AnnotateThis: Training research assistants to code subjective concepts with LLMs - a case study with climate change pessimism

Large language models (LLMs) are increasingly used for data annotation, but there is skepticism about their ability to label subjective concepts in social science work. Research assistants (RAs) commonly label data, yet little work examines RA-LLM interactions in concept iteration and annotation. We present **AnnotateThis**, a system co-designed with non-technical social scientists, that supports research assistants (RAs) in iteratively instructing LLMs to detect subjective concepts.

AnnotateThis is designed to overcome several challenges with existing approaches. First, we specifically design for the social science process - which is iterative, reflective, and deliberative. Second, we provide domain-specific expert scaffolding (information tools designed to nudge participants towards expert judgement) to overcome the problem of over-reliance on LLMs which has been well documented in the machine learning literature [citation]. We show that with our approach, participants generate much better prompts than automated prompt improvement methods - highlighting the importance of human-centered design in human-AI collaborative tools.

AnnotateThis provides 8 information features from per-instance labels, consistency across model runs to natural language explanations, and visualizations such as uncertainty trends across runs to help participants assess reliability signals and refine instructions. Our implementation uses GPT-4o-mini (2024-07-18) to produce outputs given participants' prompts; this allows participants to observe and reflect on how an LLM responds to their instructions.

We study five questions: **RQ1**: Can participants use AnnotateThis to improve their instructions to LLMs?, **RQ2**: When supplied with the prompts produced through AnnotateThis can LLMs annotate data as well as participants?, **RQ3**: Will the expert scaffolding in AnnotateThisES produce more alignment both between participants and experts, and between LLMs and experts? **RQ4**: When supplied with the prompts produced through AnnotateThis can LLMs annotate data as well as better than when supplied with prompts generated by fully automated methods?

We answer these questions through evaluating AnnotateThis with two user studies on detecting climate change mitigation pessimism in social media posts, the first provides more space for user interpretation and the second more instruction from domain experts. Answering **RQ1**, we find that participants successfully improved their LLM instructions through iteration: Accuracy and F1 scores increased by **+0.137** and **+0.106** in Study 1 and **+0.185** and **+0.108** in Study 2. **For RQ2**, when LLMs were prompted with these refined instructions, performance measured against expert annotations improved significantly with expert scaffolding: F1 increased by **+0.082**, and accuracy by **+0.117**. **RQ3** reveals that with more instruction, users aligned better with expert judgment: without scaffolding, participants' own labels showed a slight divergence from experts (Accuracy change: **-0.020**, F1 change: **−0.034**), but with scaffolding, alignment improved (+**0.036**, **+0.019**). We also compare participant-crafted prompts with those from the state-of-the-art Automatic Prompt Engineer (APE) **(RQ4)**. Participant prompts were competitive or better: in Study 2, they achieved an F1 gain of **+0.147** over APE, with minimal recall loss (**−0.011**) and a notable accuracy improvement (**+0.234**). These findings show that AnnotateThis helps RAs externalize and refine decision criteria, while expert scaffolding stabilizes those criteria, enabling stronger expert-aligned annotation on subjective social-scientific constructs.