STEP 1 Initial Prompt Creation

Chat GPT

Prompt Used: Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology, an abstract, summarize key findings, identify trends and gaps, create one graph, and one testable hypothesis. Use a sophisticated academic tone and make all sources peer reviewed.

Structure: The article did include a methodology section and used a systematic format across the article. The steps of the methodology section include a search strategy across PsycINFO, Web of Science, Scopus, and PubMed. Then, inclusion and exclusion criteria and a selection process yielding 126 studies after duplication removal and inclusion criteria. The last step in the methodology was listed as data extraction and synthesis.

Synthesis: There were key findings and limitations/ concerns listed but they did not do as good of a job as I had hoped. The paragraph didn't explain the key findings but only reiterated them.

Trends and Gaps: Trends and gaps were listed but not well explained. At most a sentence max on the limitation of the research with nothing else. The trends and gaps did not seem meaningful, they were generic conclusions that could be put towards a multitude of research articles including "machine learning has limited psychological insight" or "clinical studies often lack large, diverse samples".

Hypothesis: Multimodal machine learning models integrating linguistic, behavioral (e.g., smartphone activity), and physiological data will predict clinically diagnosed major depressive disorder with significantly greater accuracy (AUC ≥ .85) than unimodal linguistic-only models (AUC ≤ .75), controlling for demographic variables. This hypothesis can be tested using a longitudinal clinical dataset with cross-validated model comparison and external replication.

> This hypothesis is testable and feasible. I like the approach of a longitudinal dataset and replication to achieve the best results.

References: All the citations were real, accurate, and reputable journal publishers.

Overall, the article gave more of a bullet pointed list of everything I wanted included instead of writing it as an academic essay but did provide accurate citations from reputable sources.

CoPilot

Prompt Used: Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology, an abstract, summarize key findings, identify trends and gaps, create one graph, and one testable hypothesis. Use a sophisticated academic tone and make all sources peer reviewed.

Structure: The article included a methodology section and systematic format using a search strategy, inclusion criteria, exclusion criteria, screening and selection, and data extraction. The article used 112 studies, but this was not listed in the methods section, it was listed in the abstract.

Synthesis: There was a specific findings section of the article, unlike ChatGPT which had multiple paragraphs of information on machine learning and psychology but didn't list a finding section. The key findings listed were explained better and there were many compared to ChatGPT.

Trends and Gaps: The gaps listed in the research were unique. The article discussed the overreliance on reddit or twitter to collect their sample which I have not heard before in an article. The article also discusses ethical concerns of collecting the research online and I found it admirable that Copilot would think of the ethics of research without specifically being asked.

Hypothesis: Multimodal machine learning models that integrate linguistic, sensor-based, and behavioral data will significantly outperform unimodal models in predicting clinically validated mental health outcomes, even after controlling for demographic variables.

> The hypothesis is testable and relevant. It discussed multimodal vs unimodal machine learning and how machine learning is related to behavior data collection. ChatGPT was more descriptive with AUC accuracy and how the data would be collected, but both AI models provided a testable hypothesis.

References: There were no references listed or cited in the article.

I asked Copilot to create a graph, and it needed confirmation of what dataset I wanted to use before creating the graph. I had to adapt the prompt so that copilot would create the graph I wanted. I added the phrase "create one graph of publication trends over time" which was the graph copilot suggested creating.

Grok 3

Prompt Used: Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology, an abstract, summarize key findings, identify trends and gaps, create one graph, and one testable hypothesis. Use a sophisticated academic tone and make all sources peer reviewed.

Structure: The article included a methodology section including inclusion and exclusion criteria. The methodology was not very good at explaining what the article wanted to focus on or listing out the steps for replicability. There were only 15 articles analyzed and I received that information from the abstract; it was not listed in the methodology or anywhere else in the paper.

Synthesis: The key finds were present and well explained but they were not very relevant. The article discussed 4 different findings as if it was writing 4 different papers. Not once in the findings section did Grok relate them to the current topic.

Trends and Gaps: There were trends and gaps written out and they were related to the main theme of the paper. They were not unique, simple statements about bias and small samples being challenging. The article did talk about the ethical components of machine learning and psychology though.

Hypothesis: Among individuals with bipolar I disorder, machine learning models trained on event-related potentials from a go/no-go task will demonstrate superior predictive performance for suicide attempt history compared to models using solely demographic variables (age, gender, education), as measured by an area under the curve (AUC) exceeding 0.80 and sensitivity above 80%, tested via leave-one-out cross-validation on a sample of at least 50 participants (approximately 80 words).

> The hypothesis is testable and somewhat relevant to the prompt. It focuses on bipolar I but still uses machine learning models and predicting their performance in determining suicide attempts in bipolar I participants. The sample size is small of only 50 people.

References: There were no references listed but there were citations used in the article. The citations were real, however.

STEP 2 Analyze Model Responses

Strengths and Weaknesses between AI models:

ChatGPT was the only model with a reference section and scoured the most articles to find reliable information. ChatGPT had a hypothesis that was more in depth and specific. The model was lackluster in explaining the information it was giving and provided reiterations of the same information.

CoPilot had the most specific findings and listed them out as well as an explanation. The gaps in research were also the most unique citing real sources like Twitter being overused in collecting data. The hypothesis could have been more specific and explained how the study could be replicated but it was still descriptive enough that I understood what it wanted to do. There was also no reference page listed which really affected the credibility for me.

Grok 3 has the better structure because it was structured like a scientific paper instead of bullet points, but it was overall the worst model. Only 15 journal articles were analyzed, the key findings were present but not tied together, and the gaps were not unique. The hypothesis was specific towards a subgroup of psychology which would be beneficial for a study, but I am looking more towards machine learning and social data, not bipolar I. Lastly, it did not provide a references page which CoPilot also had an issue with.

STEP 3 Refine the Prompt

Tailored prompts:

ChatGPT

Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology, an abstract, summarize and explain key findings, identify trends and gaps thoroughly, create one graph, and one testable hypothesis. Use a sophisticated academic tone and make all sources peer reviewed. Use paragraph formatting with minimal bullet points.

> After revising the prompt, ChatGPT created an article that used 127 articles, provided a more detailed methodology, added a screening process and explained where they got the participant data from, added another paragraph explaining the key findings, had better organization throughout, and implemented ethical considerations. The hypothesis became more specific towards using linguistic features in machine learning to discover behavioral patterns and kept the same notion of using longitudinal data and AUC scores. The reference page was shorter but more tailored to the article.

CoPilot

Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology including the number of articles analyzed, an abstract under 300 words, summarize key findings, identify trends and gaps, create one graph, and one testable hypothesis with specific ways of scoring or testing the hypothesis. Use a sophisticated academic tone and make all sources peer reviewed with a reference section. Refrain from using bullet points, instead use paragraph form.

> After revising the prompt, CoPilot created a paper with 112 articles analyzed. The introduction was related to the article and provided insight into machine learning and data mining, and the methodology did not change much, it still focused on exclusion and inclusion criteria. The findings were well thought out and in paragraph format, contrary to what the previous prompt gave. There was also a specific trend and gap section that explained the challenges of getting reliable longitudinal data. There was also an added reference section with reliable sources.

Grok 3

Conduct a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology that explains replicability and the focus, an abstract under 300 words, summarizes key findings with explanations and relate them to the main topic, identify trends and gaps with ethical consideration, create one graph, and one testable hypothesis related to machine learning. Use a sophisticated academic tone and make all sources peer reviewed with a reference section.

> After revising the prompt, Grok 3 created an article that did not list how many articles it analyzed but did expand the years where it looked for articles. It summarized key findings and added AUC scores and mean accuracy for specific studies. It also provided in text citations, which were lacking in the previous prompt exercise. Grok created more unique trends and gaps with some focusing on the overreliance on certain data resulting in overfitting and the ethical consideration of research. The hypothesis was refined and focused on a specific age cohort, a specific machine learning model, and incorporated AUC scores. A references page was added with reliable sources.

STEP 4 Cross-Model Collaboration

The best elements of each model would be the

- Number of articles analyzed by ChatGPT
- Methodology from ChatGPT
- Findings of CoPilot and Grok
- Trend and Gap section of CoPilot
- Hypothesis of Grok
- Reference page of ChatGPT

The new prompt will be used in ChatGPT

New Prompt: Using these drafts from three AI models, combine the strongest aspects to produce a 2,000-word structured, systematic literature review on the use of data mining and machine learning in psychology or social data analytics. Include a methodology that explains replicability and the focus including the number of articles analyzed, an abstract under 300 words, summarizes key findings with explanations and relate them to the main topic, identify trends and gaps with ethical consideration thoroughly, create one graph, and one testable hypothesis with specific ways of scoring or testing the hypothesis related to machine learning. Use a sophisticated academic tone and make all sources peer reviewed with a reference section. Use paragraph formatting with minimal bullet points.

> I combined the strongest aspects of each AI model and of the prompts that resulted in the greatest change. I think the output fits exactly what I was looking for in an article about machine learning and psychology. There were aspects of each AI model that I did not particularly like, but with the prompt revision tailored to each model, I believe I was able to pull the best portion of each article to create a paper that best fits with the topic, descriptions, level of specificity, and references I would use when writing a paper.

STEP 5 Reflection

Each model took their own approach to the systematic review. Chat GPT focused on listing out the information simply, so it was easier to digest, CoPilot focused on the trends and gaps and how they relate to the topic, while Grok focused on a more niche part of psychology and machine learning instead of the overarching view of ML and psychology. Grok had the best prompt refinement, but I also think that is because it needed the most work from the beginning. Before the prompt refinement, there was little to nothing I liked about the review. It was scattered, too specific in some areas and not in depth enough in others, but with the refinement I was able to list every part that I wanted changed and it made improvements accordingly. AI can be helpful tool with the right manipulation and many drafts.

The first prompt attempt did not yield good results from any of the models but after the revision, there were aspects in all the reviews that I would find in my own writing style. It was academically sound but still easy to understand. There were descriptions and explanations where necessary and the references used were authentic and peer reviewed. I learned that the more specific you are with AI models, the more information and the more tailored their responses will be. If you are broad and looking to find quick answers without understanding what you're asking for, AI will not be much help. It is a tool, not an assistant.

Overall, I think I liked ChatGPT's review the most because of how well it combines all the aspects that I had to change with the other models. It consistently used in text citations and references, it described the hypothesis and research replicability, provided AUC scores from the beginning, and incorporated an introduction that gave a background to the review. Even though Grok had the best article at the end, ChatGPT was the most useful model when it came to providing me with what I prompted. But once again, AI tools can only bring you so far, you have to do the reading and work yourself to really understand what you want out of a tool; and AI cannot become a reliance. AI cannot write papers for you or be your final draft, authenticity still needs to be shown in the work that is completed.