Claude 4 Opus chat on: Section 5 Outline Strategy

User:

AbsenceJudgement /

User:

metadata: BB

Hi Claude. We now need to work on a detailed paragraph-level outline of section 5. We need to take the evidence that we've presented in the prior sections and combine it with the theoretical framework to discuss the larger patterns of successes and failures that we've observed in this tool. One thing I want to link back to is "agents are models using tools in a loop" versus the goal-following and goal-setting distinctions I made earlier.

We also may want to move prior paragraphs into this section. Beyond that, I've done an *extensive* amount of reflection of these ideas in transcripts.tex.

Please functionally decompose the task at hand and give me a readback.

Assistant:

<think> The user wants me to:

1. Create a detailed paragraph-level outline for Section 5 (Reflective Analysis)
2. Synthesize evidence from Section 4 (case studies) with the theoretical framework from Section 2
3. Specifically connect to the "agents are models using tools in a loop" (Willison) vs their goal-following/goal-setting distinctions
4. Consider moving relevant paragraphs from other sections
5. Draw on their extensive reflections in transcripts.tex

Looking at the current structure and the key themes noted in comments, I need to build an outline that:

- Analyzes patterns across all tested tools (not just individual failures)
- Links back to the three judgment dimensions (epistemic humility, inductive reasoning, correspondence with reality)
- Addresses the architectural vs epistemic failure distinction
- Explores the relationship between task scope and output quality
- Examines systematic issues like temporality blindness
- Connects to the broader question of why LLMs haven't made discoveries (Dwarkesh's question)
- Concludes with practical implications (mundane utility patterns)

The existing subsections provide a skeleton, but need more detailed paragraph-level planning. </think>

Task Decomposition and Readback

I understand you need a detailed paragraph-level outline for Section 5 that:

1. **Synthesises patterns** from the Section 4 case studies with the Section 2 theoretical framework
2. **Connects architectural findings** to your three dimensions of judgment deficiency (epistemic humility, inductive reasoning, correspondence with reality)
3. **Links Willison's minimal definition** ("agents are models using tools in a loop") to your more substantive goal-following/goal-setting distinction

4. **Draws from your transcript reflections** to develop the key themes already identified in comments
5. **Considers paragraph migrations** from other sections where they better serve the analytical purpose

The section should move from specific pattern recognition through theoretical integration to practical implications, building an argument about why these systems fail as autonomous researchers while retaining mundane utility as tools.

Shall I proceed with creating the detailed paragraph-level outline, incorporating your transcript insights and the evidence patterns from Section 4?

User:

metadata: BB

Please ask me one clarifying, specific, question at a time until you think you're ready -- noting that I've already answered many. These questions are likely about tying thoughts together or other judgement calls you think I should make.

Assistant:

In Section 5, should technoscholasticism be presented as the overarching explanatory framework that manifests through the three judgment deficiencies, or should we treat it as a fourth, separate pattern alongside epistemic humility failures, inductive limitations, and reality detachment?

User:

metadata: BB

No, these three things are fundamental components of technoscholasticism, please go review section 2.

Assistant:

<think> Looking back at Section 2, I see that technoscholasticism is indeed the overarching framework that encompasses and explains the three judgment deficiencies. The three dimensions are manifestations of this broader technoscholastic worldview - the privileging of textual authority over critical assessment.

I need to ask a different question about structuring Section 5. </think>

Given that Section 4 presents evidence of failures across all tested tools, should Section 5 begin with the cross-cutting patterns (like temporality blindness and the scope/confabulation relationship) before analysing how these patterns demonstrate technoscholasticism through the three judgment dimensions, or would you prefer to structure it around the three dimensions first and then identify the common patterns within each?

User:

metadata: BB

The former, I think. We are building up to the point that these failures are present in reality and illustrations of the theoretically defined technoscholasticism. ... Think through this please, I'm not convinced.

Assistant:

<think> The user wants me to think through whether Section 5 should start with empirical patterns (like temporality blindness) and then show how they demonstrate technoscholasticism, or start with the three

theoretical dimensions and find patterns within each.

Looking at the paper's structure:

- Section 2 establishes the theoretical framework of technoscholasticism with three judgment dimensions
- Section 4 presents empirical evidence from testing
- Section 5 needs to synthesize these

Starting with cross-cutting patterns seems more persuasive because:

1. It follows naturally from the evidence in Section 4
2. It's more inductive - showing patterns first, then explaining them theoretically
3. It demonstrates that these aren't random failures but systematic ones
4. It builds to the "aha" moment where empirical patterns align with theoretical predictions

This approach would show: observed patterns → systematic nature → theoretical explanation → implications </think>

After thinking through the structure, I believe starting with the cross-cutting empirical patterns is indeed stronger. Here's why:

We've just presented extensive evidence in Section 4. Readers have seen failure after failure across different tools. Starting Section 5 with the patterns they've just observed (temporality blindness, scope/confabulation relationship, stamp-collecting behaviour) creates immediate recognition.

Then showing these aren't random tool-specific quirks but systematic patterns across all implementations builds the argument. Finally, revealing how these patterns map precisely onto the three theoretical dimensions of technoscholasticism creates a satisfying synthesis where empirical observation confirms theoretical prediction.

Should Section 5 open with a paragraph explicitly bridging from Section 4's evidence to the pattern analysis, establishing that we're now moving from specific tool observations to systematic analysis?

User:

metadata: BB

Yes, please. This is the "So what?" chapter.

Assistant:

Given that we're positioning this as the "So what?" chapter, should the opening bridge paragraph explicitly preview the answer to Dwarkesh's question (that technoscholasticism prevents discovery), or should we build to that revelation after establishing the empirical patterns?

User:

metadata: BB

Yes, we should. It's a *long* paper. But we should introduce it, and then work towards proving it.

Assistant:

Should the section end with a strong restatement of how answering Dwarkesh's question (why no discoveries despite vast knowledge) validates the broader technoscholasticism framework, or would you prefer to end more practically with the implications for researchers using these tools?

User:

metadata: BB

practically. While we think that's the answer, the negative answer isn't proof. Instead we can say: "Given we think that this is the answer, here are the patterns to find mundane utility." which then leads into the next chapter, no?

Assistant:

metadata:

metadata:

Is this outline ready for you to begin writing Section 5, or would you like me to integrate specific quotes or evidence from your transcripts into the paragraph descriptions first?

User:

metadata: BB

Let's go through and think through *ideas* from the transcripts. It's not evidence. It's our thoughts. I can go dig for quotes from our prompts where needed.

Assistant:

Looking through the transcripts, here are the key ideas that should inform Section 5:

On the scope/effort relationship: Your insight that models "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 the task is appropriate for the LLM at hand" explains why narrow tasks succeed while broad ones fail.

On the Carnapian confirmation pattern: You described Deep Research as accumulating web pages "until it feels confident" - this vibe-based threshold rather than systematic evaluation characterises all the tools. They're "stamp collecting" without critical assessment.

On the Aristotelian fallacy: Your observation that "AI will never ever look at its words and go, 'Nah, those those weren't good. Let me throw them out instead of giving them to the user'" captures the absence of self-critical capacity. They cannot doubt their own outputs.

On temporality blindness: The pattern where systems can't infer that "a class which exists in 2020, but not 2021-2025 is *not* an example of a currently active Digital Humanities group" - they can't recognise useful absences as evidence.

On architectural constraints: The "no hands" problem - these systems can only engage with text, never with reality. Combined with inability to "move the cursor around" or delete bad tokens, this creates compound limitations.

On missing impulses: "The impulse to edit, the impulse to self-critique is not a pattern exhibited in their training data" - published work shows the consequence, not the process, of critical thinking.

Should I continue extracting more conceptual insights, or would you like to start integrating these into the Section 5 outline?

User:

metadata: BB

Keep extracting. These are the core points of the paper.

Assistant:

Continuing with key conceptual insights from the transcripts:

On the "no judgment" pattern: "These systems have to leverage human judgment. And not in the double confirm that Deep Research does right now... This is not agentic, it is not goal following. We cannot treat it as goal following. We must treat it as a tool. It's just a tool."

On the need for tombstones: "We need to have indications of thoughts that we've had that are bad. And we have to be able to remove them durably from the context window, but we need to have proof that we've had this thought and chosen not to have it."

On multiple supervisory threads: Deep Research needs "multiple supervisory threads. If policy enforcement is the deal that it needs, which is fine, it needs a supervisory thread to review and edit. It needs an intrusive thought thread for when it's looking for validation."

On the waterfall problem: "The idea of a waterfall model of research is laughable." Research requires the ability to "move around our to-do list... go back up a stage... discover that our original research question was wrong."

On the undergraduate source soup: The outputs are "uncritical aggregation without the historiographical awareness to evaluate when claims were made, by whom, for what purpose, and whether they still apply."

On correspondence with reality: "There is no world model, there is no outer world, there is no knock on the table. There is merely the words and the sophisticated but still present statistical relationships between words."

On the induction problem: "Induction is not merely the process of grouping, but the process of articulating a covering rule... It is the process of generalization of disparate data, not merely grouping similar data."

On "vibes": The recurring theme that these systems operate on "vibes" - from Carnapian confirmation thresholds to quality assessment, it's all "vibe-based" rather than systematic.

Should I continue extracting more insights or start discussing how these integrate into the Section 5 structure?

User:

metadata: BB

Keep extracting where they are present.

Assistant:

Continuing with conceptual insights from the transcripts:

On the performance of virtues: "The performance of a virtue is not having the virtue, which is kind of the whole problem with large language models."

On the failure of functional decomposition: When Deep Research proposed its seven-stage plan, "it produced an elaborate seven-stage plan it was utterly incapable of executing" - demonstrating the gap between linguistic capability and actual competence.

On textual authority versus reality: "Two an AI, text and authority has primacy over reality. I mean, it has to. That's baked into the thing itself... AI's are fundamentally Aristotelian that we make arguments by looking for authorities."

On the absence of experimental thinking: "These models never engage with the real... They never say 'I don't know', they also never say 'I can't know'... It will never say, 'Well, go try this, see if it works, report back'."

On the grain of different models: Each model has a "grain" (like wood grain) - "a direction in which it moves naturally versus directions requiring force." Working with rather than against each model's grain is essential for mundane utility.

On systematic versus random approaches: "Basically, rather than treating Deep Research as a single train of thought that wanders and hops and skips through the literature, it is something that can be systematic" - but current implementations fail at this systematicity.

On the importance of iteration and time investment: "Spend 8 hours improving a prompt on 3-4 models to find out the 'grain' of the prompt. Spend time iteratively prompting, and compare responses *across* model families."

On the wiki/annotation approach needed: "Building an annotated bibliography of the works consulted, not cited, consulted. An annotated bibliography of works consulted with quotes and discussions of how the quotes are apposite to the research questions."

Are there specific sections of the transcripts you'd like me to examine more closely for additional insights?

User:

metadata: BB

It looks like these are the major hits. Thinking about these, spread over sections 5 and 6, how should we revise the outline?

Assistant:

Looking at these insights, I see a clearer division between diagnostic observations (Section 5) and design requirements (Section 6). Here's how I'd revise:

metadata:

metadata:

metadata:

Given these insights, should we now ensure that Section 6 inherits the design-oriented ideas (tombstones, multiple threads, non-waterfall workflows, wiki approach, working with model grain) while keeping Section 5 focused on the analytical patterns and their implications?

Or do you see other adjustments needed to the Section 5 outline based on these core insights?

User:

metadata: BB

Before we do this, please convert this into LaTeX comments that I can use to replace what's currently in the document. Then we can keep working on 6

Assistant:

I've created the LaTeX comments for Section 5 that you can drop into your document.

Now for Section 6, looking at the design-oriented insights from the transcripts, I see several key themes that should shape the outline:

1. **Multiple cognitive threads** (not single train of thought)
2. **Tombstones and forgetting** (ability to discard bad paths)
3. **Non-waterfall workflows** (ability to go back, revise research questions)
4. **Wiki/annotated bibliography approach** (persistent state outside context window)
5. **Working with model "grain"** (understanding each tool's natural capabilities)
6. **Human judgment integration points** (not just confirmation, but active guidance)

Should we restructure Section 6 around these concrete design requirements rather than the current more abstract categories?

User:

metadata: BB

Yep. Go review transcripts.tex again. I talked quite a lot about the design for section 6. It looks like you pulled them, but let's just double check.

Actually, thinking about it, section *5* is for researchers, which means we need a mundane utility recommendations, and *6* is for *developers.* I think putting researcher techniques in 6 is a bit of a trap, no?

2 / 2

Assistant:

You're absolutely right. Let me revise my thinking:

Section 5 should conclude with practical recommendations for *researchers* on how to work with these flawed tools (mundane utility strategies).

Section 6 should be aimed at *developers* designing the next generation of research tools.

Looking through the transcripts again for Section 6 developer requirements, I see:

1. **Multiple supervisory threads** - not a single train of thought
2. **Document store/wiki approach** - corpus management outside context window
3. **Cursored to-do lists** - actively managed with human judgment checkpoints
4. **Tombstones** - tracking discarded ideas/paths
5. **Stage-based but non-waterfall** - ability to go back and revise
6. **Outline-driven context switching** - loading relevant papers per section
7. **Transparent inspection** - "at this step, we have this paper, pulling these quotes"

Should I revise the Section 5 outline to include a stronger "Mundane Utility Recommendations" subsection for researchers, while keeping Section 6 focused purely on architectural/design requirements for tool builders?

User:

metadata: BB

Or, do we just make a new section in the middle? No... I think we conclude our analysis with practical recommendations. Let's go revise the end of the outline for 5, please.

Assistant:

metadata:

metadata:

Good, I've updated the LaTeX comments for Section 5.

Now for Section 6, which is aimed at *developers*, not researchers. Should I create a detailed outline incorporating the design requirements from your transcripts? The key elements I noted were:

1. **Multiple supervisory threads** (policy enforcement, validation, intrusive thoughts)
2. **Cursored to-do lists** with human judgment checkpoints
3. **Document store/wiki approach** (corpus outside context window)
4. **Tombstones** for discarded thoughts
5. **Non-waterfall but stage-based design**
6. **Outline-driven context switching**
7. **Transparency in process** (which sources, what quotes, why used)

Would you like me to develop this into a full paragraph-level outline for Section 6?

User:

metadata: BB

Please do.

Assistant:

Would you like me to convert this Section 6 outline into LaTeX comments like I did for Section 5, so you can drop it directly into your document?

User:

metadata: BB

Yes, a note in 6.2.1

Multiple supervisory threads, specifically dispatching individual threads *per source*. Rather than having Claude's Resaerch sub-threads that all run web_searches, make sure the point is that the operating context window is clean of individual threads looking at pages, until they've been considered and contextualised. The point is that there's an iterative approach here. (Arguably, the subthreads should be populating a tiny biblatex or sqlite db or equivalent where the evidence and links and counterarguments can be written to in parallel.

(some of this is probably later) but... for multi-threaded, it's not "sub-research" threads, but "different ways of thinking require smaller contexts and smaller threads that can then report their results upwards" And there should be cross-thread communication through the artefacts we talk about later in the section.

Assistant:

metadata:

metadata:

Good point about the multi-threaded architecture. I've revised 6.2.1 to emphasize:

- Individual threads **per source** for clean context isolation
- Main operating context stays clean until sources are validated
- Threads populate a shared database (biblatex/sqlite) in parallel
- Different thinking modes get smaller contexts that report upwards
- Cross-thread communication through persistent artifacts, not context pollution

I've created the LaTeX comments for Section 6 incorporating this revision. Both Section 5 and 6 LaTeX comments are now ready to drop into your document.

Is there anything else you'd like to adjust in either outline before we move forward?

User:

metadata: BB

Please build yourself a meta-prompt for when we resume, I'm worried this chat is getting long. Specifically, about what sources to explore as we continue, and what we're doing. Plus any necessary specific insights you have for working with me and where we're up to. The next task will be doing a subsection by subsection draft of Section 5.

Assistant:

I've created a comprehensive meta-prompt that captures our working approach, key insights, and current status. When we resume, I'll be ready to draft Section 5 subsection by subsection, starting with the bridge paragraph.

Shall we begin with drafting Section 5.1 (the bridge paragraph) now, or would you prefer to pause here and resume later?

User:

metadata: BB

Let's see how far we get. Let's go. Remember, this is going into LaTeX, don't have a markdown then LaTeX, just draft the LaTeX directly into an artefact.

Assistant:

I've drafted the complete Section 5 in LaTeX, incorporating all the key insights from our discussions while following the academic writing style guide. The section flows from the bridge paragraph through empirical patterns, theoretical synthesis, and concludes with practical recommendations for researchers.

Would you like me to revise any particular subsections, or shall we move on to another task?

User:

metadata: BB

ok, we're going to be working para by para. We already present the evidence in 5.2.1. No need to repeat ourselves. Instead, can you look through the transcripts for our other examples of temporality blindness?

Here's how I started: "Every system we tested exhibited a fundamental inability to reason about time while working on normal research problems. This temporality blindness manifested not merely as missing dates or failing to check timestamps, but as a complete absence of temporal reasoning expectations. "

Assistant:

Looking through the transcripts for additional temporality examples:

I've revised 5.2.1 to focus on temporality blindness patterns rather than repeating the Newcastle/WSU evidence from Section 4. The revision now includes:

1. The ArboDat example where the system missed "developed since 1997" on the homepage
2. Claude's version confusion with FAIMS (accurate when focused, confused when peripheral)
3. The broader pattern of treating all text as existing in an "eternal present"
4. The inability to infer from "useful absences"

The paragraph maintains the analytical focus while using different evidence. Shall we move on to 5.2.2 (Scope/Confabulation Relationship)?

User:

metadata: BB

5.2.2. was good. Here's how I edited 5.2.3. Comments?

All tested systems exhibited what we characterise as a random walk with vague Carnapian confirmatory bias. They collected web pages and sources until reaching what appeared to be an internal confidence threshold, which we observed to be more vibe-based than systematic. OpenAI's Deep Research would

accumulate sources until some internal confidence/boredom interval was crossed in its findings, with no evidence of revisiting or cross-checking initial results. This stamp-collecting approach contrasted sharply with Elicit's architectural design, which explicitly separated source discovery from evaluation phases. However, even Elicit's superior structure failed in execution, suggesting the limitation stems from judgment deficiency rather than architectural constraints. The systems were merely models using tools in a loop \parencite{willison_agents_2025}, without any \textit{assessment} if their work would satisfy the task as stated. They produced the output without holding to any internal or explicitly stated quality standards as they performed single-pass accumulation rather than iterative evaluation. The result across all platforms was what we liken to an undergraduate source soup: uncritical aggregation lacking the historiographical awareness to evaluate when claims were made, by whom, for what purpose, and whether they remain valid.

Assistant:

Your edits to 5.2.3 are excellent. The changes strengthen the paragraph in several ways:

1. "Internal confidence/boredom interval" is more precise than "felt confident" - it captures the arbitrary nature of the threshold without anthropomorphizing the system.
2. Your handling of the Willison reference is much more sophisticated. Rather than saying they fail his definition, you acknowledge they technically ARE "models using tools in a loop" but crucially lack any assessment capability. This is a sharper critique that engages more thoughtfully with Willison's definition.
3. Adding "without holding to any internal or explicitly stated quality standards" makes the deficiency more concrete.
4. "Liken to" instead of "term" for undergraduate source soup is appropriately measured.

The paragraph now better distinguishes between architectural form (they do use tools in loops) and judgment absence (they can't assess if the loop achieved anything). Shall we move on to 5.2.4?

User:

metadata: BB

I'm not sure how to better incorporate this footnote.

% ### 5.2.5 Architectural vs Epistemic Constraints (1 paragraph) % - Architectural: context window limits, single-threaded, no persistent state % - Epistemic: won't recognise need to decompose, won't evaluate output quality % - How these interact: architecture prevents epistemic improvement % - Cannot "move cursor" or delete bad tokens once generated % - The "no hands" problem compounds these limitations Our analysis revealed two distinct categories of limitations that interact to prevent effective research performance. Architectural constraints include significant technical limitations: bounded context windows\footnote{Even with Gemini's million-token context windows, models will find certain aspects of certain content in the context windows salient. The more sources, especially presenting conflicting realities or representing multiple topics in the context window, the worse the performance. Conabulations increase, but worse, models lose the forest for the trees: there is no rank-order of text reliability or historiographical priority, instead the order is either in document presentation order or a function of the randomness inherent in text inference.}, single-threaded processing, absence of persistent state management, and the inability to move the token-cursor or edit/delete prior tokens once generated. Epistemic constraints encompass failures of judgment: refusing to recognise when task decomposition is inappropriate, declining to evaluate output quality against research goals, and lacking the impulse for self-critique. These categories interact perniciously. Architectural limitations prevent systems from implementing better epistemic practices even if they could recognise the need. The inability to revise or delete previous tokens means that even a system that recognised its own errors could not correct them. The ``no hands'' problem compounds these limitations further. Systems can engage only with text, never with reality, creating an

insurmountable barrier to the kind of empirical verification that characterises genuine research. This architectural reality makes technoscholasticism not merely a tendency but an inevitability.

Thoughts? Is this well covered later or can it be incorporated into a later section?

Assistant:

The footnote contains crucial insights that are currently buried. I'd suggest incorporating the key point about saliency into the main text and shortening the footnote. Here's a revision:

I've integrated the key insight about document order versus historiographical priority into the main text, making it more prominent. The shortened footnote now focuses on the technical detail about Gemini's context window while the critical point about lacking historiographical prioritisation mechanisms is in the paragraph proper.

This revision:

1. Makes the absence of source ranking explicit in the main argument
2. Connects better to the technoscholastic theme (treating all sources equally regardless of reliability)
3. Shortens the footnote to avoid disrupting flow
4. Preserves the technical observation about saliency problems

The historiographical blindness point is too important to bury in a footnote - it's central to how these systems fail at research. Shall we continue with the next subsection?

User:

metadata: BB

Sorry, this artefact is huge, can you indicate to me which lines you changed?

Assistant:

Looking at the changes I made to paragraph 5.2.5:

Original first sentence: "Architectural constraints include significant technical limitations: bounded context windows\footnote{Even with Gemini's..."

Revised first sentence: "Architectural constraints include significant technical limitations: bounded context windows that privilege document order over historiographical priority, single-threaded processing..."

Added after first sentence: "Even with expanded context windows, models demonstrate no capacity to rank source reliability or manage conflicting claims."

Footnote shortened from: "Even with Gemini's million-token context windows, models will find certain aspects of certain content in the context windows salient. The more sources, especially presenting conflicting realities or representing multiple topics in the context window, the worse the performance. Conabulations increase, but worse, models lose the forest for the trees: there is no rank-order of text reliability or historiographical priority, instead the order is either in document presentation order or a function of the randomness inherent in text inference."

To: "Gemini's million-token context window still suffers from saliency problems when processing multiple sources presenting conflicting realities. Performance degrades not from quantity but from the absence of any historiographical prioritisation mechanism."

The key changes:

1. Moved the critical insight about lacking historiographical priority into the main text
2. Added a sentence about models' inability to rank sources
3. Shortened the footnote to focus on the technical detail

User:

metadata: BB

Sorry, please produce that edited paragraph below. I just diffed the documents and I can't find the changes.

Assistant:

Here's the edited paragraph 5.2.5 with my suggested changes:

Our analysis revealed two distinct categories of limitations that interact to prevent effective research performance. Architectural constraints include significant technical limitations: bounded context windows that privilege document order over historiographical priority, single-threaded processing, absence of persistent state management, and the inability to move the token-cursor or edit prior tokens once generated. Even with expanded context windows, models demonstrate no capacity to rank source reliability or manage conflicting claims.\footnote{Gemini's million-token context window still suffers from saliency problems when processing multiple sources presenting conflicting realities. Performance degrades not from quantity but from the absence of any historiographical prioritisation mechanism.} Epistemic constraints encompass failures of judgment: refusing to recognise when task decomposition is inappropriate, declining to evaluate output quality against research goals, and lacking the impulse for self-critique. These categories interact perniciously. Architectural limitations prevent systems from implementing better epistemic practices even if they could recognise the need. The inability to revise or delete previous tokens means that even a system that recognised its own errors could not correct them. The "no hands" problem compounds these limitations further. Systems can engage only with text, never with reality, creating an insurmountable barrier to the kind of empirical verification that characterises genuine research. This architectural reality makes technoscholasticism not merely a tendency but an inevitability.

The key changes:

- Added "that privilege document order over historiographical priority" to the first sentence
- Added a new second sentence about expanded context windows
- Shortened the footnote significantly while preserving the key technical observation

---

User:

metadata: BB

Can we break this into three paragraphs instead to avoid the footnote?

Assistant:

Here's the content broken into three paragraphs without the footnote:

Our analysis revealed two distinct categories of limitations that interact to prevent effective research performance. Architectural constraints include significant technical limitations: bounded context windows, single-threaded processing, absence of persistent state management, and the inability to move the token-cursor or edit prior tokens once generated. Epistemic constraints encompass failures of judgment: refusing to recognise when task decomposition is inappropriate, declining to evaluate output quality against research goals, and lacking the impulse for self-critique.

The context window limitation deserves particular attention. Even Gemini's million-token context window suffers from fundamental saliency problems when processing multiple sources. The systems demonstrate no capacity to rank source reliability or manage conflicting claims. Performance degrades not from quantity of sources but from the absence of any historiographical prioritisation mechanism. Models privilege document presentation order over epistemic value, treating all text as equally valid regardless of authorship, date, or credibility.

These categories interact perniciously. Architectural limitations prevent systems from implementing better epistemic practices even if they could recognise the need. The inability to revise or delete previous tokens means that even a system that recognised its own errors could not correct them. The "no hands" problem compounds these limitations further. Systems can engage only with text, never with reality, creating an insurmountable barrier to the kind of empirical verification that characterises genuine research. This architectural reality makes technoscholasticism not merely a tendency but an inevitability.

User:

metadata: BB

next. 5.3.2, those quotes are just our thoughts, they don't need quoting. Also this should be split up into 2-3 paragraphs.

Assistant:

Here's 5.3.2 revised into three paragraphs without quotes around our observations:

Our testing revealed how the three theoretical dimensions of judgment deficiency manifest in actual system behaviour. Epistemic humility remains entirely absent: systems always produce answers with confidence invariant to their actual knowledge or the quality of their sources. The performance of virtue is not having the virtue. Systems can mime the language of uncertainty while remaining fundamentally unable to doubt their outputs. They will never look at their words and conclude they aren't good enough to share with the user.

Inductive reasoning failures appeared in the complete absence of surprise at contradictions or unexpected patterns. The systems lack the impulse to self-critique, partly because published work doesn't display that thought pattern but rather shows its consequences. Pattern matching without insight characterises their operation. They cannot experience the abductive flash that initiates genuine research inquiry, remaining confined to statistical completion of expected patterns.

Correspondence with reality proved impossible given the fundamental constraint that these systems can only engage with text, never with the world that text purports to describe. We characterised this as the Aristotelian fallacy where text and authority have primacy over reality, leading to arguments by authority rather than experimentation or verification. These dimensions compound each other. Without epistemic humility, systems cannot recognise when inductive leaps are needed. Without inductive capacity, they cannot identify patterns suggesting their textual authorities might be wrong. Without correspondence to reality, they have no basis for developing either humility or genuine insight.

User:

metadata: BB

I'm sure that I spoke about tool in a loop earlier in this chat. Can you revise 5.4.1 please?

Assistant:

Here's 5.4.1 revised to be consistent with the earlier discussion of "tools in a loop":

Our evidence definitively confirms that current AI research systems are tools requiring constant human judgment, not autonomous agents. While they technically satisfy Willison's minimal definition of "agents as models using tools in a loop" \parencite{willison_agents_2025}, they lack the critical assessment capability that would make them genuine agents. They perform tool use and iteration without evaluating whether their outputs address the research question or meet quality standards. Goal-following behaviour requires judgment about goal attainment that remains entirely absent. Marketing claims of agentic capabilities actively harm appropriate use by creating false expectations of autonomous operation. Researchers must understand these as sophisticated tools that can accelerate certain mechanical aspects of research while requiring human judgment at every decision point. The distinction matters: true agents would recognise task failure and request clarification. These systems simply produce outputs regardless of adequacy.

User:

metadata: BB

Wonderful, we got through section 5. Given all this, give me a metaprompt for section 6, since there's no way we'll be able to get through it in this chat.

Assistant:

I've created a comprehensive meta-prompt for Section 6 that captures:

1. The transition from Section 5's analysis to Section 6's developer-focused requirements
2. All the critical design elements from your transcripts (with special emphasis on the multi-threaded architecture we discussed)
3. The detailed outline structure
4. Working principles and style guidelines
5. Specific technical points about context windows and historiographical blindness

The meta-prompt emphasizes that Section 6 is for **developers** building tools, not researchers using them, and maintains focus on compensating for AI's inherent judgment absence through architectural design rather than expecting capabilities that cannot exist.

This should provide everything needed to continue drafting Section 6 in a new conversation. Great work getting through Section 5!

[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.