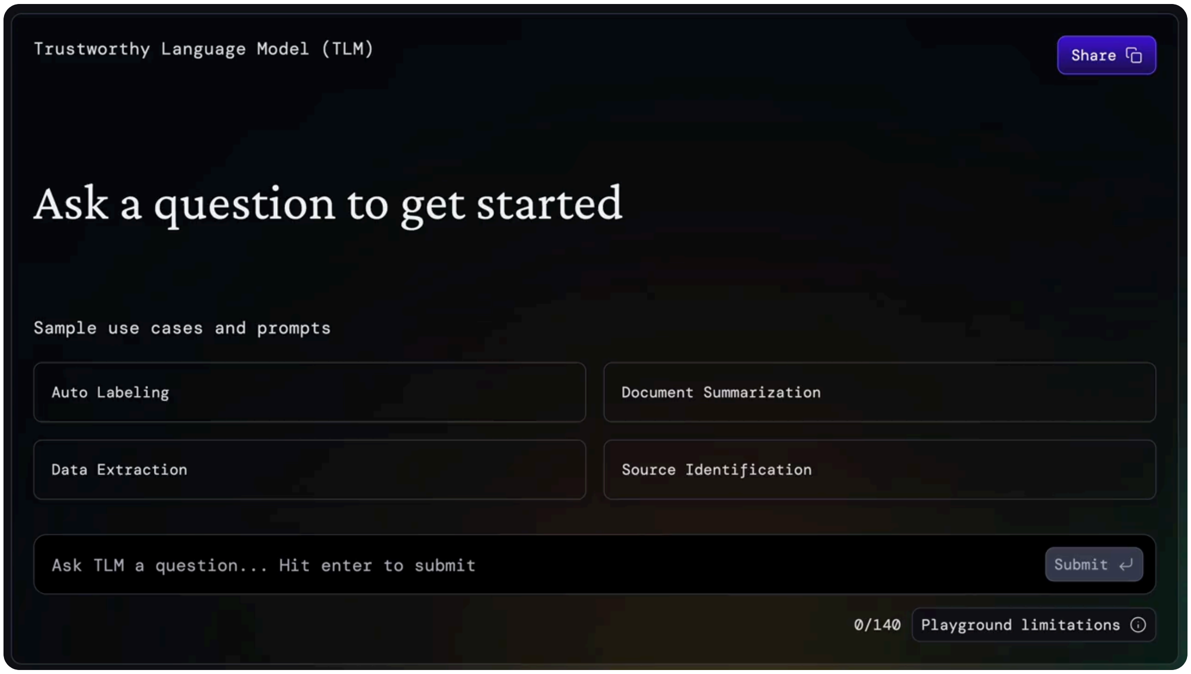
Try for free Log in

Trustworthy Language Model (TLM)

Reliability and explainability added to every LLM output. Smart-routing for LLM-automated responses and decision-making using trustworthiness scores for every LLM output. Cleanlab can also help your organization turn any LLM into a TLM.

Try for free Contact sales → Explore the tutorials ↗



Trustworthy Language Model (TLM)

Ask a question to get started

Sample use cases and prompts

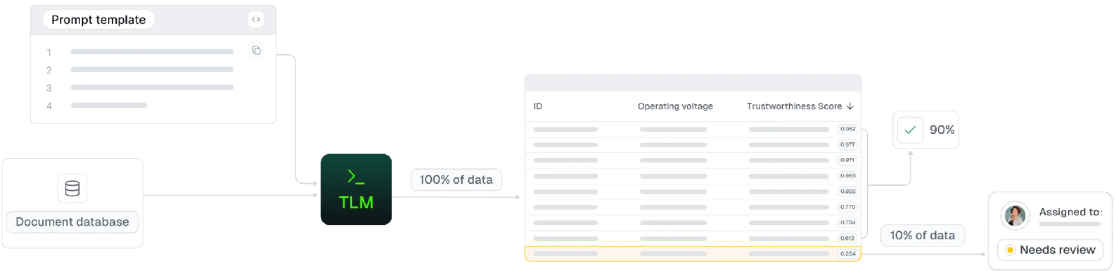
Launch TLM playground ↗

The Problem with LLMs

undermine your business.

Introducing the Trustworthy Language Model

Cleanlab's Trustworthy Language Model (TLM) is the solution. It adds a trustworthiness score to every LLM response, letting you know which outputs are reliable and which ones need extra scrutiny. TLM is a robust LLM designed for high-quality outputs and enhanced reliability—perfect for enterprise applications where unchecked hallucinations are unacceptable. Get started with our quick Python API tutorial, or read about use-cases and benchmarks here.



Key Features

- **Trustworthiness Scores:** Each response comes with a trustworthiness score, helping you reliably gauge the likelihood of hallucinations.
- **Higher Accuracy:** Rigorous benchmarks show TLM consistently produces more accurate results than other LLMs like GPT 4 / 4o and Claude.
- **Scalable API:** Designed to handle large datasets, TLM is suitable for most enterprise applications, including data extraction, tagging/labeling, Q&A (RAG), and more.

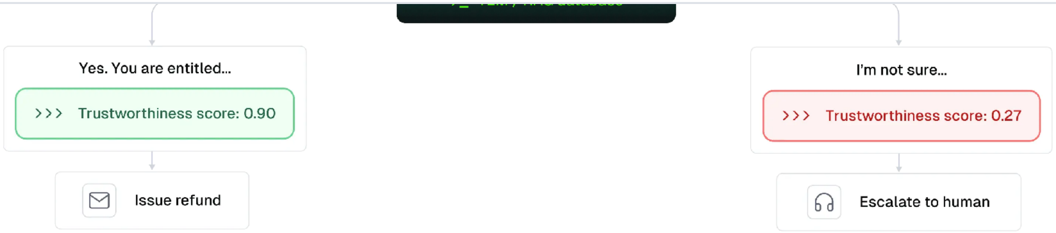
machine learning models.

Named the **Top AI Hallucination Detection Tools of 2024** in Analytics India Magazine.

Unlocking Enterprise Use Cases

With TLM, the possibilities are broader than ever before:

- **Retrieval-Augmented Generation.** TLM tells you which RAG responses are unreliable by providing a trustworthiness score for every RAG answer relative to a given question. Ensure users don't lose trust in your Q&A system and review untrustworthy responses to discover possible improvements. Explore the tutorial
- **Chatbots.** TLM informs you which LLM outputs you can use directly (refund, reply, auto-triage) and which LLM outputs you should flag or escalate for human review based on the corresponding trustworthiness score. With standard LLM APIs, you do not know which outputs to trust. Explore the tutorial
- **Auto-Labeling:** Streamline your data annotation process. TLM auto-labels data with high accuracy and reliable confidence scores. Let LLMs automatically process the subset of the data that they can reliably handle. Explore the tutorial
- **Extraction.** TLM tells you which data auto-extracted from documents, databases, transcripts is reliable and which should be double checked for review. This enables teams to transform raw unstructured data into structured data, with less errors and 90% less time spent reviewing outputs. Explore the tutorial



Explainability built-in: explore why a particular response is deemed untrustworthy

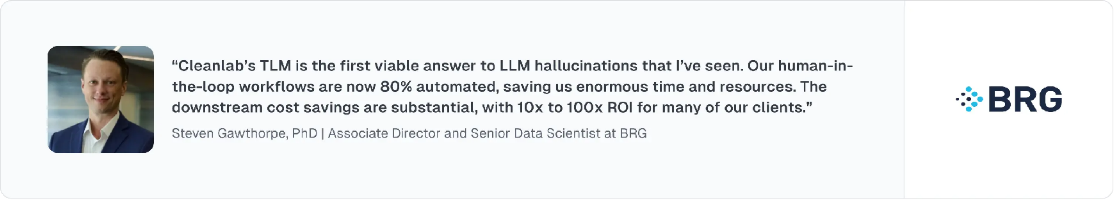
You can use TLM to not only catch hallucinations, but understand them better as well:

Proven Impact on Enterprise Deployment

Cleanlab TLM can be integrated into your existing LLM-based workflows to improve accuracy and reliability. With a trustworthiness score for each

Try for free

Log in



"Cleanlab's TLM is the first viable answer to LLM hallucinations that I've seen. Our human-in-the-loop workflows are now 80% automated, saving us enormous time and resources. The downstream cost savings are substantial, with 10x to 100x ROI for many of our clients."

Steven Gawthorpe, PhD | Associate Director and Senior Data Scientist at BRG

BRG

Turn your LLM into a TLM today

It's free to try, with no credit card required.

Try for free

Book a demo →

FAQ

How does TLM work?

The TLM scores our confidence that a response is *good* for a given request. In question-answering applications, *good* would correspond to whether the answer is correct or not. In general open-ended applications, *good* corresponds to whether the response is helpful/informative and clearly better than alternative hypothetical responses. For extremely open-ended requests, TLM trustworthiness scores may not be as useful as for requests that are questions seeking a correct answer.

TLM trustworthiness scores are a form of machine learning model uncertainty estimate. Machine learning models may produce uncertain outputs when given inputs that are fundamentally difficult (i.e. prompts that are vague or complex) or different from the model's training data (i.e. prompts that are atypical or based on niche information/facts).

TLM comprehensively quantifies the uncertainty in responding to a given request via multiple operations:

- **self-reflection**: a process in which the LLM is asked to explicitly rate the response and explicitly state how confidently good this response appears to be.

word/token in a sequence).

- **observed consistency**: a process in which the LLM probabilistically generates multiple plausible responses it thinks could be good, and we measure how contradictory these responses are to each other (or a given response).

These operations produce various trustworthiness measures, which are combined into an overall trustworthiness score that captures all relevant types of uncertainty. For instance, vague requests that could be answered in many ways generally yield lower token probabilities. Self-reflection can detect error-prone reasoning steps made when processing complex requests as well as unsupported factual statements. Observed consistency addresses fragilities in generation like tokenization/sampling that can yield vastly different responses.

For more details, refer to our research paper published at ACL, the top venue for NLP and Generative AI research. Our publication rigorously describes certain foundational components of the TLM system.

How well does TLM work?

Comprehensive benchmarks are provided in our blog. These reveal that TLM detects hallucinations/errors with significantly **higher precision/recall** than other methods, across many datasets, tasks, and LLM models.

Using the `best` or `high` quality_preset, TLM can additionally return *more accurate* LLM responses than the base LLM. Our benchmarks show that TLM can reduce the error rate (incorrect answers): of GPT-4o by 27%, of GPT-4o mini by 34%, of GPT-4 by 10%, of GPT-3.5 by 22%, and of Claude 3 Haiku by 24%.

I am using LLM model ___, how can I use TLM?

Two primary ways to use TLM are the `prompt()` and `get_trustworthiness_score()` methods. The former can be used as a drop-in replacement for any standard LLM API, returning trustworthiness scores in addition to responses from one of TLM's supported base LLM models. Here the response and trustworthiness score are both produced using the same LLM model.

your own custom (private) LLM, get in touch regarding our Enterprise plan.

What about latency-sensitive applications (like Chat)?

See our advanced tutorial for tips on trading-off performance vs. costs/latency, including quality-presets and TLM Lite.

Instead of using TLM to produce responses, you can alternatively stream in responses from your own LLM, and use `TLM.get_trustworthiness_score()` to subsequently stream in the corresponding trustworthiness score.

Cleanlab

Try for free

Case studies

Join our community

SOLUTIONS

Industries

Data and Tech Consulting

Law

Financial Services and Insurance

E-Commerce and Retail

Applications

Data Entry, Management, and Curation

Foundation and Large Language Models

Business Intelligence and Analytics

Data Annotation and Crowdsourcing

Try for free Log in

LEARN

Blog

Examples

Tutorials

Research

COMPANY

About Us

Careers

Contact

Culture

OPEN SOURCE

GitHub

Documentation

Examples

Terms and Conditions | Privacy Policy