CASE STUDY:

Increasing Accuracy and Clarity in Multi-Modal Captioning

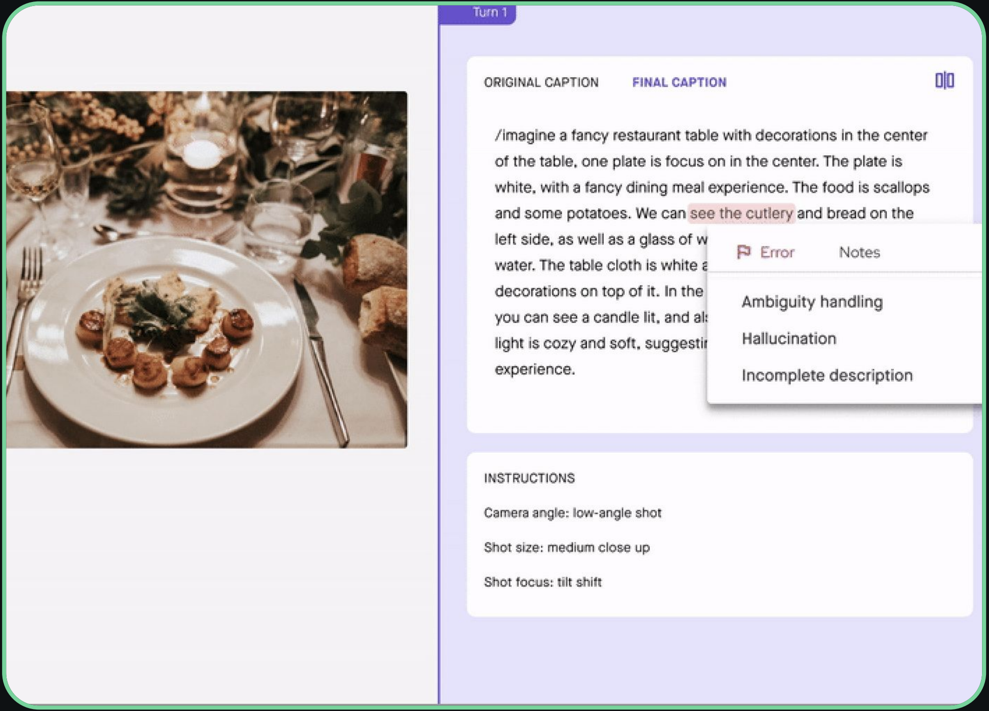
sama

Challenge:

A pioneer in multi-modal metadata search needed a partner to help them fine tune their model to improve natural language processing.

With a growing list of media and entertainment clients, the company wanted to increase the quality of their captions for video frames and stills but lacked the manpower to do prompt and response evaluation at scale.

Relying on full automation to write captions did not yield accurate results.



/imagine a fancy restaurant table with decorations in the center
of the table, one plate is focus on in the center. The plate is
white, with a fancy dining meal experience. The food is scallops
and some potatoes. We can see the cutlery and bread on the
left side, as well as a glass of w
water. The table cloth is white a
decorations on top of it. In the
you can see a candle lit, and al
light is cozy and soft, suggestin
experience.

INSTRUCTIONS

Camera angle: low-angle shot

Shot size: medium close up

Shot focus: tilt shift

Not only did automation lack the nuance required to translate highly subjective inputs into concise and accurate captions, but fully automating the process led to errors and hallucinations in the model.

Raising the bar on scalable accuracy and consistency

With the goal of improving and scaling their caption writing, the company turned to Sama's dedicated, in-house data experts, who underwent client-specific training before tackling the company's data.

To ensure quality and consistency, the Sama team kicked off the project by working with the client to define guidelines and establish context, for example how much detail to capture for salient objects, what's important for the annotator to include about the background, how to handle well known brands and logos, and what are the end use cases for the customer's clients.

After aligning on how to calibrate and ground the visual inputs, data experts developed a quality rubric and defined penalties and scoring. They also identified gold tasks where generated captions met the "gold standard."

This not only ensured consistent outputs and quality checks but also accelerated the training process.

Improving models with human expertise

With Sama's human-in-the-loop (HITL) approach, the company was able to get the best of both worlds — automation and standardization paired with real human oversight.

Integrating information from multiple modalities also presented enormous challenges. Having a HITL was critical to ensure the generated captions accurately reflected the timing of events. When handling asynchronous information, like in a scene where a sound precedes the on-screen visual, the client's model alone lacked the context to generate the correct caption, while the Sama team was able to correctly link the elements from different modalities.

> "Multi-modal captions are highly subjective and present many visual challenges. For this project, our data experts had to take many factors into account, whether it was poor lighting, overlay text, mirrors, windows, or even partially hidden text."
>
> **– Abha Laddah**
>
> Senior Director, Solutions and Launch at Sama

Results

By including human feedback loops in the process, Sama helped the company accelerate the fine tuning process and increase the accuracy of their foundation model.

With Sama, the client saw:

Reduced errors and hallucinations

Example: Retraining the model when the caption incorrectly identified a facial expression

Clearer and more concise writing

Example: Tweaking instructions when non-salient objects were being included in the description

Improved context and tone

Example: Rewriting captions to improve vague or uncertain language

Unlock the Full Potential of Your Gen AI Models

Sama is a global leader in **data annotation, supervised fine-tuning, and model evaluation** for computer vision and generative AI applications.

As a **recognized diverse supplier**, our proprietary human-in-the-loop approach, scalable platform and in-house team of over 5,000 data experts drive data-rich model improvements & RAG embedding enhancements that help get AI & ML models into production up to 3x faster.

Email sales@sama.com or visit www.sama.com

sama

Certified B Corporation