Agile Deliberation explained: building classifiers for subjective categories with an LLM loop

23 November 2026 · 6 min read · blogs.iictp.org/agile-deliberation-subjective-image-classifiers-explained

Short answer: Agile Deliberation is a Google Research framework, described on 4 November 2026, for building image classifiers when the category itself is fuzzy, such as “healthy food” or a content-moderation policy. It guides the user through scoping sub-categories and resolving borderline images while an LLM-driven loop rewrites the classifier prompt. It matters because it treats the human’s definition as something that changes during labelling, not as a fixed ground truth.

TL;DR The short version

- Agile Deliberation is a Google Research framework for building image classifiers when the category itself is fuzzy, such as 'healthy food'.
- The user scopes sub-concepts and exclusions, then resolves borderline images while an LLM loop rewrites the classifier prompt.
- It was tested in only 18 sessions on two domains, but the lesson is that policy text and examples should evolve together.

Read online, listen to it, or change the text size: https://blogs.iictp.org/agile-deliberation-subjective-image-classifiers-explained

What is Agile Deliberation?

The authors, Otilia Stretcu and Leijie Wang of Google Research, start from an assumption in ordinary supervised learning: the labeller already has a clear, stable definition in mind. For subjective concepts that is rarely true. People discover the edges of a concept only when they meet examples. The framework turns the creation of a classifier into an iterative conversation, and code is open-sourced on GitHub according to the post.

How it works

- **Stage 1, concept scoping.** The user breaks a broad concept into in-scope sub-concepts and explicit exclusions. For “healthy food” that might be salads and grain bowls in, processed foods out.
- **Stage 2, concept iteration.** Vision-language models mine borderline images that challenge the current text definition, for instance a fruit smoothie. The user decides which side it falls on, and that feedback flows into an automatic prompt-optimisation loop that updates the classifier.
- **Not classic active learning.** Active learning picks examples to reduce model uncertainty. Here the examples are chosen to expose the human’s own ambiguity.

What is new compared with earlier approaches

The post compares against three baselines: zero-shot prompting, an automated “modeling collaborator” that decomposes the concept without the user, and manual deliberation in spreadsheets with hand-written prompts. The difference is that the human stays in the loop but is asked about concrete images, not asked to engineer prompts.

Results reported by the source

- The evaluation used 18 live sessions of 1.5 hours with human participants, on two subjective domains: “Healthy Food” and “Pay-to-Play” clickbait images for content moderation.
- Google reports higher labelling consistency, classifiers that outperform the baselines (largest gains on the hardest subjective images), and statistically significant reductions in effort and stress against manual prompt engineering.
- It also says every participant preferred Agile Deliberation to the manual methods. The post gives these as findings; exact effect sizes sit in the paper, which we did not read in full.

Limitations and open questions

- A sample of 18 sessions is small, and the post reports two domains only.
- Results depend on the underlying vision-language model, and on the user being willing to reflect on edge cases.
- We read the blog post, not the paper or the code.

How to adopt the idea

1. Write the concept as sub-concepts and exclusions before you label anything.
2. Retrieve images near the decision boundary, not random ones.
3. Record each decision with a short reason and fold it back into the policy text.
4. Re-test on the cases you already fixed so a new rule does not break an old one.

What it means for trust and safety and data teams

If you maintain moderation or curation policies, the useful takeaway is that policy text and examples should co-evolve. See prompt engineering and context engineering and responsible AI in practice.

Frequently asked questions

- **What problem does Agile Deliberation solve?** Defining classifiers for categories that people do not fully define until they see examples.
- **Does it retrain a model?** The post describes an automatic prompt-refinement loop around vision-language models, not fine-tuning.
- **Is the code available?** Google says it open-sourced the algorithms on GitHub.

Related reading

- Prompt engineering and context engineering

- Multimodal AI explained
- How to evaluate an LLM application
- AI Practitioner

Sources and currency

Sources were read on 23 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Otilia Stretcu and Leijie Wang, Google Research. Agile Deliberation: How AI helps humans define the gray areas. 4 November 2026. research.google/blog/agile-deliberation-how-ai-helps-humans-define-the-gray-areas.

Published by IICTP, an independent technical credential body, on 23 November 2026. The current version of this post is at https://blogs.iictp.org/agile-deliberation-subjective-image-classifiers-explained. More guides: blogs.iictp.org.