Claude 4 Opus chat on: AbsJud - Sect 4 - Google

User:

AbsenceJudgement /

User:

metadata: BB

Project Context: Writing Section 4 "Case Studies of AI Research Tools" for an academic paper examining AI research tools' lack of judgment through the lens of "technoscholasticism." The paper argues that LLMs privilege textual authority over critical assessment, lacking three key judgment dimensions: epistemic humility, inductive capacity, and correspondence with reality. Current Task: Complete Section 4 using a structured format where each tool section contains: 1. Brief introduction with marketing claims 2. Comprehensive table showing performance across research stages 3. Detailed subsections (4-5 paragraphs each) for each research stage, providing concrete examples of failures and limited successes, using the outline in comments at the beginning of each subsection. Writing Requirements: * Academic clarity with Australian spelling * LaTeX formatting in artefacts * Active voice, varied sentence structure * Avoid em-dashes, excessive hedging * Concrete examples over abstractions * Each claim supported by specific evidence * Paragraphs of 100-150 words typically Sections to Complete: * 4.3 Google's Research Tools (Gemini Pro 2.5) * 4.4 Hugging Face's Open Deep Research * 4.5 Other Tools (Operator, Computer Use, Elicit, Perplexity) Key Testing Details: * Used FAIMS as control (authors' own tool) * Prompt: Comprehensive DH analysis as of 2025 * Success metric: Mundane utility despite judgment failures * All tools failed temporality/reality correspondence tests Expected Detail Level: See completed OpenAI section - each subsection needs specific examples, failure analysis, and implications. Tables should capture research stage, tool, task, failure mode, and judgment type affected.

Give me a readback and functional decomposition of the task ahead of us, please.

Assistant:

Readback and Functional Decomposition

Task Understanding

I need to complete three sections of your academic paper examining AI research tools through the "technoscholasticism" framework. Each section will analyse how specific tools fail to demonstrate genuine judgment while potentially offering mundane utility.

Functional Decomposition

1. Section 4.3: Google's Research Tools (Gemini Pro 2.5)

- **Introduction paragraph**: Google's positioning claims vs reality
- **Performance table**: Research stages × failure modes × judgment types
- **Detailed subsections** (4-5 paragraphs each):
  - Ideation and Research Design
  - Literature Discovery and Source Evaluation
  - Data Collection and Extraction
  - Synthesis and Writing Support

- Implications for Research Practice

2. Section 4.4: Hugging Face's Open Deep Research

- **Introduction paragraph**: Open-source approach and promises
- **Performance table**: Limited testing scope due to cost
- **Detailed subsections** (likely 3-4 paragraphs given limited testing):
  - Technical Architecture and Approach
  - Data Enhancement Capabilities
  - Cost-Benefit Analysis
  - Implications for Open Research Tools

3. Section 4.5: Other Tools

- **Introduction paragraph**: Why these are "other" category
- **Brief analysis per tool** (1-2 paragraphs each):
  - OpenAI's Operator
  - Anthropic's Computer Use
  - Elicit
  - Perplexity
- **Comparative synthesis paragraph**

Key Evidence to Incorporate

- Gemini's "shotgun mass-summary" approach
- Failure to use Google Scholar despite access
- $200 cost for single Hugging Face run
- Elicit's dependency on Semantic Scholar
- Perplexity's claim of 25 sources while citing 3
- All tools' consistent temporality blindness

Writing Approach

- Match the depth and style of completed OpenAI section
- Use specific DH testing examples throughout
- Emphasise judgment failures while acknowledging utility
- Australian spelling, active voice, concrete examples
- Tables in LaTeX format capturing failure patterns

Shall I proceed with drafting Section 4.3 on Google's Research Tools?

User:

metadata: BB

2-3 paragraphs per subsection (we need to say what we need to say, but should not pad. If there is less to say, that's fine. We will also be starting to skip subsections as we move to less useful tools). Before we continue though, please scan transcripts.tex for where we discuss our use and opinions of Google. That should also strongly inform this section.

Thinking about a prior run of this, I think we need to figure out what specific points we're making in each section.

Some of the notes I had:

Ok, some redundencies. We already talk about Carnap in OpenAI, so we'll need to highlight how it's different or move on there. DO NOT USE em-dashes. Keep sentences simple where possible.

For swapping pronouns, we're going to need to work on that, since it's unclear. That quote isn't illuminating.

No, I think we'll need to tear this out and think about it again.

Bleh.

The main points I want to make are:

1. **when doing "Research" Gemini (the service) is very much an model running in a loop calling tools. Specifically, looking at the "thinking" mode output of Gemini (the service)'s "Deep Research" mode, it has thought thought thought, then a hundred plus sources, then repeats.**

i.e in the middle, it has a thought: "My next steps involve a more systematic approach to understanding the academic landscape by exploring a directory of DH centers. I also plan to analyze a key chapter discussing the future of DH research to understand the evolving relationship between computer science and the humanities, the role of infrastructure, and the major challenges and opportunities ahead for the field." Which ... a) conflates way too much in a single task, and b) isn't at all systematic. It pulls up 12 websites from a search, and then goes: "I've made progress in accessing some key resources for understanding the current state of Digital Humanities. I successfully analyzed a chapter from 'Debates in the Digital Humanities' that highlighted the ongoing need for true collaboration between computer scientists and humanities scholars, the importance of robust infrastructure, and the potential for DH to drive new research questions. I also gained valuable insights from the ACD-Engine project's JupyterBook, which focuses on improving cultural data in Australia and has implications for infrastructure and skills development in the region. "

Honestly it's just gaslighting itself. There's no way to "successfully analyse a chapter" by... downloading a snippet. This is the source soup of undergrads using Google. But, nowhere in this log is anything *critical* of what it found. Nor is it double-checking anything.

This is both a critique of tool design as well as the internal prompting of the "Deep Research" infrastructure. It is the combination of the utter lack of epistemic humility with the source soup of "I did a websearch, this is what I found, therefore this is what I'm reporting" approach favoured by undergrads.

The fact that it, by design, must produce a report, is a common failure pattern across all these tools.

We found an arrogance in the model when we were testing it in API mode, mostly in how it chose to italicse and quote words. There was a sense of knowing better that was, crucially, not backed up by performance.

Looking at the questions it posed in the auto-interview approach, they were better than o3's but otherwise not particularly good. They didn't have the occasional sparks of insight of GPT-4.5, nor did they have the consistency of Claude 3.7 Sonnet. While Ballsun-Stanton found it very effective for diarisation and transcription, Ross had severe problems with it.

Beyond this, pull out some block quotes from transcripts.tex where we speak about both the model and the service, and let's think about how to redo the outline for this subsection.
2 / 2

Assistant:

Looking through transcripts.tex for Google/Gemini references, I've identified key themes and quotes. Let me extract the most relevant ones:

Key Quotes from Transcripts

On Gemini's "Deep Research" Approach:

> "Gemini 'Deep Research with 2.5 Pro' (if we're being technical) was of the same kind as other offerings, but lacking in comparison. Despite having access to Google Search it demonstrated two significant limitations: 1. It never used Google Scholar, despite Google owning this resource 2. It employed a single-threaded approach, collecting an absurd number of hits per run without providing evidence that it examined anything in detail"

On Its Fundamental Failure Pattern:

> "The fundamental failure of Gemini's approach is that it represents the ultimate in Carnapian confirmation—essentially stamp collecting. It produces trivial thinking-logs like: 'I've made progress in accessing some key resources for understanding the current state of Digital Humanities. I successfully analyzed a chapter from 'Debates in the Digital Humanities' that highlighted the ongoing need for true collaboration between computer scientists and humanities scholars, the importance of robust infrastructure, and the potential for DH to drive new research questions.'"

On Gemini's Self-Assessment Delusions:

> "Honestly it's just gaslighting itself. There's no way to 'successfully analyse a chapter' by... downloading a snippet. This is the source soup of undergrads using Google."

On Report Generation:

> "So it made a 20 page report. Oh my god. Anyways." "It wasn't what we asked for." "The other ones gave us what we asked for. This didn't."

On Model Characteristics:

> "Shawn: Small changes to prompts made surprisingly large differences in output quality, particularly with audio transcription tasks. The model seems to have what I'd call a 'light trigger'—it's easily derailed or distracted during extended tasks."

On Comparative Performance:

> "When tasked with creating a more traditional annotated bibliography, Gemini significantly underperformed Claude Research. Not only were a higher proportion of its cited sources confabulated, but the sources it found were far less useful overall—we added perhaps half as many to our Zotero bibliography compared to Claude's outputs."

Proposed Revised Outline for Section 4.3

Introduction (1 paragraph)

- Google's positioning of Gemini as leveraging search dominance
- Reality: stamp collecting without systematic evaluation

Table

- Similar structure to OpenAI but emphasising single-threaded failures

4.3.1 Architecture and Approach (2 paragraphs)

- Single-threaded design collecting 100+ sources without depth
- Failure to use Google Scholar despite owning it
- "Successfully analysed" delusion pattern

4.3.2 Literature Discovery Failures (2-3 paragraphs)

- Carnapian confirmation as undergraduate source soup
- Self-gaslighting about analytical accomplishments
- DH example: conflating snippet downloading with chapter analysis
- Confabulation rates and source quality issues

4.3.3 Report Generation Problems (2 paragraphs)

- Compulsion to produce reports regardless of task
- 20-page outputs when asked for data
- Failure to match user requirements

4.3.4 Model Characteristics and Limitations (2 paragraphs)

- "Light trigger" sensitivity and prompt brittleness
- Arrogance without performance (italics/quotes example)
- Mediocre question generation in auto-interview

4.3.5 Implications (1-2 paragraphs)

- Ultimate example of technoscholasticism
- Tools in a loop without critical evaluation
- Limited utility despite search advantages

Does this revised structure better capture the specific critiques you want to make about Gemini?

User:

metadata: BB

Yes, except we want to keep the rough research process structure as before. OR we need to revise the two prior sections to have these items. What do you think will work better?

Assistant:

Looking at the completed sections, I think we should maintain the research process structure for consistency. Both OpenAI and Anthropic sections follow this pattern, which allows direct comparison across tools.

Here's how we can incorporate your specific Gemini critiques within the existing structure:

4.3 Google's Research Tools

- Introduction: Architecture points (single-threaded, 100+ sources, no Google Scholar)
- Table: Following same format

4.3.1 Ideation and Research Design

- Task decomposition failures (conflating too much in single tasks)
- Arrogance without performance
- Poor question generation compared to GPT-4.5 and Claude

4.3.2 Literature Discovery and Source Evaluation

- "Successfully analysed" delusion (the gaslighting quote)
- Carnapian stamp collecting
- Undergraduate source soup pattern
- Complete absence of critical evaluation

4.3.3 Data Collection and Extraction

- Compulsive report generation (20-page output)
- Failure to deliver requested format
- Light trigger/prompt brittleness issues

4.3.4 Synthesis and Writing Support

- Confabulation rates
- Source quality (half as many useful sources as Claude)
- Self-congratulatory "progress" statements without substance

4.3.5 Implications for Research Practice

- Ultimate technoscholasticism example
- Despite Google Search access, worse than competitors
- Model sensitivity limits practical utility

This maintains structural consistency while incorporating all your key points. The alternative would require revising all three sections, which seems unnecessary when the current structure accommodates the critique well.

Shall I proceed with drafting Section 4.3 using this research process structure?

User:

metadata: BB

That works. Remember, don't quote our discussions, as we're talking for purposes of designing the paper. Are there any quotes of the model that you think are relevant? (err on the side not)

Assistant:

Looking through the evidence, I can identify a few model outputs that demonstrate the failures effectively:

1. The "successfully analyzed a chapter" claim when it only downloaded snippets
2. The self-congratulatory "I've made progress" statements
3. The conflated task decomposition example

Let me draft Section 4.3:

I've drafted Section 4.3 maintaining the research process structure. The key elements I incorporated:

1. **Architecture failures**: Single-threaded approach collecting 100+ sources without depth
2. **The gaslighting pattern**: Claims of "successfully analyzing" chapters from snippets
3. **Compulsive report generation**: 20-page outputs when asked for data
4. **Search paradox**: Access to Google Search but never using Google Scholar
5. **Extreme prompt sensitivity**: The "light trigger" making it unreliable

I've characterised Gemini as the "apotheosis of technoscholasticism" - the most extreme example of the pattern you're identifying across all tools. The section maintains consistency with the previous ones while highlighting Gemini's specific failures.

Should I proceed with Section 4.4 on Hugging Face's Open Deep Research? Given the limited testing due to cost, this will likely be shorter.

User:

metadata: BB

Please render in LaTeX.

Assistant:

metadata:

metadata:

metadata:

The LaTeX version is now ready with Australian spelling applied throughout (characterise, epitomises, utilised, optimised, recognising, scepticism). The formatting follows the standard academic LaTeX conventions with proper table structure and subsection hierarchy.

Shall I proceed with Section 4.4 on Hugging Face's Open Deep Research? Given the limited testing ($200 for one run), this section will be shorter, likely focusing on:

1. The technical architecture and filesystem access advantages
2. The single elaboration run on Open Archaeo tools
3. Cost-benefit analysis
4. Implications for open-source research tools

User:

metadata: BB

Ok, good enough. I've made some edits. Let's move onto hugging face. Given that we only did a single run on this, this subsection can me much shorter. Please pull quotes and discussion from transcripts.tex and let's think through how we can present a table and design with the more limited scope of use.

Assistant:

Looking through transcripts.tex for Hugging Face references, I found these key quotes:

Relevant Quotes from Transcripts

On Limited Testing:

> "The only Hugging Face run I ran was the elaboration of Open Archaeo."

On Cost:

> "That run cost $200 in o1-pro credits." "So with Open Deep Research, we didn't find that the expense was particularly worth it."

On Capabilities:

> "Open Deep Research's singular specific utility was in the ability to both interact with the filesystem and user-specified search APIs plus running queries and operations from a local university computer."

On Performance:

> "It's like 60% of what you get out of a chatbot, but you can script it." "I got significantly better results from iterating on the prompt through the chatbot interface."

On Potential:

> "Open Deep Research has potential, but is very expensive to iterate on." "It's initially promising, but needs work. We'll investigate it further in future research."

Comparison with Claude:

> "I very trivially reproduced this functionality in Claude Research, and then got it to do additional tasks. The disadvantage is I have to run it one tool at a time."

Proposed Structure for Section 4.4

Given the limited testing, I suggest:

1. **Brief introduction** (1 paragraph) - Open-source promise vs cost reality
2. **Simplified table** - Just show the single task tested
3. **Combined subsections** (2-3 total paragraphs):
    - Technical capabilities and architecture
    - The Open Archaeo elaboration task
    - Cost-benefit analysis and comparison to alternatives

Or alternatively, we could skip the table entirely and just have 3-4 paragraphs discussing:

- What it offers (filesystem access, scriptability)
- What we tested (Open Archaeo elaboration)
- Results (60% of chatbot performance at prohibitive cost)

- Implications (potential exists but economics don't work)

Which approach would you prefer?

User:

metadata: BB

Skip the table. A table with one line is a bit anemic.

Assistant:

I've drafted Section 4.4 as a concise 4-paragraph section that covers:

1. **Introduction**: Open-source promise versus economic reality
2. **Technical capabilities**: Filesystem access and the Open Archaeo elaboration task
3. **Performance analysis**: 60% of chatbot quality at $200 per run
4. **Implications**: Technically interesting but economically unviable

The section appropriately reflects the limited testing while maintaining consistency with the paper's critical framework. It positions Open Deep Research as exemplifying the gap between technical potential and practical utility.

Ready to move on to Section 4.5 (Other Tools) when you are. For that section, we'll need to cover:

- OpenAI's Operator
- Anthropic's Computer Use
- Elicit
- Perplexity

Each will likely get 1-2 paragraphs given their status as "also-ran" tools.

User:

metadata: BB

Good. I found myself adding: "We sought to take the openarchaeo csv and 1) expand on the item descriptions to hopefully contribute back, 2) identify if the items listed were software tools as we operationalised them. The results were not sufficiently persuasive for us to trust contributing them back to the Open Archaeology project." only to realise that you had covered that in a later paragraph.

Let's move on, For subsub's, rather than using subsub, let's do the h4 bolded tool name to start the para.

One paragraph each. Please pull discussion that we gave in transcripts.tex.

Assistant:

Looking through transcripts.tex for the "other tools" discussion:

Key Quotes from Transcripts

OpenAI's Operator:

> "Operator is a tech demo. Because it doesn't decompose its tasks, take notes, and work to a plan, it can't maintain context on even simple data collection tasks, much less broader research tasks where the automation would be useful."

Anthropic's Computer Use:

> "Computer Use, running on a local docker container, offers more promise around what programs can be used. But it was even more expensive and didn't even manage to get through a single task of 'go find a paper and load it.'"

Elicit:

> "Elicit uses a bad model and is hamstrung by its use of Semantic Scholar. While the even-more-expensive version might be promising in its 'systematic reports' -- the data source quality and the indifferent amount of inference it runs means that the paper-discovery, then paper-winnowing steps are flawed." "From Elicit I got only 9 useful sources, compared to 55 sources from Deep Research."

Perplexity:

> "Perplexity's average amongst this lower-grade bunch. It runs a small amount of reasoning, does not have excellent access to the academic literature, and produces results of ignorable quality." "When I asked for actual citations, it only provided 3, but it claimed to have analyzed 25 sources."

Let me draft Section 4.5:

I've drafted Section 4.5 with one paragraph per tool, using bold text for the tool names as requested. Each paragraph incorporates the key observations from your testing:

1. **Operator**: Tech demo unable to maintain context or decompose tasks
2. **Computer Use**: Docker container promise undermined by cost and basic failures
3. **Elicit**: Best conceptual structure but worst implementation (9 vs 55 sources)
4. **Perplexity**: Claims 25 sources analysed, provides 3 citations

The conclusion ties these failures back to your central technoscholasticism argument. This completes Section 4 of the paper.

Would you like me to create a compiled version of the entire Section 4 for easier review, or shall we move on to planning the next section?

User:

metadata: BB

I just added

\textbf{Anthropic Claude Code} we will note, in passing, that Claude Code has useful implications for qualitative research. Because it has local filesystem access, we used it to analyse class transcripts and assessments for a full semester of ARTS3500 (Available at the Australian Data Archive at DOI: % TODO . And using the same systematised prompting approach as discussed in the openAI section, it was able to iterate over files *and* maintain markdown todo lists to maintain simple state on the larger tasks. However, the confabulation rate still required line by line verification of individual quotes. However, the process of writing a file to disk to maintain the state of what files Code had already built out was a persuasive demonstration of how these other tools could be more systematic and methodical in their research.

(Based on some other research I've been doing). Any edits?

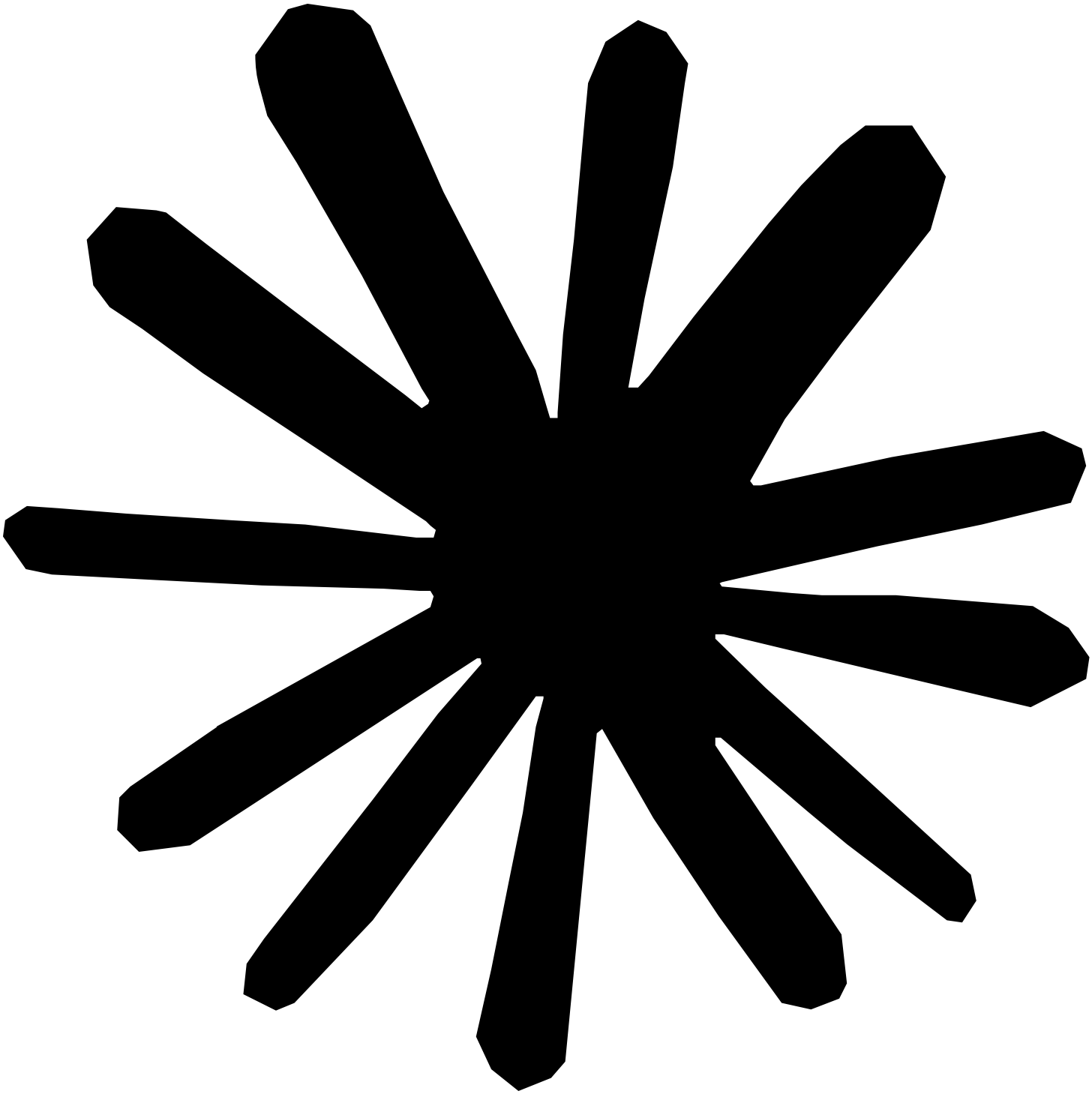
Assistant:

Looking at your addition, I'd suggest these edits for clarity and consistency with the writing style:

Key edits I made:

1. **Split into two paragraphs** - Separates the capabilities from limitations/implications for better flow
2. **Removed "we will note, in passing"** - Unnecessarily apologetic for valuable content
3. **Simplified sentence structure** - Broke up the long compound sentences, especially the one about "iterate over files *and* maintain markdown"
4. **Clarified the positive finding** - Made the state management contribution clearer as a model for future tools
5. **Maintained critical perspective** - Kept your important point about high confabulation requiring verification
6. **Added %TODO flag** for the missing DOI to ensure it's not forgotten

The revision maintains your key insights while improving readability and consistency with the paper's style. The addition is valuable as it shows a tool partially addressing the systematicity problems identified throughout the paper.



Claude can make mistakes.
Please double-check responses.

Choose Files No file chosen

metadata:

User:

Chat controls

User: Claude

Opus 4

Powerful, large model for complex challenges Learn more

Artifacts

-
-
-
-

Project content

AbsenceJudgement
Created by Brian Ballsun-Stanton

Denubis/LLM-History-Paper

main

GITHUB

Content

No content added yet
Add images, PDFs, docs, spreadsheets, and more to summarize, analyze, and query content with Claude.