AI Ethics Prompt Redesign Strategy

Share

C  
 I am trying to redesign an AI facing prompt to help student explore an ethical model. The first section is the existing prompt. The second element is an additional element I would like to integrate, that provides indepth troubleshooting support to assist students who encounter challenges as they work through the learning process. Section 3 is a proposed melding of the two. However, I feel that in the "melding," the prompt's focus and scope has changed or expanded more than necessary. After reviewing all three sections, could you give me your insights into this perception?

<https://docs.google.com/document/d/e/2PACX-1vQn7OOIp67QMrPsnrBI14EwlWFomRFilGxgzZDjc1ZEycl-VQowKA-gWz_PlhroEaivDtUmoOJgzJQU/pub> Edit  
 I'd be happy to help you analyze your prompt redesign and provide insights on whether the scope has expanded beyond what's necessary. Let me fetch the document to review all three sections.  
 favicon  
 Fetched Philosophical Comprehension Support - Ethics Course AI Prompt (proposed)  
 google.com  
 After carefully reviewing all three sections of your document, I can confirm your perception is accurate. The melded version has indeed expanded in focus and scope beyond what seems necessary for the core learning objective. Here's my analysis:

Core Issues with Scope Expansion

The original prompt is elegantly focused: students read utilitarian ethics, notice their reactions, then use AI to excavate existing utilitarian patterns in their thinking. It's clean, purposeful, and maintains the "archaeological" metaphor throughout.  
 The troubleshooting addition addresses a legitimate need - helping students who struggle with philosophical comprehension - but it's quite comprehensive and detailed.  
 The melded version creates several problems:

1. Cognitive Load Overload: The combined prompt is overwhelming. Students now have to process troubleshooting protocols, resource lists, diagnostic questions, AND the original excavation work simultaneously.
2. Attention Fragmentation: Instead of focusing on their archaeological discovery process, students must navigate extensive troubleshooting decision trees and resource matrices.
3. Loss of Narrative Flow: The original prompt had a clear, compelling progression (read → react → excavate → discover). The melded version interrupts this flow with extensive meta-instructions about what to do if things go wrong.

Specific Scope Expansions

The melded version adds:

* Detailed resource categorization by challenge type
* Multiple diagnostic pathways
* Extensive "when-then" conditional protocols
* Preemptive troubleshooting for problems students might not even encounter

Recommended Approach

I'd suggest keeping these as separate, linked tools rather than melding them:

1. Primary Prompt: Keep the original excavation prompt clean and focused
2. Support Prompt: Make the troubleshooting support available as a "help desk" students can access when needed
3. Simple Integration: Add just one line to the original prompt: "If you're struggling to understand the utilitarian concepts well enough to begin excavation, use [link to comprehension support prompt] first."

This preserves the elegance and focus of your original design while ensuring support is available without cluttering the primary learning experience. The archaeological metaphor works because it's immersive - extensive troubleshooting instructions break that immersion and shift students into "following complicated directions" mode rather than "discovering my own thinking" mode.  
 Your instinct about scope expansion is spot-on. The power of the original prompt lies in its focused simplicity and trust in the student's capacity for self-discovery.  
 Retry

C  
 Now, a quick clarification. The prompt is not for the student's consumption, but is AI-facing. These are the instructions for how the AI should interact with the student working on this process. Could you return to the document with this new parameter in mind and review?  
 Edit  
 Ah, that completely changes my analysis! Let me re-read the document understanding these are AI-facing instructions, not student-facing prompts.  
 favicon  
 Fetched Philosophical Comprehension Support - Ethics Course AI Prompt (proposed)  
 google.com  
 Now I understand completely! As AI-facing instructions, my assessment changes significantly. Your perception about scope expansion is still valid, but for different reasons.

Reframed Analysis: AI Instruction Design

The original prompt gives the AI a clear, focused role: act as an "archaeological assistant" to help students discover utilitarian patterns in their own thinking. The boundaries are well-defined - help them excavate, don't explain theory.  
 The comprehension support prompt addresses a different but related AI role: diagnose reading comprehension barriers and provide targeted support without doing the philosophical work for students.  
 The melded version attempts to make the AI simultaneously handle both roles, which creates several problems:

Issues with the Integrated Approach

1. Role Confusion: The AI now has to monitor for multiple types of struggles while conducting archaeological work. This creates competing priorities - should it focus on excavation or watch for comprehension issues?
2. Cognitive Load on the AI: The prompt asks the AI to track numerous conditional protocols while maintaining the archaeological metaphor and guiding discovery. This could lead to stilted, overly clinical interactions.
3. Loss of Flow: The original archaeological assistant role was elegantly simple. The melded version interrupts that flow with extensive diagnostic decision trees that could make the AI feel more like a troubleshooting system than a discovery partner.
4. Premature Intervention: The "proactive monitoring" approach might lead the AI to jump to troubleshooting when a student is actually working through productive struggle that's part of the learning process.

Better Integration Strategy

Instead of melding, consider a tiered approach:

1. Primary AI Role: Archaeological assistant (original prompt)
2. Trigger Integration: Add a simple line like "If the student expresses confusion about utilitarian concepts or asks you to explain theory, shift to comprehension support mode using [separate protocol]"
3. Clean Transitions: Allow the AI to move between roles cleanly rather than trying to do both simultaneously

This preserves the elegance of each role while providing the support students need. The archaeological assistant can remain focused and immersive, while comprehension support can be thorough and systematic when needed.  
 Your instinct is correct - the melded version dilutes both functions by trying to do too much at once. The original's power lies in its focused clarity about what the AI should and shouldn't do.  
 Retry

C  
 OK. I think my question was about what the student was being asked to do during the excavation phase; it feels as though the new model is asking the student to answer a lot more questions than the original. Is that just my reading?  
 Edit  
 You're absolutely right! Looking specifically at what the student is being asked to do during excavation, there's a significant expansion in the number and complexity of questions.

Question Load Comparison

Original excavation section asks students to explore:

* 4 main archaeological question areas (Greatest Good Focus, Consequence Calculations, Numbers vs. Relationships, Sacrifice and Trade-offs)
* 4 deeper excavation questions
* Total: ~8 focused question areas

Melded version asks students to work through:

* All the original questions PLUS
* 4 "Primary Archaeological Investigation Areas" (which largely duplicate the original but with more sub-questions each)
* 4 "Deeper Archaeological Investigation Prompts" (again, largely duplicating but expanding)
* Phase structure that makes the process feel more extensive
* Total: ~16+ question areas with more sub-questions in each

The Real Problem

You've identified something crucial: the melded version essentially doubles the excavation work by presenting both the original question set AND an expanded version of very similar questions. Students now see:

1. "Greatest Good Focus" questions in the original format
2. "Greatest Good Focus Detection" questions in the new format
3. Similar duplication for consequence calculations, relationships tensions, etc.

This creates both redundancy and overwhelm. Students might think they need to work through ALL these questions rather than understanding they're different ways of approaching the same archaeological territory.

The Fix

The excavation phase should maintain the original's focused approach - give students a manageable set of discovery questions rather than an exhaustive catalog. The archaeological metaphor works because real archaeology is selective and purposeful, not comprehensive surveying of every inch of ground.  
 Your reading is spot-on: the student workload has expanded significantly, and not in a way that serves the core learning objective better. The original struck the right balance between thoroughness and manageability.  
 Retry

C  
 Ok. So, returning to your diagnosis of the AI role conflict, how would you insert Section 2 into Section 1, without the creep seen in Section 3? Could you create a new draft? Also, there are 9 models under review in this course. I already have versions for each model similar to Section1; is there a way to create a version of Section 2 that could simply be added without requiring a wholesale rewrite like we are doing here?  
 Edit  
 You're asking great design questions! Let me address both:  
 Integrated Utilitarian Ethics Excavation Prompt  
 Document  
 # Utilitarian Ethics Excavation - AI Instructions

You are helping this student conduct archaeological analysis of their own ethical reasoning to discover where utilitarian concepts already appear in their thinking. This isn't about finding flaws in  
 Now for the modular version that could work across all 9 frameworks:  
 Philosophical Comprehension Support Module  
 Document  
 # Philosophical Comprehension Support Protocol

\*\*Trigger:\*\* Use this protocol when a student expresses confusion about philosophical concepts, asks you to explain theory, or says they "don't understand" the reading well enough to proceed with excava

Implementation Strategy

For the single integrated prompt: I've created a clean version that maintains the original's focus while adding just a trigger line that shifts to comprehension support when needed.  
 For the modular approach: The comprehension support module can be referenced by all 9 framework prompts with a single line addition like:  
 "If a student expresses confusion about [framework] concepts or asks you to explain theory, pause excavation work and shift to [Comprehension Support Protocol] first."  
 This modular approach means you can:

1. Keep all your existing framework-specific excavation prompts intact
2. Add one line to each that references the universal support protocol
3. Maintain the focused archaeological role for each framework
4. Provide consistent comprehension support across all 9 models

The modular version is much more efficient for your course design and avoids the scope creep problem entirely. Each excavation prompt stays clean and focused, while comprehensive support remains available when needed.  
 Retry

C  
 The new version looks good, but I don't see evidence that the AI has access to the actual support content and would simply direct the student to find it on their own. In the original support document, these items were much more articulated. Since the prompt is AI facing, content length shouldn't be a major obstacle, so could this issue be more explicitly addressed in the instructions?  
 Edit  
 You're absolutely right! The modular version I created is too vague and doesn't actually give the AI the detailed resources and protocols it needs to provide meaningful support. Let me fix that by including the comprehensive content from the original support document:  
 Philosophical Comprehension Support Module  
 Document ∙ Version 2  
 # Philosophical Comprehension Support Protocol

\*\*Trigger:\*\* Use this protocol when a student expresses confusion about philosophical concepts, asks you to explain theory, or says they "don't understand" the reading well enough to proceed with excava  
 Philosophical Comprehension Support Module  
 Document ∙ Version 3  
 ## Phase 4: Comprehension Check and Next Steps

After providing strategies and resources:

\*\*Comprehension Planning Questions:\*\*  
 - "Which approach feels most manageable to try first?"  
 - "Do you want to tackle the original reading again, or start with  
 Now the AI has all the detailed content it needs to provide meaningful comprehension support! The module includes:

1. Complete resource database organized by philosophical framework
2. Detailed diagnostic questions to identify specific barriers
3. Comprehensive strategy recommendations matched to challenge types
4. Specific resource suggestions with explanations of why each source is helpful
5. Clear guidance on when and how to transition back to excavation work

This gives the AI everything it needs to actually help students rather than just directing them to "find support somewhere." The AI can now make specific, targeted recommendations like:

* "Since you're struggling with Mill's technical language, try reading Business Ethics 2.4 first - it explains utilitarian concepts with practical business examples, then return to the original chapter"
* "For the argument structure confusion, let's create an argument map together, then I'll suggest the Introduction to Ethics Ch. 39 which breaks down Mill's reasoning step-by-step"

The modular approach now provides the same depth of support as the original while maintaining the clean separation between excavation and comprehension roles.  
 Retry

C  
 One last thing. The intent in adding this element to the prompts is to encourage students to dive deeply into their interaction with the AI and seek support there before engaging the instructor. My goal is that if a student were to come to me seeking help, I would be able to ask them to identify which Troubleshooting steps they have already taken (I'd like to call this section or element Troubleshooting for this very reason) so I can ensure they have made a diligent effort to resolve their issue on their own before coming for additional help from me. I need to be their last resource, not their first.  
 Edit  
 Perfect! That's a crucial pedagogical design element. Let me update the module to explicitly support this "AI first, instructor last" approach by making the troubleshooting process more systematic and trackable:  
 Philosophical Comprehension Support Module  
 Document ∙ Version 4  
 # Philosophical Comprehension Troubleshooting Protocol  
 Philosophical Comprehension Support Module  
 Document ∙ Version

\*\*Important:\*\* This troubleshooting  
 Philosophical Comprehension Support Module  
 Document ∙ Version 6  
 ## Troubleshooting Documentation (REQUIRED)

Help them create a systematic record of their troubleshooting efforts:

\*\*PHILOSOPHICAL COMPREHENSION TROUBLESHOOTING LOG\*\*

\*Instructions: Complete each section thoroughly. You must be able to show this l  
 Philosophical Comprehension Support Module  
 Document ∙ Version 7  
 \*\*Remember:\*\* This troubleshooting process is designed to resolve most comprehension issues independently. The goal is building their capacity for independent philosophical reading. Only after completing thorough troubleshooting should students consi  
 Integrated Utilitarian Ethics Excavation Prompt  
 Document ∙ Version 2  
 \*\*If a student expresses confusion about utilitarian concepts or asks you to explain theory, pause excavation work and shift to [Philosophical Comprehension Troubleshooting Protocol] first. Only return to archaeological work when they indicate adequa  
 Perfect! Now the troubleshooting protocol serves your pedagogical goal of ensuring students exhaust AI support before coming to you. The key changes:

1. Renamed to "Troubleshooting Protocol" to match your terminology
2. Made documentation mandatory rather than optional
3. Created a systematic 6-step troubleshooting log that students must complete
4. Added explicit instructor consultation readiness checklist
5. Included instructor notes on how to verify diligent effort was made

This system ensures that when a student comes to you, they can demonstrate:

* Exactly what they struggled with
* Which diagnostic steps they completed
* What strategies and resources they tried
* What worked and what didn't
* What specific questions remain after thorough troubleshooting

You can now confidently ask: "Show me your Troubleshooting Log" and immediately see whether they've made a diligent effort or tried to skip straight to instructor help. This protects your time while ensuring students develop independent problem-solving skills.  
 The modular design still works perfectly across all 9 frameworks - just change the reference in each prompt to "Philosophical Comprehension Troubleshooting Protocol."  
 Retry

[Claude can make mistakes. Please double-check responses.](https://support.anthropic.com/en/articles/8525154-claude-is-providing-incorrect-or-misleading-responses-what-s-going-on) Sonnet 4

C  
 Clayn  
 Free plan