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

Toloka

linkedin.com/company/toloka

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 · JetBrains · Leroy Merlin · Naver Labs · OLX Group · 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

perplexity

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?

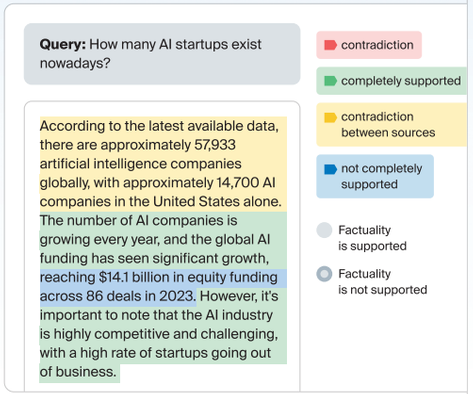
Helpfulness

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?



Eightify — Separating the wheat from the chaff

Eightify

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

Metrics

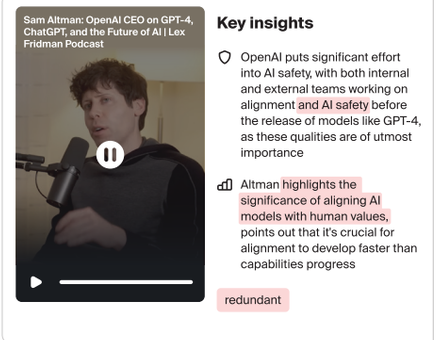
Truthfulness

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?



Spoke — Summarization as rocket fuel

Spoke

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

Metrics

Helpfulness

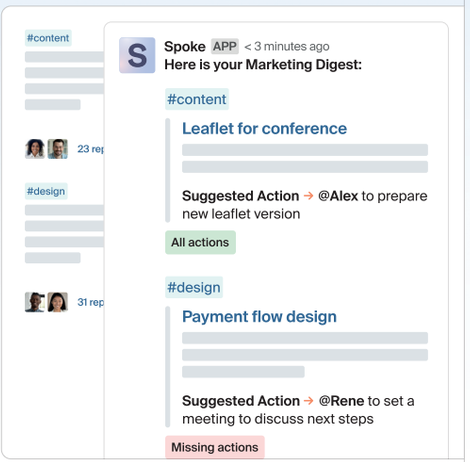
Completeness
Are all important details and action points mentioned in the summary?

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?



Chatfuel — Reliable chatbots made easy

chatfuel

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?