Title: Notes on Annotation appendix for "How alt-tech users evaluate search engines: Cause-advancing audits"
Authors: Evan M. Williams (1), Kathleen M. Carley (1)
Date: February 11th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix: Notes on annotation

Prompt creation was an iterative process, where we iteratively added example comments and procedures to address GPT-4o-mini (GPT-4) errors on a sample of unseen comments. We first included several offensive comments containing hate-speech, to ensure that GPT4 would be able to correctly annotate them. To ensure that our few-shot examples were not biasing results, we tried annotating all data with two alternative prompts that had some examples swapped out for different ones. In the second prompt, we added two additional annotated comments that contain edge cases where we saw the model struggle—ties and responding to "green text."

In annotations 1 and 2, we noticed that GPT4 often failed to map "Jewgle" to "Google," so we created a third prompt where we both swapped out several of the original annotated comments and standardized engine names. In all three prompts, Yandex was by far the top-ranked search engine (min = 321, max = 331). The other top seven rankings were largely the same, but in the second prompt, Searx and Startpage had swapped rankings. All prompts and annotations are released in supplementary materials. As both annotations for prompts 2 and 3 had lower Jaccard similarity inter-annotator agreement for first-place rankings, we elected to base findings on the annotations for our original prompt. Much of the disagreement can be traced back to inherent ambiguity in the nature of the task and comments.

Extracting relative preference ranks was a challenging task both because comments could contain substantial ambiguity in relative rankings and because we were discretizing a continuous space. On the former point, for example, some comments recommend using different search engines for different functions or topics (e.g., using Bing for image search but Searx for web queries). Additionally, as we were extracting relative ranks within a comment, what "worst" means could be ambiguous. For example, if a user were to recommend Yandex over Bing, whether that user says Bing is "okay" or "the worst," Bing still receives a relative rank of 2 under our annotation schema.

This annotation approach also presented a challenge in terms of calculating inter-annotator agreement, as a normal engine ranking for a single comment could look something like: {"yandex":1, "searx":1, "bing":2, "startpage":2, "google":3}. This prohibits the use of standard inter-annotator agreement metrics, and even indeterminate ranked list comparison metrics like Rank-Biased Overlap fail as ranks are shared by elements. Our interest is in relative rankings, and there can be disagreement about both rank and whether an engine is being compared. As our primary research question concerns which engine is most promoted, we elected to evaluate inter-annotator agreement using the Jaccard similarity of engines each annotator assigned a rank of 1, i.e., the set of engines users most strongly recommended in each comment. The Jaccard similarity between two sets is simply the cardinality of the intersection divided by the cardinality of the union. We additionally report the precision and recall of GPT-4 on query identification (Task 3) while treating annotator 1 as ground truth. Precision and recall are defined as:

$$Precision = \frac{\#\ True\ Positives}{\#True\ Positives + \#True\ Negatives}$$

$$Recall = \frac{\#\ True\ Positives}{\#True\ Positives + \#False\ Negatives}$$