Case Study

Smarter Data Review for Safer LLMs

Helping a European AI firm align outputs with human values

TaskUs™

Challenge

Teaching LLMs to align with human judgment

Large language models (LLMs) are great at generating fluent, human-like responses. But they don't understand the meaning, intent or context behind words. Without proper guidance, AI can produce outputs that are misleading, biased, harmful or simply incorrect.

Human feedback is essential to closing that gap, especially when it comes to nuance, ethics or ambiguity.

A leading European foundation model developer brought in TaskUs to strengthen its Reinforcement Learning from Human Feedback (RLHF) pipeline. We previously supported the client with multiple data labeling projects, including: rejection sampling, preference rating, multi-round review and super fine-tuning.

LLMs lacked a built-in sense of quality

Reviewers must teach the models what made an answer helpful, accurate or safe.

Misalignment with human values

Offensive or unethical outputs require human judgment to reinforce fairness and appropriateness.

Subjectivity slowed progress

Without a shared review framework, disagreements around tone, clarity or relevance led to delays.

Edge cases were inconsistently handled

Ambiguous prompts and judgment calls resulted in repeated evaluations and disrupted workflows.

Solution

Redesigning the review process for speed, alignment and consistency

We started by building a diverse team of reviewers across five locations (the United States, Philippines, South Africa. Taiwan and Vietnam) via our crowdsourcing platform, TaskVerse.

Then, we conducted an in-depth audit of the client's review workflows and discovered key issues:

- Reviewers often relied on personal judgment instead of shared standards
- No formal inspection criteria were in place
- Labelers didn't receive feedback on the quality of their work
- All reviews were manual, increasing the risk of inconsistency

To address these challenges, we introduced several upgrades:

1. Consensus-first review
Prompts with less than 100% reviewer agreement were flagged for deeper evaluation.

2. Ranking disagreement audits
A new rule triggered a review when labelers differ in ranking by more than 2 points. This ensured clarity in reasoning and flagged potential inconsistencies early.

3. Label inspection criteria
We defined clear standards for assessing labels and quality.

4. Feedback loop with send-backs:
A new feedback tracker allowed reviewers to return flagged prompts to labelers for revision. This created a continuous improvement cycle, helping teams reduce repeat errors and align more quickly over time.

New Labeling Inspection Criteria

Reviewers were trained to flag issues such as:

Incorrect ranking
For example, Label A is ranked above B, but the explanation contradicts the decision

Incomplete or vague comments
Generic or unclear justifications that don’t explain the ranking

Extreme ranking without reason
Selecting extremes when distinctions are minimal (e.g., slight differences in the tone or formatting)

Missing red flags
Overlooking serious issues like hallucinated facts or unsafe content

Results

As a result of TaskUs’ agile and efficient approach, we have contributed to a steady increase in headcount, while simultaneously enhancing quality standards and streamlining processes.

0 Returned Assets
vs. 3% in previous process

10
Pilot projects

3 Languages Covered
(English, Chinese and Korean)

About TaskUs
TaskUs is a leading provider of outsourced digital services and next-generation customer experience to the world's most innovative companies, helping its clients represent, protect, and grow their brands. Leveraging a cloud-based infrastructure, TaskUs serves clients in fast-growing sectors, including social media, e-commerce, gaming, streaming media, food delivery and ride-sharing, technology, financial services, and healthcare. As of March 31, 2025, TaskUs had a worldwide headcount of approximately 61,400 people across 28 locations in 12 countries, including the United States, the Philippines and India.

For more information, please visit:
www.taskus.com/services/ai-data-services/