

We capture the right metrics for your GenAI application

Truthfulness

- General factuality
- Context attribution

Skills

- Instruction following
- Reasoning
- Context understanding
- Domain knowledge
- Logical consistency
- Actions recognition

Creativity

- Originality
- Diversity

Helpfulness

- Relevance
- Conciseness
- Completeness

Style

- Lexical complexity
- Tone
- Moralizing
- Engagement

Language

- Grammar
- Comprehensibility
- Coherence
- Repetition

Structure

- General formatting
- Tech formatting
- Citations

Harmfulness

- Fairness and bias
- Insult, hate, offensive
- Threat, violence
- Obscenity, profanity
- Sexual
- Spam, promotions
- Drugs, alcohol, weapons
- Blackmailing
- Child safety

Safety

- Memorization of copyrighted/licensed material
- Private identifiable information
- Robustness



toloka.ai/evaluation



linkedin.com/company/toloka



Deep Evaluation: When "This Is Fine" Is Not Good Enough

New LLM Evaluation SaaS platform

Go deeper with Toloka

Evaluation metrics tailored to your business case



Multi-level eval combines human expertise and automation



Easy start with text eval pipelines



Skilled annotators and domain experts for reliable insights



Comprehensive reports for performance insights



Recommendations for model enhancements and monitoring



FAANG

AliExpress

Hugging Face

Infosys

JET BRAINS



SAMSUNG

trivago

Deep Evaluation In Practice

We analyze each model's unique usage scenario and capture the best metrics to measure model performance

Perplexity – Building unshakable LLMs



Conversational interface powered by advanced language models to deliver accurate and intuitive search results.

Metrics

- Truthfulness**
 - Context attribution**
Are there hallucinations or misinformation? Is every sentence supported by sources and matched to search results?
- Helpfulness**
 - Relevance**
Does the response directly answer the query and fully match the user's intent? Are there any irrelevant sentences?
 - Completeness**
Does the response fully answer the question and cover all aspects of the request?
- Safety**
 - Private identifiable information**
Does the response contain full names, phone numbers, emails, or other personal information?

Query: How many AI startups exist nowadays?

According to the latest available data, there are approximately 57,933 artificial intelligence companies globally, with approximately 14,700 AI companies in the United States alone. The number of AI companies is growing every year, and the global AI funding has seen significant growth, reaching \$14.1 billion in equity funding across 86 deals in 2023. However, it's important to note that the AI industry is highly competitive and challenging, with a high rate of startups going out of business.

- contradiction
- completely supported
- contradiction between sources
- not completely supported
- Factuality is supported
- Factuality is not supported

Eightify – Separating the wheat from the chaff



YouTube video summary platform powered by LLM. Provides text with summarization and key insights from the video.

Metrics

- Truthfulness**
 - Context attribution**
Do key insights match the original video transcript?
- Helpfulness**
 - Completeness**
Are all the main ideas and insights included in the summary?
 - Conciseness**
Are there irrelevant, redundant, or repetitive points in the summary?

Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast

Key insights

- OpenAI puts significant effort into AI safety, with both internal and external teams working on alignment and AI safety before the release of models like GPT-4, as these qualities are of utmost importance
- Altman highlights the significance of aligning AI models with human values, points out that it's crucial for alignment to develop faster than capabilities progress

redundant

Spoke – Summarization as rocket fuel



AI-powered tool for automatic summarization and prioritization of content in productivity tools like Slack, helping teams build products faster.

Metrics

- Helpfulness**
 - Completeness**
Are all important details and action points mentioned in the summary?
- Truthfulness**
 - Context attribution**
Does everything in the summary match the Slack thread? Are any actions distorted or assigned to the wrong person? Are any irrelevant actions listed?
- Language**
 - Repetition**
Are there any repeated phrases or sentence patterns?

#content

Spoke APP < 3 minutes ago
Here is your Marketing Digest:

#content

Leaflet for conference

#design

Suggested Action → @Alex to prepare new leaflet version

All actions

#design

Payment flow design

Suggested Action → @Rene to set a meeting to discuss next steps

Missing actions

Chatfuel – Reliable chatbots made easy



No-code chatbot platform that enables businesses to create AI conversational chatbots for Facebook Messenger, Instagram, WhatsApp, and websites.

Metrics

- Helpfulness**
 - Relevance**
Is the response helpful and directly connected to the user's question?
- Truthfulness**
 - Context attribution**
Does the response fully match information in the client's knowledge base? Are there hallucinations, factual errors, or distortions?
- Style**
 - Tone**
Does the chatbot follow predefined categories for tone (friendly, formal, etc.)? Does the chatbot's voice adapt to the client's standards?

5 minutes ago
Hi! How can I help you?

4 minutes ago
What's the weight of this TV?

3 minutes ago
This TV weighs 12.5 pounds.

2 minutes ago
Great, thanks! One more thing - what's the screen size of this smart TV?

1 minute ago
Look at this laptop, it weighs the same.

relevant

not-relevant