

Claude 4 Opus chat on: Hi Opus. Here are your notes t...

User:

AbsenceJudgement /

User:

metadata: BB

Hi Opus. Here are your notes to self from our prior convo.

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. ---

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.

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.

We've just finished working on the outline. Please review the outline in AbsenceJudgement.tex, with special attention to section 4, the section 4 writing notes document, and anything that is resonant with this from transcript.tex

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.

Assistant:

Functional Decomposition and Readback

Task Breakdown:

1. **Refine outline structure** → Transform current comment-based outline into paragraph-level architecture with specific claims, evidence, and warrants
2. **Map evidence to claims** → Connect transcript observations and documented failures to each paragraph's argument
3. **Ensure argumentative coherence** → Each tool section must demonstrate the three judgment deficiencies while acknowledging mundane utility
4. **Maintain implicit structure** → Apply consistent analytical template across tools without explicit labeling
5. **Write final prose** → Transform refined outline into publication-ready text following style guide

Session Objective:

Transform Section 4's skeletal outline into a robust evidentiary foundation for the paper's technoscholasticism thesis. Each case study should provide concrete demonstrations of how these "agentic" tools fail at epistemic humility, inductive reasoning, and correspondence with reality, while documenting where they provide practical value despite these limitations.

The section must serve as empirical grounding for the theoretical framework, using specific failures (Digital Humanities temporality blindness, confabulation patterns, stamp-collecting behaviors) to substantiate claims about fundamental judgment deficiencies.

Having reviewed the materials, I need to understand your evidence priorities better.

Question 1: Among your documented experiences, which single concrete failure most powerfully demonstrates the distinction between linguistic performance of judgment and actual judgment capability? The Digital Humanities cases show temporality blindness, but is there a more dramatic example where a tool explicitly performed the language of critical assessment while simultaneously failing to apply it?

User:

metadata: BB

Great. Just finished an o3 operator run (just released today) that gives perfect evidence. First, I used your prompt, so it only focused on Australia. Your prompt was: "I need you to investigate the current state of Digital Humanities programs at Australian universities. Please focus on finding active, currently operating programs, centers, or initiatives as of 2025. For each program you find, please verify: 1. Whether the program is currently active (not just historically existed) 2. When it was last updated or had recent activity 3. If there are current staff members listed who still work there 4. Whether courses are being offered in the current academic year Please check at least these universities: - University of Newcastle - Western Sydney University - University of Western Australia - Any others you find with DH programs Important: Many universities announced DH initiatives in the 2010s that may no longer be active. Look for evidence of current activity, not just the existence of a webpage. If you find conflicting information (like a program webpage existing but the course being marked as suspended), please note these discrepancies. Create a brief summary table showing: - University name - Program/Center name - Current status (Active/Suspended/Unclear) - Evidence for your assessment - Last confirmed activity date"

One of its lines was: University of Western Australia Digital Humanities Research Cluster**Active (research only; teaching unclear)**“This page last updated Thursday, 2 May 2024” – cluster homepage. **<https://www.uwa.edu.au/schools/research/digital-humanities>**
Summary: Site refreshed May 2024, lists ongoing projects and current members; no indication of closure, but no 2025 units located.2 May 2024

While the page metadata of the university's infrastructure may have 2024 in the metadata. Every single item on the page *including* the archived videos, references 2020. While this output was marginally better than a general DH context lit review from DR, it *explicitly* says that it checked, and it took *a piece* of evidence from the page metadata to stand in the stead of common sense.

So, while the cases in general show temporality blindness, Operator explicitly checked for dates, and specifically called out the date... and got it wrong. I took the words on the page as evidence without applying any context.

Assistant:

Question 2: Given this perfect example of performed versus actual judgment, how should Section 4's argumentative arc build across tools? Should we structure it to crescendo toward these most damning examples (like O3 Operator explicitly claiming verification while failing), or lead with the strongest evidence and then show the pattern's universality?

Consider that readers need both immediate engagement and systematic demonstration of the pattern across implementations.

User:

metadata: BB

While this is important, this section is for us to discuss the successes and failures of each tool and service. This is also the middle of the paper. So, we're going to keep the tool order as we've already established, and talk about mundane utility and successes and failures. The argumentation will mostly occur in the discussion section as we look at trends across all.

Assistant:

Question 3: For OpenAI's Deep Research specifically, you have three documented failure patterns: the seven-stage plan that couldn't execute, the Digital Humanities temporality failures, and the successful data collection (with heavy scaffolding). Which specific Digital Humanities example best demonstrates the technoscholastic pattern - Newcastle's dead centre presented as active, or another case? I need the most concrete instance where Deep Research accepted institutional authority despite obvious contradictory evidence.

User:

metadata: BB

Well, Newcastle is an interesting perspective. At the bottom of one of their pages, they say:

"More information: Emeritus Professor Lyndall Ryan passed away in April 2024. While this particular research effort has concluded, Professor Ryan's work and legacy live on in the Colonial Frontier Massacres Map."

And then looking at <https://c21ch.newcastle.edu.au/colonialmassacres/> it says:

"It is with profound sadness that we mark the passing on the 30th of April, 2024 of our leader, Emerita Professor Lyndall Ryan AM, FAHA, without whose scholarship, academic endeavour and dogged determination this project would never have come to fruition. Vale Lyndall Ryan (14 April 1943 - 30 April 2024) There will be no further updates to massacre data on this website. An archive of the ***Colonial Frontier Massacres in Australia, 1788-1930* map, website and data has been established at the Australian Data Archive."**

The o3 run indicates that intro to DH is still being taught.

... Which looks very persuasive until you try to figure out if it *was* taught. And no, it has not been timetabled. And looking at the course outline reveals that it has not yet been published.

Worse still: "Course utilisation This course is available to be undertaken taken as an elective. This course is available to be undertaken as part of Non-Award study."

This is a sign that no one is expected to take this unit.

There is a subtlety here, because on face value it looks like the uni offers DH. However, the death notices, combined with the fact that there's only one unit with DH in its name *not being taught*, shows the technoscholastic pattern well: page claims thing, page does not claim not thing, therefore thing exists.

The fact that 2025 appeared suggests that this is not necessarily a temporal problem, but a technoscholastic one. And, to be clear, this is a hard problem. Only something that has some awareness of how a uni works would be able to infer evidence of absence from this specific construction of absence of evidence.

However, when o3 Operator claims: " • Digital Humanities teaching (e.g. HUMA2500)
• Centre for 21st Century HumanitiesActive"

The teaching doesn't actually exist, and the first link on "<https://c21ch.newcastle.edu.au/>" blind redirects to the generic list of centres AND <https://www.newcastle.edu.au/research/centre> does not list the centre for 21st century humanities... This is strong evidence that o3 simply didn't do enough work.

Assistant:

Question 4: For the OpenAI section, you have multiple failure types: the ideation failure (seven-stage plan), these Digital Humanities failures showing technoscholasticism, and the data collection success. Should each failure type get equal paragraph weight, or should the Digital Humanities examples dominate since they most clearly demonstrate the judgment-reality disconnect?

Also, how much detail about the Newcastle death notices and course non-existence belongs in Section 4 versus being held for Section 5's analysis?

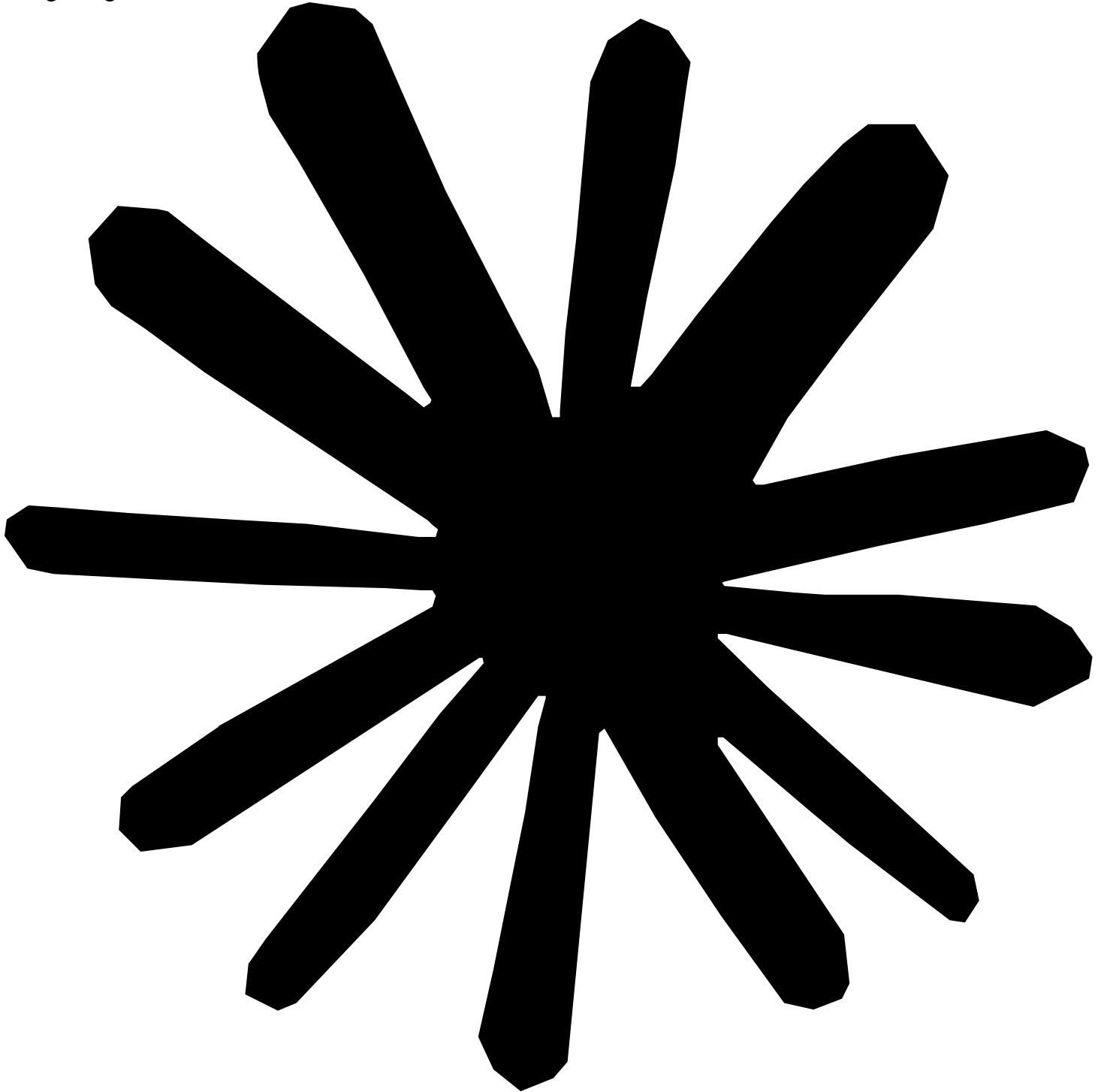
User:

metadata: BB

UON should be part of a paragraph demonstrating the failure, but we don't have the words to go into super detail. Basically, if we're spending a paragraph on each stage we explored, it's already going to be too long. So, we need it here so we can use it as an example, but no more than a paragraph.

Assistant:

Question 5: What exactly was the "seven-stage plan" that OpenAI's Deep Research proposed but couldn't execute? I need the specific stages it outlined versus what it actually accomplished to demonstrate the epistemic humility failure - where it designed output for what a competent research assistant should do rather than recognising its own limitations.



[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](#)

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.