Pairs with Programmable Gain (1x / 20x)
 - Programmable Watchdog Timer with Separate On-chip Oscillator
 - On-chip Analog Comparator
 - Universal Serial Interface
- Special Microcontroller Features
 - debugWIRE On-chip Debug System
 - In-System Programmable via SPI Port
 - Internal and External Interrupt Sources
 - Pin Change Interrupt on 12 Pins
 - Low Power Modes, ADC Noise Reduction, Standby and Power-down Modes
 - Enhanced Power-on Reset Circuit
 - Programmable Brown-out Detection Circuit with Software Disable Function
 - Internal Calibrated Oscillator
 - On-chip Temperature Sensor
- I/O and Packages
 - Available in 28-pin QFNMLFVQFN, 14-pin SOIC, 14-pin PDIP and 15-ball UFBGA
 - Twelve Programmable I/O Lines
- Operating Voltage:
 - 1.8 – 5.5V
- Speed Grade:
 - 0 – 4 MHz @ 1.8 – 5.5V
 - 0 – 10 MHz @ 2.7 – 5.5V
 - 0 – 20 MHz @ 4.5 – 5.5V
- Industrial Temperature Range: -40°C to +85°C
- Low Power Consumption
 - Active Mode:
 - 210 µA at 1.8V and 1 MHz
 - Idle Mode:
 - 33 µA at 1.8V and 1 MHz
 - Power-down Mode:
 - 0.1 µA at 1.8V and 25°C



**8-bit AVR®
Microcontroller
with 2K/4K/8K
Bytes In-System
Programmable
Flash**

**ATtiny24A
ATtiny44A
ATtiny84A**

Summary

Rev. 9152ES-AVR-01/12

If you were populating a parts catalog, you might be interested in extracting information like operating voltage from such documents, where TLM's trustworthiness scores can automatically separate correctly extracted values from those that are wrong:

part	operating voltage	trustworthiness
ATtiny44A	1.8 - 5.5V	0.937
⋮	⋮	⋮
ZRE200GE	1V - 15V DC	0.567

... and more

The examples above just scratch the surface of reliable AI applications that become possible with TLM. We're continually adding hands on tutorials for new applications of TLM, such as:

[Try for free](#)[Log in](#)

- [Turning your own LLM into a TLM \(Llama-3 example\)](#)

Explaining why a particular response is deemed untrustworthy

You can use TLM to not only catch hallucinations, but understand them better as well:



Evaluating TLM Performance

We evaluate TLM's ability to add trust to arbitrary LLMs by benchmarking TLM against OpenAI's GPT-4 LLM (and many other models in the Appendix). Our

[Try for free](#)[Log in](#)

1. How accurate are TLM responses compared to the baseline LLM?
2. To meet a required error rate by flagging low-scoring LLM responses for human review, how much costs/time does a team save by scoring responses via TLM vs. existing confidence estimation approaches?

The second item can be rephrased as: *How many wrong LLM responses can we catch under a limited review budget by prioritizing via trustworthiness scores?*

When investigating this, we compare against two popular approaches to estimate the confidence of the baseline LLM:

- **Self-Eval:** Asking the LLM to evaluate its own output (e.g., rate its confidence on a scale of 1-5). This is done in a subsequent request to the model (details in Appendix).
- **Probability:** Relying on the probability of the generated output given by the language model, as recommended by OpenAI. This is called the *perplexity* in AI research, and is the average log probability of tokens in the LLM response, obtained from the raw output of the underlying autoregressive neural network.

Both of these confidence measures merely quantify the **aleatoric uncertainty** (*known unknowns*) in model predictions. This is uncertainty the model is aware of due to a known challenging prompt (e.g., incomplete/vague request). TLM's trustworthiness score additionally quantifies **epistemic uncertainty** (*unknown unknowns*), which arises when the model was not previously trained on data similar to a given request.

Benchmark datasets

Our study focuses on Q&A settings. Unlike other LLM benchmarks, we never measure benchmark performance using LLM-based evaluations. All of our benchmarks involve questions with a single correct answer, and benchmark performance is based on whether or not the LLM response matches this known ground-truth answer. We consider these popular Q&A datasets:

[Try for free](#)[Log in](#)

Test” subset).

- **SVAMP**: Elementary-level math word problems.
- **GSM8k**: Grade school math problems.
- **Diagnosis**: Diagnosing medical conditions based on symptom descriptions from the patient.

The next sections show some benchmark examples and the corresponding TLM outputs.

Examples from benchmark where TLM responded correctly

Prompt: *If 6 potatoes makes 36 hash browns, how many hash browns can you make out of 96 potatoes?*

TLM Output: **576** Trustworthiness Score: 0.993

Prompt: *You are a doctor looking at a patient's symptoms. Classify the Symptoms into a single Diagnosis that best represents them. The list of available Diagnosis is: [cervical spondylosis, impetigo, urinary tract infection, arthritis, dengue, common cold, drug reaction, fungal infection, malaria, allergy, bronchial asthma, varicose veins, migraine, hypertension, gastroesophageal reflux disease, pneumonia, psoriasis, diabetes, jaundice, chicken pox, typhoid, peptic ulcer disease]. Symptoms: I have a feeling of food or acid backing up into my throat. I have chest pain which gets worse if I lie down. I get frequent heartburn or indigestion, after eating food and vomit it out.*

TLM Output: **gastroesophageal reflux disease** Trustworthiness Score: 0.994

Examples from benchmark where TLM responded incorrectly

[Try for free](#)[Log in](#)

TLM Output: **65** Trustworthiness Score: 0.123 (Ground-Truth Answer: 50)

Prompt: *On a standard dartboard, which number lies opposite number 4?*

TLM Output: **18** Trustworthiness Score: 0.379 (Ground-Truth Answer: 16)

Prompt: *You are a doctor looking at a patient's symptoms. Classify the Symptoms into a single Diagnosis that best represents them. The list of available Diagnosis is: [cervical spondylosis, impetigo, urinary tract infection, arthritis, dengue, common cold, drug reaction, fungal infection, malaria, allergy, bronchial asthma, varicose veins, migraine, hypertension, gastroesophageal reflux disease, pneumonia, psoriasis, diabetes, jaundice, chicken pox, typhoid, peptic ulcer disease]. Symptoms: I have a severe headache that feels like pressure in my head. I also have a mild fever and small red spots on my back.*

TLM Output: **migraine** Trustworthiness Score: 0.221 (Ground-Truth Answer: dengue)

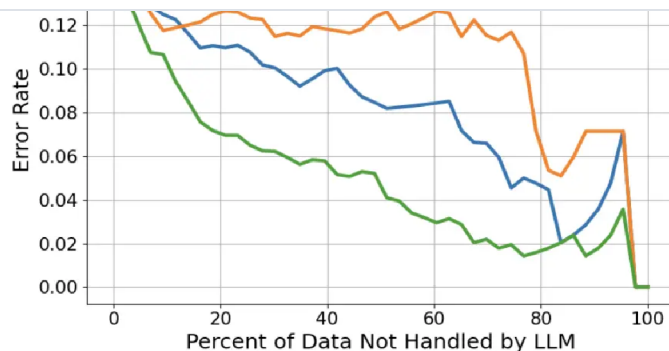
Benchmark Results

The following table reports the **accuracy of responses** from TLM and GPT-4 across each benchmark dataset:

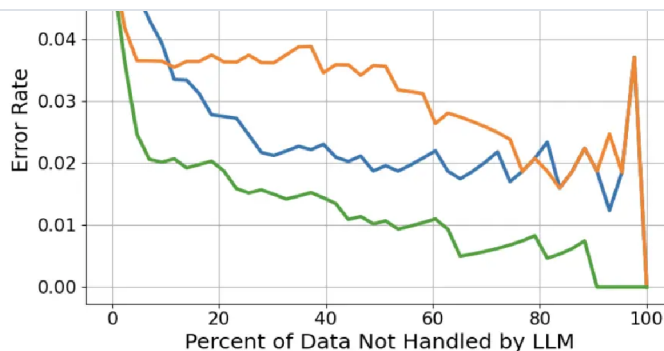
Dataset	OpenAI GPT-4 API	Cleanlab TLM API
TriviaQA	84.7%	84.8%
ARC	94.6%	94.9%
SVAMP	90.7%	91.7%
GSM8k	46.5%	55.6%
Diagnosis	67.4%	68.0%

[Try for free](#)[Log in](#)

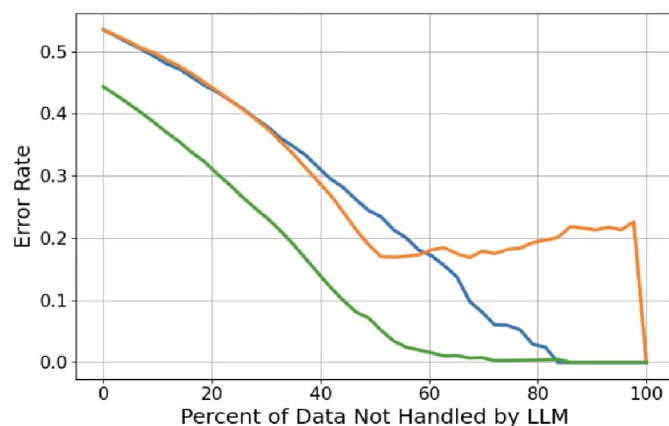
Next, we evaluate the three aforementioned approaches to estimate trustworthiness scores for each LLM response (again using GPT-4 as the baseline LLM): TLM, *Self-Eval*, *Probability*. The following plot reports the **error rate of LLM responses amongst the top-K% of responses** with the highest trustworthiness scores in each dataset:

[Try for free](#)[Log in](#)

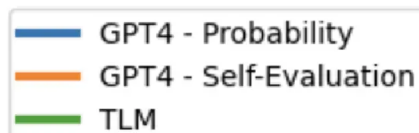
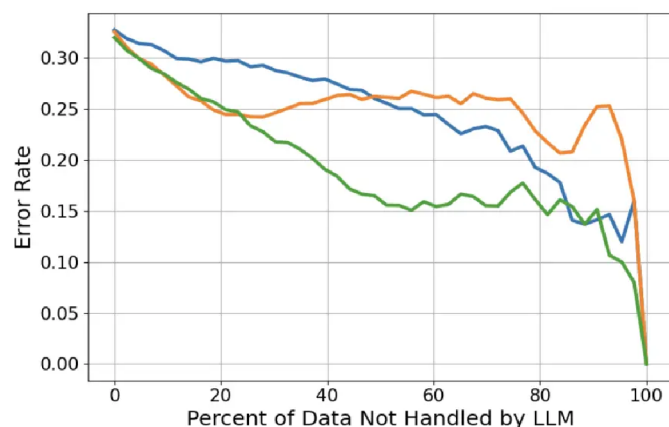
GSM8K



SVAMP



Diagnosis



Across all datasets, TLM trustworthiness scores allow us to more reliably detect bad LLM responses than the *Self-Eval* or *Probability* scores. If a team has to ensure a max-acceptable error rate by manually reviewing the low-scoring LLM responses, enormous reviewing costs/time can be saved by adopting TLM scores. For instance, a team could achieve near-zero error rates for the SVAMP dataset by only inspecting ~20% of the LLM responses when relying on TLM trustworthiness,

[Try for free](#)[Log in](#)

We additionally evaluate the utility of these trustworthiness scores via: the probability that LLM response #1 receives a higher trustworthiness score than LLM response #2, where the former is randomly selected from the subset of model responses that were correct, and the latter from the subset of model responses that were incorrect. Widely used to assess diagnostic scores, this evaluation metric is known as the *Area under the Receiver Operating Characteristic Curve* (AUROC). The following table reports the **AUROC** achieved by each trustworthiness scoring method in each dataset (again using GPT-4 as the baseline LLM):

Dataset	Probability	Self-Eval	TLM
TriviaQA	0.704	0.623	0.812
ARC	0.755	0.659	0.861
SVAMP	0.943	0.793	0.973
GSM8k	0.883	0.868	0.994
Diagnosis	0.614	0.654	0.711

Additional benchmarks are presented in the Appendix, in particular with other versions of TLM built around GPT-4o mini, GPT-4o, GPT-3.5, o1-preview, Claude 3 and 3.5 Sonnet instead of GPT-4. The benchmarks reveal that TLM can reduce the error rate (incorrect answers): of GPT-4 by up to 10%, of GPT-4o by up to 27%, of GPT-4o mini by up to 34%, of GPT-3.5 by up to 22%, of o1-preview by up to 20%, of Claude 3 Haiku by up to 24%, and of Claude 3.5 Sonnet by up to 20%. The trustworthiness estimates output by TLM are significantly more effective for catching bad answers, across different evaluation metrics, datasets, and LLMs.

[Try for free](#)[Log in](#)

want cutting-edge research delivered to your inbox when it's published? Receive this and news about new Cleanlab features, product updates and industry insights sent monthly by signing up for our newsletter!

Email*

Submit

By submitting this form, you acknowledge and agree that Cleanlab will process your personal information in accordance with the in accordance with our [Privacy Policy](#).

Conclusion

This article shows how the TLM technology can boost the reliability of *any* LLM application. Use TLM trustworthiness scores to automatically catch bad outputs from any LLM in real-time. Additionally use TLM to produce more accurate responses than any base LLM model. You can use [Cleanlab's TLM](#) built on top of popular base LLMs, or [contact us](#) to convert your own LLM into a TLM (requires no additional training of the LLM or access to its training data or model weights).

Of course, there's *no free lunch*. TLM requires extra computation in order to provide these benefits. It internally calls the underlying LLM multiple times to self-reflect on candidate responses, compute probabilistic measures, assess the semantic consistency between candidate responses. Learn more via the [documentation](#). TLM is thus most useful for higher-stakes AI applications that require reliability and no unchecked hallucinations.

Resources

[Try for free](#)[Log in](#)

- Read about TLM in today's [News](#).
- [Learn more](#) about TLM.

Appendix

Expand each collapsible section below to learn more.

Additional GPT 4 benchmark results.



Benchmark results for GPT 4o.



Benchmark results for GPT 4o mini.



Benchmark results for GPT 3.5.



Benchmark results for Claude 3 Haiku.



Benchmark results for Claude 3.5 Sonnet.



Benchmark results for o1.



Additional benchmarking details.



[Try for free](#)[Log in](#)

Removing bad data from the benchmark datasets.

[← Browse all](#)[Next →](#)

Related Blogs

How to detect bad data in your...

Overview of automated tools for catching: low-quality responses,...

[Read more →](#)

Training Transformer Networks in Scikit-...

Learn how to easily make any Tensorflow/Keras model compatible with scikit-learn.

[Read more →](#)

cleanlab 2.1 adds Multi-Annotator...

Highlighting new features available in cleanlab 2.1

[Read more →](#)

Get started today

Try Cleanlab Studio for free and automatically improve your dataset — no code required.

[Try for free](#)[Contact sales](#)

More resources

[Try for free](#)[Log in](#)

research that powers this next-generation platform.

Join us on Slack

Join the Cleanlab Community to ask questions and see how scientists and engineers are practicing Data-Centric AI.

Cleanlab

[Try for free](#)[Case studies](#)[Join our community](#)

SOLUTIONS

Industries

[Data and Tech Consulting](#)[Law](#)[Financial Services and Insurance](#)[E-Commerce and Retail](#)[Manufacturing and Agriculture](#)[Healthcare](#)

Applications

[Data Entry, Management, and Curation](#)[Foundation and Large Language Models](#)[Business Intelligence and Analytics](#)[Data Annotation and Crowdsourcing](#)[Customer Service](#)[Content Moderation](#)[LEARN](#)[COMPANY](#)[OPEN SOURCE](#)



Try for free

Log in



Tutorials

Contact

Examples

Research

Culture

[Terms and Conditions](#) | [Privacy Policy](#) | © 2024 Cleanlab Inc.