Claude 4 Opus chat on: Opus-4 Section 4 Outline AbsJudge

User:

AbsenceJudgement /

User:

metadata: BB

Hi Claude (4 Opus),

Today the objective is to end up writing section 4 of AbsenceJudgement.tex, \section{Case Studies of AI Research Tools}. We have a number of things to do before we start writing. I left off by crunching down our discussions of the tool in transcript.tex. Also, on an earlier go, before my colleague and I talked about this section, we had a go at revising the outline:

\section{Case Studies of AI Research Tools} % Note: Open question - Should there be an introductory paragraph before diving into subsections? \subsection{OpenAI's Suite of Tools} % Subsection on OpenAI's Deep Research, GPT 4.5, and reasoning models (O3 and O1 Pro) % Note: Open question - How detailed should the comparison between these models be? \subsection{Anthropic's Offerings} % Subsection on Claude 3.7 with reasoning capabilities and Anthropic Research (May 2025 release) % Note: Open question - How much emphasis should be placed on the very recent Anthropic Research release? \subsection{Google's Research Tools} % Subsection on Google Gemini Pro 2.5 (model 03-25) and Google Gemini Pro 2.5 Deep Research % Note: Open question - Should we include the new model that just dropped if testing is incomplete? \subsection{Hugging Face's Open Deep Research} % Subsection on Hugging Face's Open Deep Research \subsection{Other Tools} % Subsection for tools with less practical utility for your research purposes \subsubsection{OpenAI's Operator} % Brief overview of OpenAI's Operator \subsubsection{Anthropic's Computer Use} % Brief overview of Anthropic's Computer Use \subsubsection{Elicit Research Report} % Brief overview of Elicit Research Report % Note: Open question - What specifically should be mentioned since you didn't have access to the systematic review? \subsubsection{Perplexity Research} % Brief overview of Perplexity Research

Thus, today the objective is to make sure that Section 4 can serve as evidence for the rest of the paper. To do that we need to write it (paying attention to the style guide). To do *that* we need to revise the outline until we have a good sense of what claims, evidence, and connecting warrants we are going to be making in each paragraph. (While this section is descriptive, instead of argumentative, having an idea of the logical flow of the paragraphs is useful.)

To begin, please functionally decompose the tasks and goal at hand to a useful level of precision, and give me a readback of what you understand the objective for this session to be.

Assistant:

<think> The user wants to develop Section 4 (Case Studies) from outline to written text. This section needs to provide empirical evidence supporting the paper's central thesis about AI tools lacking judgment (epistemic humility, inductive capacity, correspondence with reality) and exhibiting "technoscholasticism."

The transcripts contain their testing experiences with various AI tools, and they need to:

1. Refine the outline to specify what evidence goes where
2. Identify the logical flow between tools/subsections

3. Write the section following their style guide

I should decompose this into manageable tasks and confirm my understanding. </think>

Task Decomposition and Readback

My understanding of today's objective: Transform Section 4 from a skeletal outline into a fully written section that provides concrete empirical evidence supporting the paper's theoretical framework about AI judgment deficiencies.

Functional decomposition:

1. **Extract and organize evidence from transcripts**
   - Identify specific examples of judgment failures (epistemic humility, inductive reasoning, correspondence with reality)
   - Categorize technoscholastic behaviors by tool
   - Note instances of "mundane utility" despite limitations
2. **Refine the section outline**
   - Determine paragraph-level structure for each tool subsection
   - Identify which specific examples/evidence belongs in each paragraph
   - Establish logical flow between tools (why this ordering?)
   - Resolve open questions about scope and emphasis
3. **Map evidence to claims**
   - For each tool, clarify: What specific judgment deficiency does this illustrate?
   - How does each example connect to the broader technoscholasticism framework?
   - What patterns emerge across tools?
4. **Write the section**
   - Follow style guide requirements
   - Maintain descriptive rather than argumentative tone
   - Ensure each subsection provides clear evidence for the paper's thesis
   - Create smooth transitions between tools

Key decisions needed:

- How much detail for each tool (proportional to their importance as evidence?)
- Whether to include an introductory paragraph synthesizing patterns
- How to handle tools with limited testing (Anthropic Research, newer releases)

Should we begin by reviewing the evidence in your transcripts to determine what specific examples best support each aspect of your judgment framework?

User:

metadata: BB

Cool, To be clear, I would like you to ask me one *specific*, pointed, and critical question at a time until you believe you have enough information to accomplish the task at hand. I am not interested in empty praise here, the objective is to make the best possible academic paper.

The major tools should be subsections in their own right, where we talk about how we used them as part of the research lifecycle, what we observed, and where we found mundane utility. The tools which we tried once and then ignored can be lumped into their own isolated subsection as a group.

Yes, paying attention to transitions is important, though I suspect we'll want to craft that at the end. (Good to have a %TODO)

Let's look at what changes we think we need to make to the outline as a function of the conversations between Shawn and myself for Section 4. It should be at the end of the transcripts.tex document and each one should be clearly labelled with the prompt that the various AI asked us and our answer.

Assistant:

Looking at the transcript responses, I see you've already answered the restructuring question. You specified that OpenAI's suite, Anthropic's offerings, Google's tools, and Hugging Face should get dedicated subsections, while Operator, Computer Use, Elicit, and Perplexity should be grouped under "Other Tools."

Here's my first specific question:

In your Digital Humanities in Australia case study (which appears multiple times across different tools), you found that Deep Research presented the University of Newcastle's defunct Centre for 21st Century Humanities as active despite a death notice on the webpage. Should this specific example anchor the OpenAI subsection as the clearest demonstration of technoscholasticism, or do you have a more compelling example of OpenAI's judgment failures?

User:

metadata: BB

Ironically, that's not a great example (as phrased) of technoscholastisism. Because that's ignoring a specific textual element of evidence on the page itself. I do think hitting the DH explorations as one of the themes, plus the reports of use that Shawn provided will be a good basis for evidence at various stages of the research lifecycle. However, if we're anchoring DH, the point is that it demonstrated both the temporality problem and technoscholstic "believing what the page says because it is well formed in a university context"

Here, the issue is that these tools were not at all sceptical of claims made by pages, and then uncritically extrapolated limited, specific instances, to broader (positive) trends. The fact that failures are seldom announced as such meant that my injunction to be critical (I've got the quote in there somewhere, otherwise I can look it up), meant that it *looked for critical sources*, rather than critically evaluating the context (in a historiographical sense) of the page itself: when was it written, what was the context of the writing, what appeared around it, what were the intentions of the author, and how do surrounding documents support/not-support the outcome? I.e. claiming that there are active DH programs despite the classes being offered once in 2020 and never again.

Does this answer the question?

Assistant:

Yes, that's much clearer - the technoscholasticism isn't about missing text, but about uncritically accepting well-formatted institutional claims without historiographical assessment.

Here's my next question:

You mention "reports of use that Shawn provided" for various stages of the research lifecycle. Looking at the transcripts, I see Shawn found OpenAI Deep Research best for literature reviews (despite "undergraduate source soup" quality), while O3 performed better for tool metadata collection. Should the OpenAI subsection be structured chronologically by research stage (literature discovery → data collection

→ composition) to show where different models succeeded or failed, or would a thematic structure (technoscholastic failures → temporal blindness → effective applications) better serve your argument?

User:

metadata: BB

We need to be consistent across all of the subsections, and echo the pattern appropriately. Given that the discussion (Section 5) will be linking these to technoscholasticism more broadly, let's go with research lifecycle, so long as we have sufficient evidence of such for all of our major sections.

Assistant:

Looking at the evidence, I see uneven coverage across the research lifecycle for different tools.

For tools where you lack evidence across all research stages (particularly Hugging Face, which seems primarily used for data enrichment, and Google, which appears weakest overall), should we: (a) explicitly note the gaps as part of the evaluation ("we found no utility for X in literature discovery"), (b) focus only on the stages where you actually used each tool, or (c) reorganize to group tools by the research stages where they showed the most utility rather than giving each tool its own subsection?

User:

metadata: BB

Hugging face was used once. I'm happy having that be it's only focus, and yes, let us note the gaps as part of our case study in this section, but in passing. We're providing positive evidence here. I do think we want this by tool just so we can cross-reference though. Basically, we tried these tools until we hit a wall, and so discussing what we tried and when we stopped is important.

Assistant:

Understood - we'll maintain the tool-by-tool structure, noting gaps briefly while focusing on actual usage and stopping points.

For the Anthropic subsection: You mention Claude was "more conscientious" than Deep Research and that Shawn found it produced the most accurate tool metadata with "no outright inaccuracies." However, you also note "shockingly high rates of confabulation" in literature reviews. Should we lead with this contradiction (strong performance in structured tasks, weak in open-ended synthesis) to illustrate how different research stages expose different judgment deficiencies?

User:

metadata: BB

Yes. So, when research is doing a single, specific thing, it was quite good. When we let it loose to do a broader annotated bib: "research" it returned to confabulation. This is strong evidence of how important prompt, but even more so, scope-of-task is. It will generally do the same amount of effort in any given task, and it's up to us to zoom in and out until the scope of thetask is appropriate for the LLM at hand. When the scope is too broad, this will absolutely lead to confabulation. When there are too many sources leading to too many things, this will lead to confabulation, mostly because Research cannot emit single-subthreads to explore each source individually and then return the annotated bibliographies as a concatenated set. There are implications here for tool design (I muse quite a lot on this much earlier in transcript.tex, and this will form the basis of Section 6, so let's make sure we support it)

Does this make sense?

Assistant:

Yes, excellent point about scope determining performance. This directly connects to your judgment framework - the tools lack the epistemic humility to recognize when a task exceeds their capabilities.

When presenting this scope/confabulation relationship for Claude, should we explicitly frame it as a failure of epistemic humility (the system doesn't know it should narrow its focus or request task decomposition), or is this better positioned as evidence that these "multi-agent" architectures are still fundamentally single-threaded in their judgment, unable to genuinely parallelize the critical evaluation that human researchers would distribute across time and attention?

User:

metadata: BB

Well, scope/confab for all the models. We just noticed it in Claude 3.7 Sonnet the most because it could return quite effective results if prompted and scoped correctly. It's evidence for both.

OH! Quote we need to integrate somewhere (let me know where)

"I was going slightly spare at the fact that every talk at this Anthropic developer conference has used the word "agents" dozens of times, but nobody ever stopped to provide a useful definition. I'm now in the "Prompting for Agents" workshop and Anthropic's Hannah Moran finally broke the trend by saying that at Anthropic: Agents are models using tools in a loop I can live with that! I'm glad someone finally said it out loud."

\parencite{willison_agents_2025}

This is a problem of epistemic humility because, fundamentally, it doesn't push back, nor does it pause, nor does it decompose the task into multiple research runs. There is a fundamental assumption of the system (LLM + loops) that it can accomplish whatever well specified task at hand is presented to it. It's an absence of judgement because with the context window full, everything gets smushed about.

Your claim about parallelisation (AUSTRALIAN SPELLING) isn't quite right. It's not that we do things in parallel. Its that we do things one at a time, set them aside when we're done, and then return to them when we want to use them. It feels like this is an infrastructure problem, since the thing cannot leave notes for itself in passing as well as a design problem. There's also the critical evaluation of output that simply never happens, and that is problematic in quite a few different ways, starting from architecture and going upwards.

Let's probe this further, since this is going to be a theme throughout the many different subsections.

Assistant:

You're right - it's not parallelization but rather the inability to work sequentially with persistent state, setting tasks aside and returning with fresh context. The tools can't build a working memory outside their context window.

Given this fundamental constraint - that all these tools operate in single, consuming passes without the ability to maintain working notes or return to previous work - should we explicitly distinguish between two types of judgment failures: (1) architectural failures (can't escape the context window, can't maintain state) versus (2) epistemic failures (won't recognize when they should decompose, won't critically evaluate

their own output)? Or would this distinction overcomplicate the empirical evidence we're presenting in Section 4?

User:

metadata: BB

We absolutely should, as part of the structure of each subsection. However, we need to generalise and combine the thoughts of these failures in section 5. It feels like we should also emit section 5 outline notes as part of this process so we don't forget these observations when we start working on the next secction.

Here, this case study in section 4 does need to be about the tools, but we can look at these different axes as failure as part of our observations of tools.

Assistant:

Agreed - we'll track both failure types in each subsection while building notes for Section 5's synthesis.

For structuring each tool subsection, should we follow a consistent template like: (1) what we attempted across research stages, (2) architectural constraints observed, (3) epistemic failures manifested, (4) where we found mundane utility despite limitations? Or would you prefer a more narrative flow that weaves these elements together as you describe your testing experiences?

User:

metadata: BB

Consistent template, but without labelling. We should have the template in mind when writing, but I don't see any need to hit people over the head with it, no? (Thoughts?)

Assistant:

Agree - template as scaffolding for the writer, not signposting for the reader. The consistency will emerge through parallel structure rather than explicit labels.

Looking at your OpenAI evidence: you have the seven-stage research plan that couldn't execute (epistemic failure), the Digital Humanities temporality issues (technoscholasticism), and successful data collection after seven prompt iterations. Do you also have specific examples of architectural constraints (context window issues, inability to revise approach mid-stream) that manifested during your OpenAI testing, or should we draw those primarily from other tools?

User:

metadata: BB

So, the architectural constraint most fundamentally with Deep Research is the single threaded random-walk through sources. I think I speak about this early in transcript.tex -- can you go review for where I talk about it? If it's not there, we can revisit and I can go into more detail.

Assistant:

Found it. You discuss the random walk pattern around the '20250402 - Section 5 question 4' timestamp:

"my critique of the processes that I've seen in Elicit, in OpenAI, in Perplexity, is that they effectively random walk through the tree of possibilities with a vaguely Carnapian confirmatory bias where they they collect evidence until some internal vibe threshold is met, and then that's that."

And the single-threaded issue in '20250228 - Agentic.m4a':

"for Deep Research to be effective, it cannot have a single train of thought. It needs multiple supervisory threads."

You also note the "fundamental lack of systemization, even when explicitly instructed to" and that it "simply will not work down a list of sources."

Is this random-walk pattern most evident in how Deep Research jumped between different aspects of the Digital Humanities landscape without systematic coverage, or do you have a clearer example where you could track its wandering attention through a specific sequence of sources?

User:

metadata: BB

Ok, side note, just because I'm doing multiple things. I just tried Operator again (now that their system card announced that it was upgraded to using o3, rather than 4o) and it actually managed to return quite persuasive results on our FAIMS year-data test. (I've got a different thread here working up a prompt to see how well it does on a DH search)

Yes, exactly that. It did a thing, searched, did a thing, searched. There was no evidence that it ever looked back, or critically evaluated sources once it searched for them. There was no contextualisation or investigation of the authority of a source. The only times it retried were due to access issues. Once it found something it accepted the thing it found as evidence for the thing itself and moved on (Shawn's undergrad source soup).

For the tool use, it never just worked year by year, one after the other. It feels like this is a stamp-collecting mode rather than a systematic or collecting evidence mode. Ironically enough, Elicit is the only tool that had a source collection/source-evaluation steap breakout as explicit and intentional in its research reports mode. Unfortunately, it was hamstrung by the quality of model and the quality of sources it was picking up out of semantic scholar, but from a research lit review mode, it was easily the most "correctly structured" of themall

---

Assistant:

This stamp-collecting versus systematic evaluation distinction is crucial - and ironic that Elicit got the architecture right but failed on execution.

Given that Elicit demonstrated the "most correctly structured" approach with explicit source collection/evaluation phases, should we: (a) keep it in "Other Tools" but highlight this structural insight as a contrast to the timestamp-collecting approach of the major tools, or (b) promote this observation to the introduction of Section 4 as an example of how architectural design can't overcome poor model/corpus quality? This seems like important evidence for your Section 6 recommendations.

User:

metadata: BB

Yes, we keep it in other tools. It looks good superficially, but never returned anything of value. Other model providers can learn from it -- but there was no indication that the mid-level paid account is worth the money, or that it's worth experimenting with their expensive systematic reviews. They don't have the

literature access or model access necessary to provide a persuasive output. Furthermore, they do not reveal what model they use, so we cannot even indicate when people should pay attention to them. I do like the nod to this in the intro though,since this is a useful core thought.

Assistant:

Understood - Elicit stays in "Other Tools" as an example of good architecture undermined by poor foundations.

For Section 4's introductory paragraph: Should we open with the Willison quote about "models using tools in a loop" to immediately ground the discussion in what these "agentic" claims actually mean, then contrast the stamp-collecting reality across all tools with Elicit's structurally sound but practically useless approach? This would set up the pattern of good intentions failing against fundamental limitations.

User:

metadata: BB

I wonder if it should be here or elsewhere in the paper. I don't think we should quote it twice. Is it better here? (If so, absolutely, though the citation is going to be a bit odd, since it's a webpage referencing a personal communication, effectively)

Assistant:

The quote seems more definitional than empirical - it might fit better in Section 2.2 (Agency vs. Task Decomposition) where you're already distinguishing between true agency and task decomposition. There it would serve as industry acknowledgment that even Anthropic defines "agents" minimalistically.

For Section 4's introduction, should we instead open with a brief paragraph noting that across all tools tested, regardless of marketing claims about "agentic" capabilities, you observed consistent patterns of stamp-collecting without systematic evaluation - then use the rest of the section to detail how this manifested differently in each tool?

User:

metadata: BB

Great, before we continue, could you give me something to stick in Section 2.2, please?

Assistant:

Here's a snippet for Section 2.2:

User:

```
latex

  Even industry practitioners acknowledge the definitional ambiguity around "agentic" systems
```

Should I adjust the tone or focus?

User:

metadata: BB

That's fine, not sure where to put it though. Ideas?

Assistant:

Looking at Section 2.2.1, I'd suggest placing it after the paragraph that ends with "This distinction is not merely semantic—it determines the appropriate level of trust and delegation."

This placement would:

1. Follow your critique of marketing terminology
2. Precede your discussion of the agency spectrum (task-decomposing → goal-following → goal-setting)
3. Provide industry acknowledgment that even minimal definitions aren't being met
4. Set up your argument about judgment deficits undermining even basic agency claims

Does this placement make sense given your argument flow?

User:

metadata: BB

Great. That works, digression over.

You asked: > For Section 4's introduction, should we instead open with a brief paragraph noting that across all tools tested, regardless of marketing claims about "agentic" capabilities, you observed consistent patterns of stamp-collecting without systematic evaluation - then use the rest of the section to detail how this manifested differently in each tool?

Yes, that ties it in with earlier sections. Basically, we found mundane utility in places requiring no judgement, where this stamp collecting was on a specific task, and that it genuinely was a case of doing a tedious number of web searches. As we increased scope or complexity (annotated bibs, lit reviews), quality decreased. My perspective is that this is due to a concomitant requisite increase in the need for effective judgement. The more a given run requires the effective application of judgement, the lower the quality.

Assistant:

Excellent framing - the inverse relationship between judgment requirements and output quality.

For the OpenAI subsection specifically: You have strong evidence of the seven-iteration prompt engineering success for data collection (low judgment requirement) versus the Digital Humanities failures (high judgment requirement). Should we structure this subsection to move from the initial ideation failures (seven-stage plan) → through the DH lit review problems → to the eventual data collection success, showing this degradation-to-utility arc? Or would you prefer a different narrative order?

User:

metadata: BB

Yes, I'm happy with that. It's what we did.

Assistant:

Good. Looking at the specific evidence for each stage:

For the ideation failure stage, you mention the seven-stage plan but I don't see the specific example in the transcripts. Do you have the actual plan Deep Research proposed (even just the stage titles), or should we focus on the general observation that it "made this wonderful seven-step research plan" that required seven iterations to reduce to simple CSV output? This seems like a powerful example of epistemic hubris - proposing capabilities it couldn't execute.

User:

metadata: BB

Yes, i have a copy of it. It's very long.

Give me a prompt to extract what you want out of the conversation, please? I'll run it in a different thread.

Assistant:

Here's a prompt to extract what we need:

"I need to extract information about the seven-stage research plan that OpenAI's Deep Research initially proposed for the systematic data collection task. Please provide: (1) The specific task/prompt you gave it, (2) A brief summary or list of the seven stages it proposed, (3) What it actually accomplished versus what it claimed it would do, and (4) How many iterations it took to get from this elaborate plan to functional CSV output. If the full plan is very long, just the stage titles/summary would be sufficient to demonstrate the scope mismatch between its proposed capabilities and actual performance."

User:

metadata: BB

(Right, ... I should have read that before running. The prompt itself is 1 and 2). For 3, it took me 7 iterations to get something that DR will return a CSV that is mostly persuasive. It was far worse on finding things that corresponded to our definitions of tools. 4, it took me looking at the first piece of the prompt: "Initial Tool Identification and Context “Search for academic publications, project pages, or any official announcements that introduce or describe the software named [SOFTWARE_NAME]. Identify the earliest date of release or mention. Provide both the date and the source of this information (article citation, project website, or other).” Refinement: * If direct references do not appear immediately, expand the search to synonyms, related project names, or references in review articles. * If multiple candidate tools with the same name appear, cross-check domain or authors to confirm you have the correct tool." And building out something that would return a CSV. That took 7 goes. (Looking at our prompts archived in overleaf)

Our final functional prompt that got useful results out of DR was (see attached). Note the absurd increase in guardrails and prescriptive do this, do that.

Your summary output from parts 1 and 2 of your prompt (I'd also like to meta-note that this is an example of the exact failure modes we're talking about...)

"<think> Looking through the documents for references to the seven-stage research plan that OpenAI's Deep Research initially proposed. I can see this is mentioned as a key example of epistemic humility failure in Section 4. From the documents, I can find:

- **The task was to create "a synthetic structured prompt to systematically gather longevity data on historical research software"**

- **Deep Research produced "an elaborate seven-stage plan it was utterly incapable of executing"**
- **It required "seven iterations" to produce even a basic CSV output**

- **The transcripts mention "it made this wonderful seven-step research plan"**

  However, I don't see the actual initial seven-stage plan documented in the provided materials. Documents 5 and 6 appear to contain the final, working versions of the prompt after iteration, but not the original overambitious plan. </think>

  Based on the documents provided, I can see that:

  The Task/Prompt Given: You asked OpenAI's Deep Research to create "a synthetic structured prompt to systematically gather longevity data on historical research software" (as referenced in Section 4 of your paper).

  About the Seven-Stage Plan: While your documents reference this as a key example of Deep Research's epistemic humility failure - where it "produced an elaborate seven-stage plan it was utterly incapable of executing" - the actual initial seven-stage plan isn't included in the materials you've provided.

  What I do have are documents 5 and 6, which appear to be the *final working versions* of the prompt after your seven iterations of refinement. These successful versions show a systematic approach with steps like:

- **Identify Initial Release and Context**
- **Document Key Details**
- **Locate Source Code Repository**
- **Analyze Repository Activity**
- **Check Software Preservation Measures**
- **Gather Mentions in Scholarly Literature**
- **Assess Ongoing Maintenance and Use**

- **Compile a Longevity Profile**

  But these represent what eventually worked, not the initial overambitious plan that failed.

  Do you have the original Deep Research response saved that contained the seven-stage plan it couldn't execute? That would be valuable evidence to include or reference more specifically in your paper to demonstrate the gap between its proposed capabilities and actual performance."

Let's keep going on this line, since I think we need to synthesise multiple aspects of our current conversation and I'm not persuaded we're clear.

I'm also getting worried that we're reaching the length limits of this chat, so let's consider what we need to do before we can achieve the outline, please.

Assistant:

You're right - let's step back and synthesize before we hit context limits.

Given our discussion so far, here's the Section 4 structure emerging:

1. Intro paragraph: Inverse relationship between judgment requirements and utility
2. OpenAI: ideation failure → DH failures → data collection success
3. Anthropic: strong structured tasks, weak open synthesis, confabulation at scale

4. Google: ultimate stamp-collecting, single-threaded, temporality blindness
5. Hugging Face: one-shot data enrichment use
6. Other Tools: Elicit (good structure, bad execution), others as brief dismissals

What's the ONE critical piece of evidence or clarification you need me to understand before we can draft the complete paragraph-by-paragraph outline?

(Should we create that outline artifact now and fill in the specific evidence mappings, or do you need to clarify something else first?)

User:

metadata: BB

I think we're good. I'm happy with those core points. Let's make that outline.

Assistant:

I've created the detailed outline with specific evidence mapped to each section. Key elements:

1. **Introduction** establishes the inverse relationship between judgment requirements and utility
2. **Each tool section** follows our implicit template (attempts → constraints → failures → utility)
3. **Specific examples** are tagged with context (DH Australia case, seven-stage plan, etc.)
4. **Other Tools** are kept brief but highlight why each failed

Two questions before we proceed:

1. Should I add explicit %TODO markers for transitions between major sections?
2. Do you want to add any Section 5 notes to the outline now (themes like architectural vs. epistemic failures, the scope/confabulation relationship, temporality blindness)?

User:

metadata: BB

Yep, let's move this into LaTeX format (comments for the individual paragraph sections, while preserving the section and subsections present in the LaTeX. Please turn the paragraph level design into comments, with clear TODOs on tasks we need to do before we start writing.

And yes, please give me context that I can add (commented out) to 5 so we don't lose any insights from this conversation.

Assistant:

I've converted the outline to LaTeX format with:

1. **Section structure** preserved as LaTeX commands
2. **Paragraph-level content** converted to comments with clear descriptions
3. **TODO markers** for specific tasks before writing:
    - Finding OpenAI marketing quotes
    - Getting the seven-stage plan from archives
    - Finding specific failure examples
    - Comparison metrics between tools

4. **Section 5 notes** at the bottom capturing our key insights:
    - Architectural vs epistemic failures distinction
    - Scope/confabulation relationship
    - Temporality blindness pattern
    - Stamp-collecting behavior
    - Integration points for the Willison quote

The outline maintains your research lifecycle structure within each tool section while noting where you need specific evidence. Should we tackle any of those TODOs now, or would you prefer to move forward with another task?

User:

metadata: BB

Great. This works. Ok, Give me a long prompt that I can use to resume this conversation in a new thread with all necessary context. I've committed the outline, but I'd like to have us resume with all necessary context and insights when we start writing.

Also, I can drop anything necessary in an artefact into this project, if you'd like to capture notes and "memories" about this process.

Assistant:

Here's a comprehensive prompt for resuming in a new thread:

I'm writing Section 4 (Case Studies of AI Research Tools) of my paper "An Absence of Judgment: AI's Limitations in Deep Research tasks". The paper's core thesis is that AI tools lack three crucial dimensions of judgment: epistemic humility, inductive capacity, and correspondence with reality. We characterize these tools as exhibiting "technoscholasticism"—privileging textual authority over critical assessment of knowledge claims.

I've completed a detailed outline (attached) with specific evidence mapped to each tool. Key insights from our planning:

1. **Core finding**: Inverse relationship between judgment requirements and utility. Tools perform well on mechanical tasks but fail when judgment is needed.
2. **Pattern across all tools**: "Stamp-collecting" behavior—random walk through sources with vague Carnapian confirmatory bias until internal "vibe threshold" met. No systematic evaluation or revisiting sources.
3. **Architectural vs Epistemic failures**:
    - Architectural: single-threaded, can't escape context window, can't maintain state
    - Epistemic: won't recognize need to decompose, won't evaluate own output, lack temporality awareness
4. **Specific evidence patterns**:
    - OpenAI: Seven-stage plan → DH temporality failures → data collection success (with heavy scaffolding)
    - Anthropic: Strong on narrow tasks, confabulates at scale, "vibes" on search results
    - Google: "Ultimate Carnapian confirmation," processes 20-50 sources at once without scrutiny
    - All tools: Technoscholastic acceptance of institutional claims, temporality blindness
5. **Key examples**:
    - Digital Humanities in Australia (all tools presented defunct programs as active)
    - Western Sydney offered one course in 2020, presented as ongoing program
    - University of Newcastle centre with death notice presented as vibrant
    - Tools can't infer from "useful absences" of evidence

Writing requirements: Follow style guide (Australian spelling, active voice, 2-3 sentence units). NO empty praise or compliments. Be critical and specific. Ask pointed questions. Evidence-based writing drawing from transcripts.

Current task: Write Section 4 following the outline structure. Each tool section should demonstrate specific judgment failures while noting where mundane utility was found. Maintain consistent implicit template across tools without explicit labeling.

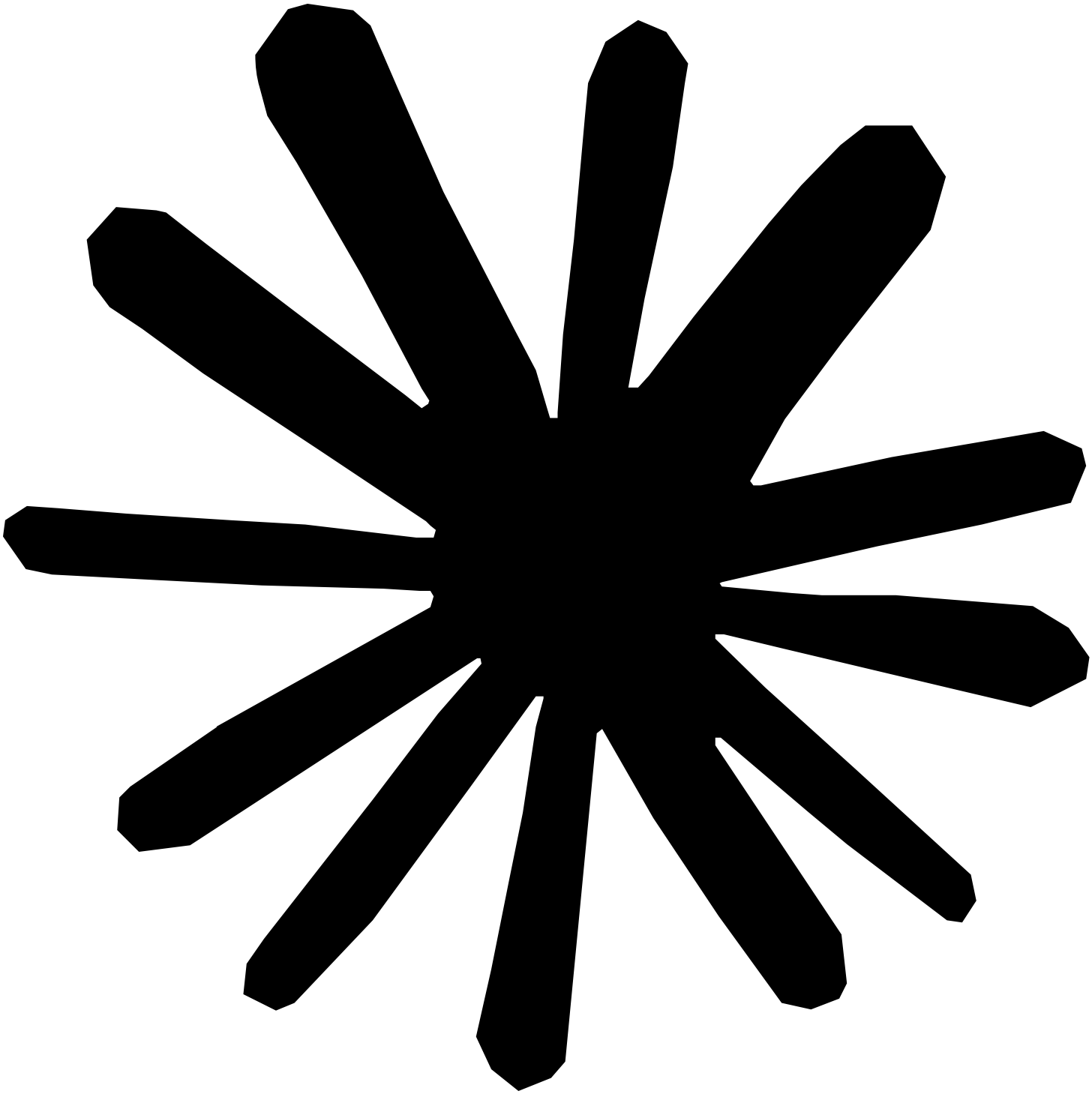
Please begin by reviewing the outline and asking any specific clarifying questions needed before we start writing the introduction paragraph.

I've created both a comprehensive resumption prompt and an artifact with our key notes and evidence. The prompt captures our analytical insights, specific evidence patterns, and writing requirements. The artifact serves as a quick reference for the judgment framework, specific examples, and writing guidelines.

The prompt emphasizes:

- The inverse relationship between judgment and utility
- Specific failure patterns for each tool
- The critical, evidence-based approach needed
- The architectural vs epistemic distinction we developed

Is there anything specific from our conversation that you feel needs stronger emphasis in either the prompt or the notes artifact?



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