Notes on Achieving 10,000x training data reduction with high-fidelity labels (for better fine-tuning)

by Markus Krause et al, Google Ads; foreword by Charles H. Marin, 8/7/2025

# Foreword

Google introduces a new active learning pipeline designed to dramatically reduce the quantity of training data needed for fine-tuning large language models (LLMs), especially in tasks like classifying unsafe or clickbait ads—where high-quality labels are hard and costly to generate

• 1. Start with a zero- or few-shot LLM ("LLM-0"), prompted to classify content (e.g., “Is this ad clickbait?”).

• 2. Generate an initial labeled dataset, though it's often imbalanced and noisy due to low precision.

• 3. Cluster examples labeled as clickbait and benign. Identify clusters that overlap, signaling examples the LLM finds confusing.

• 4. Sample pairs of examples from those overlapping clusters that are close in feature space but differ in labels. Send these to expert annotators for high-fidelity labels.

• 5. Split the expert-labeled data: One part for evaluation, tracking how well model aligns with human experts (using metrics like Cohen’s Kappa). The other part for fine-tuning the model.

• 6. Iterate: Repeat clustering and sampling over several rounds until model–human alignment plateaus or matches expert–expert alignment.

By honing in on the most informative examples—those around the decision boundary—you can train better-performing and more aligned models with far less data.

# References

[1] [Achieving 10,000x training data reduction with high-fidelity labels, by Markus Krause et al, Google Ads, August 7, 2025, Google blog](https://research.google/blog/achieving-10000x-training-data-reduction-with-high-fidelity-labels/)

[2] [Concept drift, Wikipedia](https://en.wikipedia.org/wiki/Concept_drift)

[3] [Active Learning, Wikipedia](https://en.wikipedia.org/wiki/Active_learning_(machine_learning))

[4] [Cohen’s kappa, Wikipedia](https://en.wikipedia.org/wiki/Cohen%27s_kappa)