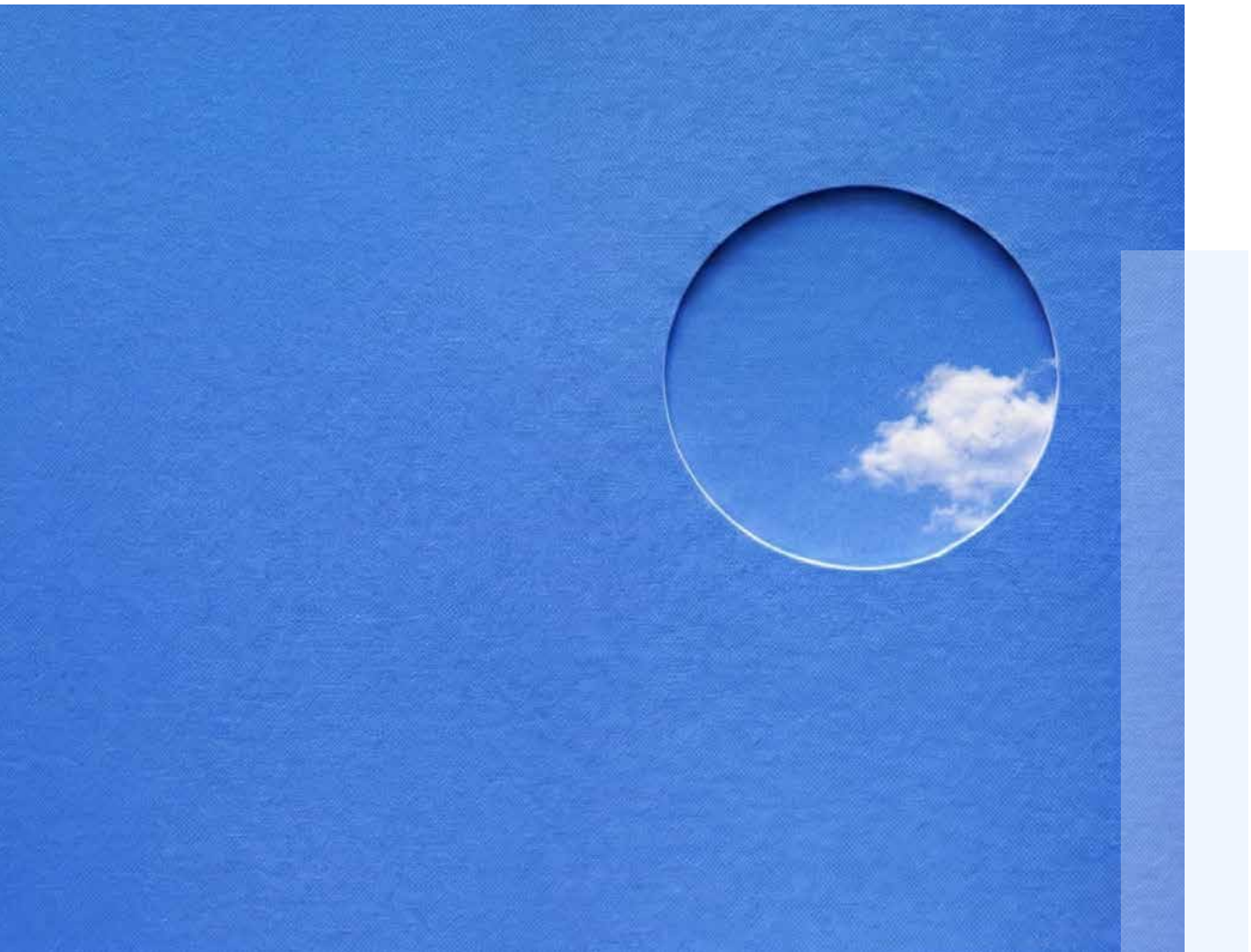
Appen

MULTILINGUAL LLM-AS-A-JUDGE MANAGED SERVICE FOR EVALUATION AT SCALE



Appen's Multilingual LLM-as-a-Judge (LLMaaJ) Service is a fully managed endpoint service for evaluation at scale across locales and use cases. The service follows a two-phase approach: first, calibrating the LLM judge against human-annotated golden sets, then running ongoing quality assurance in production to ensure it stays aligned. Backed by Appen's 30+ years of multilingual expertise across 500+ languages, this service delivers automated evaluation speed with targeted human oversight.

01 Our Core Offering

Instead of deploying teams of human annotators to evaluate every LLM query-response pair, clients send evaluations to an Appen-hosted LLM Judge endpoint and receive structured, rubric-based assessments in seconds. The service is designed around the three core pillars below: Multilingual intelligence, Locale-specific trusted sources, and End-to-end managed service.

A) Multilingual intelligence

Each locale-specific LLMaaJ endpoint deployment is configured to understand the cultural nuances, idiomatic expressions, and figurative language that characterise authentic communication in that market. Research from the MMLU-ProX benchmark which evaluated 36 state-of-the-art LLMs across 29 languages found performance gaps of up to 24.3% between high-resource and low-resource languages (Xuan et al, 2025). Appen's approach directly addresses this gap through locale-aware prompt engineering, model selection and ongoing monitoring and iteration.

B) Locale-specific trusted sources

A distinctive feature of Appen's offering is the integration of locale-specific trusted sources into the LLM judge system. The LLMaaJ endpoint employs tool use with web search to ground its evaluations in authoritative, region-appropriate sources. This is especially important for time-sensitive content such as news, current events, and trending topics, where reference date awareness is essential.

Appen's experience across diverse locales enables us to curate localised source lists for LLM tool use by locale. For example, while ESPN is a go-to sports authority in the United States, an Italian locale judge would reference Gazzetta dello Sport, and a Japanese locale judge might reference Nikkan Sports. These locale-specific sources are curated by human experts to ensure that factuality judgments are meaningful within each market.

C) End-to-end managed service

Under this managed service model, clients receive a turnkey evaluation endpoint with no need to manage prompts, models, or calibration infrastructure internally.

Appen owns the entire LLM judge pipeline covering:

- **Prompt Engineering:** Iterative refinement of system prompts tailored to each locale and evaluation rubric
- **Model Selection:** Choosing and configuring the optimal judge model(s) per use case and language
- **Search Provider Tuning:** Configuring locale-appropriate web search and trusted source feeds
- **Rubric Optimisation:** Adapting evaluation rubrics to reflect locale- and language-specific standards and expectations
- **Ongoing Monitoring:** Continuous tracking of agreement scores, drift detection, and performance metrics, enabled by continual human alignment through QA sampling and confidence based human adjudication

02 Our Two-Phase Approach: From Calibration to Production

A) Calibration

The calibration phase establishes the foundation for reliable automated evaluation. The client provides a golden set of human-annotated samples that represent an expected distribution of query types and domains. Appen's team then undertakes an iterative refinement process through:

01

System Prompt Engineering:

Crafting and iterating on prompts that align the judge's evaluation behaviour with the client's rubric and quality standards

02

Model Testing:

Evaluating candidate judge models against the golden set to identify the best-performing configuration

03

Parameter Tuning:

Adjusting levers such as temperature and output schemas to maximise alignment

04

Agreement Validation:

The calibration process continues until the LLM judge achieves 90%+ agreement with the human-annotated ground truth

B) Production

Once calibrated, the LLMaaJ endpoint moves into production with a robust quality assurance framework designed to maintain alignment over time. This framework covers two core components which are:

01

Ongoing Human QA Sampling:

Weekly human QA sampling verifies continued alignment between the LLM judge and human evaluators. A stratified sample of evaluation instances representing the full distribution of query types and domains is independently scored by human reviewers and compared against the LLM judge's outputs. This structured audit protocol is essential for detecting drift, emerging biases, and edge cases that automated monitoring alone might miss.

02

Confidence-Based Human Adjudication:

Appen has developed a proprietary confidence scoring algorithm that identifies low-confidence judgments where there is uncertainty in the assessment. These uncertain cases are automatically routed to expert human reviewers for adjudication, ensuring that the most ambiguous or challenging evaluations receive the scrutiny they require.

03 Why Partner with Appen

A) 30+ years of multilingual data experience

Appen brings over 30 years of multilingual expertise and a global contributor workforce spanning 500+ languages and 100+ countries to this challenge. This coverage extends beyond high-resource languages to include low-resource languages such as Zulu, Swahili, Catalan, Welsh, and Basque which ensures that LLMaaJ deployments are robust regardless of market.

The company's roots in linguistics run deep: Appen was founded by Dr. Julie Vonwiller, a linguist who built the company to create multilingual data solutions for training machine learning and AI models. This has allowed Appen to build a comprehensive native-speaker contributor network which provides the expertise needed to calibrate and validate LLM judges.

B) Hybrid approach combining automated LLM-based evaluation with human review

Appen's LLMaaJ Managed Service delivers the speed and scalability of automated evaluation with the precision and cultural sensitivity of human expertise. The confidence scoring mechanism ensures that straightforward evaluations are handled instantly by the LLM judge, while genuinely difficult cases receive expert human attention.

Under this managed service model, expert human review is applied in targeted, high-impact areas and edge cases. For example, human experts handle novel or ambiguous cases that fall outside the judge's established rubric, feeding insights back into the system to improve future performance.

C) Equipped for time-sensitive content

Appen's LLMaaJ service incorporates tool use with web search, enabling the judge to evaluate time-sensitive content such as news, current events, and trending topics with full reference date awareness. The judge can verify whether a model's response accurately reflects the state of the world at the relevant point in time, drawing on locale-specific trusted sources to validate claims.

Talk to an Expert

Learn how Appen's Multilingual LLMaaJ Managed Service can deliver reliable, locale-calibrated evaluation at production scale—combining automated speed with the cultural precision that multilingual markets demand.

appen.com | sales@appen.com