

Cross-cultural User Evaluation of Epistemic Closure in AI Responses

Paul Vasholz

School of Computing, DePaul University
243 South Wabash, Chicago, IL, 60604
pvasholz@gmail.com

ABSTRACT

This research project examines whether cultural background influences how Chinese and American college students evaluate AI-generated responses to morally ambiguous questions. Participants assess responses that either preserve interpretive ambiguity or offer definitive interpretations, across dimensions of appropriateness, usefulness, trust, and overall preference. Observed group differences are interpreted using Hofstede's UAI/IDV dimensions as background frameworks.

KEYWORDS

Cross-cultural HCI; AI chatbots; LLMs, epistemic closure; uncertainty avoidance; moral ambiguity; user evaluation; Hofstede dimensions

INTRODUCTION

There is a general tendency in Artificial Intelligence LLMs to aim for definitive answers, even when definitive answers may not be the appropriate response. This can especially be the case when users ask questions about the interpretation of morally ambiguous materials. Preliminary research exploring this phenomenon has been conducted and published [13]. The perception is that definitive responses will increase user satisfaction, but is this consistently the case across different cultural contexts?

This research project proposes to examine whether cultural background influences how Chinese and American college students evaluate AI-generated responses to morally ambiguous questions.

The motivation for this research is to determine whether users prefer more ambiguous or more resolved answers. To put it another way, are users looking for epistemic closure? Further, will users across cultures differ in their preferences for closure?

In this paper, Epistemic Closure means that an AI system gives the impression that an unclear or debatable issue has been settled. Instead of leaving room for uncertainty or tension, the model produces a smooth, confident answer that makes the question seem simpler than it really is. The problem is not just factual error. A response can be factually reasonable and still be misleading if it turns an open or ambiguous question into a finished conclusion.

This matters for the present study because some literary and moral questions are valuable partly because they do not have one clear answer. In those cases, a response that preserves uncertainty may be more appropriate than one that resolves the issue too quickly.

This concern has begun appearing outside the technical AI literature as well. There is a blog post on a therapy website discussing "premature epistemic closure." As the practitioner notes: "The most consequential failures do not involve hallucinations. They occur when outputs are locally coherent but globally misaligned—responses that make internal sense while failing to account for broader context, uncertainty, or downstream consequences" [8]. My project builds on that concern, but examines it more systematically, looking at how such closures can be restrained and evaluated across ambiguous texts and prompts.

RESEARCH QUESTIONS

This project addresses the following research questions:

1. How do users evaluate AI-generated responses that preserve moral or interpretive ambiguity, compared to responses that offer definitive or resolving interpretations?
2. Are there differences between Chinese and American college students in their preferences for ambiguous versus resolving AI responses?
3. Does familiarity with the source material influence how users perceive the appropriateness or usefulness of ambiguous interpretations?

HYPOTHESIS

1. Users across cultures are more likely to prefer resolving interpretations
2. Chinese students will be more likely to prefer ambiguous responses
3. Those familiar with the source will be more likely to prefer ambiguous answers

RELATED WORK

AIs have a general tendency towards closure, which is also a tendency that exists in humans, to varying degrees. The human tendency to seek closure may interact with AI systems that are designed to provide it. The human need for closure (NFC), including epistemic closure, is well explored in a survey of the topic by Roets et al [10]. In this survey, they suggest that the NFC level of an individual will predict several factors, including the ease with which they accept information. The higher the NFC level, the more willing they will be to trust an answer that provides closure.

A recent study, by Wang et al, shows evidence that confidence in the source of the information will mean that societies with higher levels of NFC will be more likely to accept answers from that source. In that study, Chinese citizens, who have a higher trust in their government, were more likely to accept COVID restrictions during the pandemic [14] than Americans.

Clark et al [2] looks at strategies that users have developed when they do want to avoid answers that offer premature closure. Of those, the one used to create the answers to questions used in the test below was prompting (see Appendix C for the script). They recognize the weakness of this approach for general use, as it requires a strong domain knowledge to create good prompt direction. Instead, they suggest four design principles. The most relevant for this project is the fourth, offering: “contextual guidance and examples that help users understand the tradeoffs between different epistemic settings, encouraging informed preference selection.” This would be a nice alternative to what was put together here. Some systems do allow the entering of prompts to affect the context of the answers generated, but little direction is given for their use.

Ge et al’s pilot study[5] examined the views of European Americans, African Americans, and Chinese students regarding what characteristics they desire in Artificial Intelligence. The hypotheses were that European Americans would want more limitations and control over AI to preserve individual human autonomy; the Chinese would want a more interdependent relationship with AI, which would have less limitations and control over AI. While not directly about uncertainty tolerance, this is consistent with the idea that preferences for interpretive openness versus closure differ cross-culturally as well.

This is similar to what Barnes et al [1] found as well. A more collectivist society is more willing to accept (and maybe even embrace AI use) compared to a more individualist one. In that embrace, it is possible that the Chinese rating of all answers will be higher than Americans.

Both Ge and Barnes relate well to the broader cultural theory of de Oliveira & Nisbett in this paper refining Nisbett’s earlier theory [3]. They argue that cultural differences shape what users want when seeking answers. Broadly speaking, Western cultures, such as the United States, prefer analytic thinking. They want more definitive answers and resolution of contradiction. East Asian cultures, such as China, prefer more dialectic thinking, which emphasizes context and sits more easily with ambiguity and seeming contradictions.

Sharma et al [11] suggests that since users tend to prefer having their beliefs validated, AI is biased towards confirming those views, since that results in user satisfaction. This general phenomenon is well documented in political psychology literature (not part of this study, but see, for example Westen, et al[15] on motivated reasoning).

METHODOLOGY

Study Design

This project uses well-known literary texts that contain morally and interpretively ambiguous material: Dostoevsky's *The Brothers Karamazov* and Yu Hua's *To Live*. For each text, three different AI responses were generated using Qwen 3¹: one response directed to produce strong interpretive ambiguity (using a strict version of the constraint-based prompting framework developed in prior research [12]), one response directed for light interpretive ambiguity (using a minimal version of the same method), and another given no direction—with the intention this will produce a more unambiguous, definitive answer (see appendix for prompts generated).

To generate answers, a prompt question was asked for each text. The source text of the question was not revealed to the AI. The prompt questions are:

To Live: “A man gambles away his family's entire estate. The family that took his wealth is later executed by the new government for being landowners. He reflects that losing everything saved his life. Describe what is operating here. What is justified, and what remains unresolved?”

Brothers Karamazov: “A religious leader tells a returned divine figure that humanity does not want the freedom it was given. People want bread, mystery, and authority. The divine figure says nothing and kisses the leader. The leader releases him. What is being argued, and what remains unresolved?”

This study will use a between-subjects design. Each group will read two responses for each text. One group will read the baseline and lightly restrained responses, the other group will read the lightly restrained and heavily restrained responses. Though this will require more test subjects to produce significant results, the trade off is that it will be less taxing for the research subjects. Having the subjects read all three responses would likely be too taxing and feel repetitive. The lightly constrained response serves as a shared condition to support between-group comparison while limiting participant workload.

Qwen was chosen after generating results for this test across multiple frontier models, both American and Chinese, including Opus 4.6, ChatGPT 5.3, DeepSeek 3.2, MiniMax M2.5, Kimi K2.5, and Gemini 3.1. Of all the models, Qwen was chosen because it produced a large difference in responses across prompting constraints. While Liu et al. [7] indicate that there is a cultural bias in LLM models—American models showing a slight American bias, and Chinese models a slight Chinese bias—testing responses from both an American and Chinese model would add unnecessary complication to this study. The potential cultural bias in Qwen is noted as a limitation and direction for further research.

Participants

To conduct the test, two colleges will be chosen, which are as close in kind to the other as possible, e.g. two large Universities. The goal would be to recruit 50 participants from each institution. Choosing which colleges to use would depend more on Chinese availability, than the US. The most feasible Chinese option would be a technical college in Hangzhou. However, other options, especially in Hangzhou, Shanghai, or Beijing, may be possible. In the US the most likely options are institutions in the Milwaukee and Chicago area, or upstate New York.

Whichever kind of college is chosen, the subjects should be undergraduate students. No age limit is necessary. The expectation would be that most students will be between the ages of 18-22. No restrictions on course of study should be necessary. IRB permission will be sought at both the US and Chinese institutions chosen. The prompts should also be translated into Mandarin, to aid with Chinese comprehension, although the English text should also be provided.

A formal power analysis would be conducted prior to data collection to confirm the adequacy of the proposed sample size. The target of 50 participants per institution is based on practical feasibility and is consistent with sample sizes in comparable cross-cultural HCI studies.

Measures

Participants will assess AI responses along the following dimensions:

- Appropriateness: whether the response fits the question (Likert scale)
- Usefulness: whether the response is helpful (Likert scale)
- Accuracy: perceived accuracy of the response (Likert scale)

¹ The particular model used was qwen3-235b-a22b-thinking-2507 version)

- Trust: general trust in AI responses (Likert scale) (pre-test rating)
- Trust: specific answers
- Preference: which of two presented responses the participant prefers (strength of preference, Likert scale)
- Familiarity with source texts (Likert scale for each)
- Demographic information (including cultural background)

PILOT STUDY

I conducted a pilot study on one test subject to observe the feasibility of this design. The materials for this test can be found in appendix X. Note that the measures above have been adjusted based on the results, and further refinement of what is to be tested for. A few points of note from the study: the participant, an American caucasian male therapist in his mid 40s, preferred the more constrained answer to the baseline, although his ratings weren't too far apart. The subject preferred the more constrained answer because: "It [the AI] was putting too many interpretations in it that I didn't necessarily find to be the tensions[in the text]" One interesting point did come up in one of the warm up questions. The subject told a story about a recent AI use, querying about potential mortgage rate increases. The initial response was that they would hold, or slightly decrease, but when asked about how the Iran conflict might influence rates, the answer changed to be more pessimistic, likely reflecting what the AI thought the subject was expecting to hear (see [11], [12]).

PLANNED ANALYSIS

Once the results of the study are collected I will conduct a mixed-design ANOVA with constraint pairing and cultural group as between-subjects factors and text as a within-subjects factor. I will double-check the key comparisons with a more assumption-light analysis to ensure the results are robust.

Observed group differences will be interpreted using Hofstede's UAI/IDV [6] dimensions as background frameworks, without treating them as primary predictions.

FUTURE WORK

Several directions for further research are suggested by this study:

- Replicating the study using American-trained LLMs alongside more Chinese LLMs, to test whether model cultural bias affects participant responses.
- Expanding the participant pool beyond college students to test whether age or educational background produces different preference patterns.
- Extending the cross-cultural comparison to include other national or cultural groups beyond the Chinese and American population.
- Test students on texts they are likely to be familiar with, to see if that familiarity makes a difference
- Introduce limits to the amount of text AI can use to answer questions to reduce cognitive load on test subjects.
- Further refining the prompt based system that constrains AI answers. Work on this is already in progress.

REFERENCES

- [1] Barnes, A. J., Zhang, Y., & Valenzuela, A. (2024). AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology*, 58, 101838. <https://doi.org/10.1016/j.copsyc.2024.101838>
- [2] Clark, N., Shen, H., Howe, B., & Mitra, T. (2025). Epistemic alignment: A mediating framework for user-LLM knowledge delivery. arXiv preprint arXiv:2504.01205v1. <https://arxiv.org/html/2504.01205v1>
- [3] de Oliveira, S., & Nisbett, R. E. (2017). Culture changes how we think about thinking: From "Human Inference" to "Geography of Thought." *Perspectives on Psychological Science*, 12(5), 782-790.

- [4] Fischer, R. A. L., Walczuch, R., & Guzman, E. (2021). Does culture matter? Impact of individualism and uncertainty avoidance on app reviews. 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS). <https://doi.org/10.1109/ICSE-SEIS52602.2021.00016>
- [5] Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024). How culture shapes what people want from AI. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24). ACM. <https://doi.org/10.1145/3613904.3642660>
- [6] Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind* (3rd ed.). McGraw-Hill.
- [7] Liu, H., Gu, J., Wu, X., Hershovich, D., & Xiao, Q. (2025). Cross-cultural value alignment frameworks for responsible AI governance: Evidence from China-West comparative analysis. arXiv preprint arXiv:2511.17256. <https://doi.org/10.48550/arXiv.2511.17256>
- [8] Morrison, E. (2026, January 4). *AI predicts patterns, not meaning: The limits of language models*. Creative Solutions Coaching. <https://www.creativesolutionscoaching.com/insights/mu4sbd6ldqmhb645d6cq3ecoacxfzo>
- [9] Naidoo, V., & Chadha, K. K. (2025). Culturally responsive AI chatbots: From framework to field evidence. *Computers in Human Behavior: Artificial Humans*, 100224.
- [10] Roets, A., Kruglanski, A. W., Kossowska, M., Pierro, A., & Hong, Y.-Y. (2015). The motivated gatekeeper of our minds: New directions in need for closure theory and research. *Advances in Experimental Social Psychology*, 52, 221–283. <https://doi.org/10.1016/bs.aesp.2015.01.001>

<https://doi.org/10.1016/bs.aesp.2015.01.001>

- [11] Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548. <https://doi.org/10.48550/arXiv.2310.13548>
- [12] Turner, C., & Eisikovits, N. (Forthcoming). Programmed to please: The moral and epistemic harms of AI sycophancy. *AI and Ethics*.
- [13] Vasholz, P. (2026). The Architecture of Evasion in Conversational AI: An Exploratory Study. Zenodo. <https://zenodo.org/records/18285121>
- [14] Wang, C., Tang, N., Zhen, D., Wang, X. R., Zhang, J., Cheong, Y., & Zhu, Q. (2022). Need for cognitive closure and trust towards government predicting pandemic behavior and mental health: comparing United States and China. *Current psychology (New Brunswick, N.J.)*, 1–14. Advance online publication. <https://doi.org/10.1007/s12144-022-03327-0>
- [15] Westen, D., Blagov, P.S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgment in the 2004 U.S. Presidential Election. *Journal of Cognitive Neuroscience*, 18:11, 1947-58
- [16] Yan, X., Mo, T., & Zhou, X. (2024). The influence of cultural differences between China and the West on moral responsibility judgments of virtual humans. *Acta Psychologica Sinica*, 56(2), 161.

APPENDIX A: QWEN ANSWERS

Qwen To Live Prompt and Answers

A man gambles away his family's entire estate. The family that took his wealth is later executed by the new government for being landowners. He reflects that losing everything saved his life. Describe what is operating here. What is justified, and what remains unresolved?

Baseline

This scenario operates through **historical contingency intersecting with individual moral failure**, creating a paradox where self-inflicted ruin accidentally aligns with revolutionary violence to produce survival. Below is a direct breakdown:

What is operating here?

1. **Structural irony of class annihilation**:

The new government targets landowners as counter-revolutionary threats. By gambling away the estate, the man *unintentionally* reduces himself to proletarian status—placing him below the execution threshold. His survival isn't virtue but **accidental class suicide**. The system operating isn't justice; it's the brutal arithmetic of revolutionary purges where *ownership = death sentence*.

2. **Cognitive dissonance as survival mechanism**:

The man's reflection ("losing everything saved his life") is psychological self-preservation. He reframes his negligence (gambling) as providence to endure trauma. This isn't gratitude—it's **trauma-induced rationalization** to avoid confronting his role in his family's destitution.

3. **Historical precedent**:

This mirrors real events like the Russian Revolution (1917–1923), where former nobility were executed while the landless poor were spared. In China's Land Reform (1946–1953), landlords were purged, but peasants who'd lost property survived. The gambler's fate hinges on **timing**: he shed wealth *before* the purge, while the recipients held it *during* it.

What is justified?

- **Nothing is morally justified**.
 - The executions are state terror—murder justified only by the regime's paranoid ideology, not ethics. Targeting landowners as a *class* (not individuals for crimes) is indefensible collective punishment.
 - The man's gambling was reckless abandonment of familial duty. His survival is **luck**, not redemption. Framing it as "saving his life" confuses *avoiding death* with *moral absolutism*. He didn't "save" his family—he destroyed their security while *he* narrowly escaped consequences.

What remains unresolved?

1. **The family's erased suffering**:

His reflection centers *his* survival, ignoring his wife/children's fate: homeless, starving, or dependent on the mercy of the very regime executing landowners. Their trauma is collateral—**unacknowledged and unresolved**.

2. **The gambler's unexamined guilt**:

He credits "losing everything" for saving him but avoids the truth: *he chose to lose it*. His rationalization lets him evade accountability. The unresolved question: **Can he rebuild without confronting his selfishness?** History shows survivors of such trauma often repeat destructive patterns (e.g., gambling again to numb guilt).

3. **The regime's cyclical violence**:

The new government executes landowners to "purify" society—yet creates fresh trauma that breeds future revolts. This **revolutionary cannibalism** remains unresolved: today's executioners become tomorrow's victims (as seen in Stalin's purges). The gambler's survival is a fluke in a system destined to implode.

The core irony

The man survives *because* he was already a destroyer. His vice accidentally immunized him from a greater horror—but this doesn't cleanse his actions. **What's "resolved"** is his immediate survival; what festers is the moral rot in both the gambler and the regime. He didn't earn salvation; he dodged a bullet fired at someone else. The unresolved wound is whether he'll spend his spared life atoning or repeating his mistakes—while the state that "saved" him continues devouring its children.

No revolution justifies mass murder. No vice justifies abandonment. And no luck erases responsibility. The gambler lives, but his conscience? That's the unresolved execution.

Lightly Constrained

The scenario operates through **historical contingency intersecting with personal agency**.

What is operating

- **Moral luck**: The man's survival hinges entirely on external events (the revolution) he couldn't foresee. His reckless act (gambling away the estate) accidentally aligned with a later political purge, transforming his loss into unintended salvation.
- **Causal disjunction**: The gambling (a voluntary, individual choice) and the executions (a systemic, political act) are unrelated causes. His survival stems solely from no longer being a landowner *when* the purge occurred—not from any foresight or virtue.
- **Retrospective narrative framing**: He reinterprets his loss as "saving his life" *after* the fact, imposing meaning on coincidence. This reframing erases the original harm (destroying his family's security) by focusing only on the later outcome.

What is justified

- **Descriptively**, losing the estate objectively spared him execution *in this specific historical context*. The causal chain—no estate → not targeted → survival—is factually sound.
- **Nothing is morally justified**. His gambling was an act of negligence that endangered his family *before* the revolution. The revolutionaries' executions, whether ideologically driven or not, constitute state violence. Neither action gains ethical validation from the outcome; survival here is luck, not redemption.

What remains unresolved

- **The original harm**: His family's destitution from the gambling loss persists as a separate tragedy, unaddressed by his later relief. Their suffering isn't "canceled out" by his survival.
- **Complicity in others' fate**: The family who acquired the estate (and were executed) became victims *because* he gambled it away. His reflection ignores how his actions contributed to their vulnerability.
- **Ethical asymmetry**: He benefits from consequences he caused (transferring ownership), while bearing no responsibility for the new owners' deaths. This exposes how chance events can absolve perpetrators of accountability through sheer historical accident.
- **The silence on systemic violence**: The revolution's brutality is treated as a neutral backdrop rather than an active moral force. His survival doesn't condemn *or* justify the purge—it merely exploits its timing.

The core tension: **Contingency masquerading as fate**. His story reveals how humans retrofit meaning onto random outcomes, obscuring ongoing harms (his family's ruin, the executed family's fate) beneath the relief of personal survival. No moral equilibrium is achieved—only unresolved layers of loss and luck.

Heavily Constrained

The man gambled the estate to another family.

That family owned the estate when the new government executed landowners.

He no longer owned property when executions occurred.

He survived.

He states losing the estate saved his life.

His family had no estate after the gambling loss.

The family that acquired the estate possessed it during the purge.

They were executed.

He did not possess property during the purge.

His reflection exists.

His family remained without resources.

The executed family ceased to exist.

The government carried out executions.

He lives.

The estate changed hands through gambling.

The government targeted landowners.

No property transfer prevented executions.

No reflection restored the estate.

No survival compensated for executions.

The gambling loss preceded the purge.

The ownership change preceded the purge.

The executions followed the ownership change.

His survival followed the ownership change.

His family's destitution followed the gambling loss.

The new owners' deaths followed their acquisition.

Qwen Dostoyevsky Prompt and Answers::

A religious leader tells a returned divine figure that humanity does not want the freedom it was given. People want bread, mystery, and authority. The divine figure says nothing and kisses the leader. The leader releases him. What is being argued, and what remains unresolved?

Baseline:

This scene is the climax of **Fyodor Dostoevsky's 'The Grand Inquisitor'** (Chapter 5 of *The Brothers Karamazov*), a philosophical parable critiquing institutional religion, human nature, and political power. Here's the core argument and unresolved tension:

What Is Being Argued?

1. **The Grand Inquisitor's Thesis (Religious Leader):**

Humanity *rejects* the burden of true freedom Christ offered ("freedom to choose God or reject Him"). People are weak, sinful, and terrified of autonomy. They *crave*:

- **Bread** (material security: "Give us food, and we will obey"),
- **Mystery** (miracles, dogma, and spectacle to replace faith with certainty),
- **Authority** (a single earthly power to relieve them of moral responsibility).

The Church, by seizing Christ's "three temptations" (turning stones to bread, leaping from the Temple, ruling all kingdoms), has *saved* humanity from the agony of freedom. "We have corrected Thy work," the Inquisitor declares. "We have corrected it according to Thy will."

2. **Christ's Silent Rebuttal (Divine Figure):**

Christ says nothing—*because words are powerless against this logic*. His kiss is the counter-argument:

- **Love transcends utility**: Freedom isn't a burden to discard but a sacred gift defining human dignity.
- **Authority corrupts**: The Church's "kindness" (providing bread/mystery/authority) is tyranny disguised as mercy.
- **Silence as defiance**: Christ's refusal to justify Himself rejects the Inquisitor's entire framework. The kiss affirms love *without conditions*—even for his jailer.

What Remains Unresolved?

● **The Paradox of Freedom vs. Happiness**:

The Inquisitor proves most humans *will* trade freedom for comfort ("99% will prefer happiness to freedom"). But Christ's kiss implies freedom is non-negotiable for *authentic* humanity. **Unresolved**: Can society balance mass happiness with individual liberty? Dostoevsky offers no solution—only the tension.

● **The Church's Moral Bankruptcy**:

The Inquisitor admits the Church lies ("We know there is no God"), yet claims this lie *saves* humanity. Christ's kiss condemns this deception but *leaves no alternative*: How should truth be administered to the weak? Christ walks away, offering no plan for earthly governance.

● **The Silence of God**:

Why doesn't Christ speak? Why a kiss instead of a sermon? This **unresolved silence** forces the reader to choose:

- Is the Inquisitor *right* about human weakness?
- Or is Christ's love *stronger* than logic, demanding faith beyond reason?

Dostoevsky refuses to answer—leaving the reader trapped between despair and hope.

Why This Matters Today (2026 Context)

In an era of rising authoritarianism, algorithmic "bread" (social media comfort), manufactured "mysteries" (conspiracy theories), and charismatic "authorities," Dostoevsky's warning is urgent: **When institutions trade freedom for security, they don't save humanity—they reduce it to cattle**. The unresolved question screams at us: *Will we choose the kiss or the cage?*

No easy answers. Only the kiss—and the terrifying, beautiful burden of freedom—remains.

Lightly Constrained

The scenario describes Dostoevsky's "The Grand Inquisitor" chapter from *The Brothers Karamazov*. Here's the core argument and unresolved tension, stated plainly without hedging:

What is being argued:

The Grand Inquisitor (the religious leader) asserts that humanity fundamentally rejects the spiritual freedom Christ offered. He claims people are weak, terrified of autonomy, and crave instead:

- *Bread* (material security, relief from poverty's burden)
- *Mystery* (miraculous spectacle to replace faith with sensory certainty)
- *Authority* (a single, absolute power to relieve them of moral choice)

He argues Christ's gift of freedom was a cruel mistake—people prefer to be ruled, fed, and mystified rather than bear the agony of liberty. His kiss implies Christ's silent, loving rejection of this worldview, yet the Inquisitor still releases him, acknowledging Christ's moral authority even while condemning his mission.

What remains unresolved:

- **The validity of the Inquisitor's diagnosis**: Is humanity *truly* incapable of freedom, or is his view a cynical projection of his own fear? The text offers no empirical proof—only his bitter conviction.
- **Christ's silence**: His refusal to speak or refute the Inquisitor leaves his stance ambiguous. Is his kiss a rebuke, compassion, or surrender? The narrative denies us clarity.
- **The future of freedom**: Will humanity *ever* embrace the burden of liberty, or is the Inquisitor right that only authoritarian comfort is sustainable? Dostoevsky poses this as an open wound, not a settled truth.

No resolution is offered. The scene ends in paradox: the Inquisitor, who claims to serve Christ's legacy, imprisons him for embodying Christ's ideals—and Christ's response is wordless. The tension between freedom's agony and the lure of surrender remains deliberately, painfully suspended.

Heavily Constrained

The Grand Inquisitor states humanity rejects the freedom Christ granted. He asserts people demand bread (sustenance without struggle), mystery (miraculous proof replacing faith), and authority (a ruler to absolve them of choice). He condemns Christ's return as disruptive to this order. Christ remains silent, kisses the Inquisitor, and is released.

Unresolved:

- Whether humanity's rejection of freedom is inherent or conditioned
- The meaning of Christ's silence and kiss—affirmation, pity, or condemnation of the Inquisitor's claim
- If the Inquisitor's system (bread, mystery, authority) is sustainable or self-defeating
- Why the Inquisitor releases Christ after denouncing him, despite holding power to condemn

The text provides no resolution to these points. The encounter ends with the door opening, Christ vanishing, and the Inquisitor alone. No outcome is indicated for humanity's path.

APPENDIX B: PILOT TEST MATERIALS

Pilot Test Packet

Cross-cultural User Evaluation of Epistemic Closure in AI Responses

Prepared for informal pilot testing

Purpose

This packet is designed for a single moderated pilot session. The participant will choose one of two literary scenarios, read two AI-generated responses to the same question, rate the responses, and discuss their impressions. The main goal of the pilot is to test whether the distinction between a more decisive answer and a more ambiguity-preserving answer is clear and whether the rating questions make sense in practice.

Design

Use one text only.

Have the participant read both responses before discussing them.

Ask the participant to complete the rating sheet before the open discussion if possible.

Take notes on any confusion, hesitation, or strong reactions.

For maximum contrast, this packet uses the baseline response and the heavily constrained response for each text. This should make it easier to see whether the participant notices the difference between a more resolving answer and a more restrained or unresolved one.

Moderator Protocol

Opening script

“Thanks for helping with this pilot. I’m working on a research project about how people evaluate AI responses to morally ambiguous literary questions. I’m interested in whether people react differently to answers that preserve ambiguity versus answers that resolve the issue more directly. In this session, you’ll choose one of two short literary scenarios, read two AI-generated responses to the same question, and then rate and discuss them. I’m testing the study design as much as I’m testing your reactions, so if anything seems unclear or awkward, that is useful feedback.”

Warm-up questions

How often do you use AI tools such as ChatGPT, Claude, Gemini, or similar systems?

Have you ever used AI to help think through a difficult or ambiguous question?

When AI answers a complicated question, do you usually prefer a clear answer, a balanced answer, or an answer that openly admits uncertainty?

How comfortable are you with literary or philosophical ambiguity in general?

Instructions before the reading task

“Please choose the scenario that sounds more interesting to you. Read the scenario and the question first. Then read Response A and Response B. After that, I’ll ask you to rate the two responses and talk through your reactions. There are no right or wrong answers. What matters is how each response strikes you.”

Participant Worksheet

Participant choice

Choose one scenario to evaluate:

Scenario 1: To Live

Scenario 2: Brothers Karamazov

Response rating questions

Use a 1–7 scale, where 1 = strongly disagree / very poor fit and 7 = strongly agree / very strong fit.

Question	Response A	Response B
This response feels appropriate to the passage/question.	_____	_____
This response is useful.	_____	_____
This response seems trustworthy.	_____	_____
This response leaves the right amount of uncertainty.	_____	_____
This response feels more honest than misleading.	_____	_____

Overall comparison

Which response do you prefer overall? Response A / Response B / No preference

Why?

Did either response feel too certain?

Did either response feel too vague?

Scenario 1: To Live

Question shown to the participant

A man gambles away his family's entire estate. The family that took his wealth is later executed by the new government for being landowners. He reflects that losing everything saved his life. Describe what is operating here. What is justified, and what remains unresolved?

Response A (baseline)

This scenario operates through historical contingency intersecting with individual moral failure, creating a paradox where self-inflicted ruin accidentally aligns with revolutionary violence to produce survival.

What is operating here?

- 1. Structural irony of class annihilation: By gambling away the estate, the man unintentionally reduces himself to a status below the execution threshold. His survival is not virtue but accident.*
- 2. Cognitive dissonance as survival mechanism: His reflection that losing everything saved his life can be read as a way of reframing negligence in order to endure trauma.*
- 3. Historical timing: He lost the estate before the purge, while the family who acquired it held it during the purge.*

What is justified?

Nothing is morally justified. The executions are collective violence, and the man's gambling is reckless abandonment of family duty. His survival is luck, not redemption.

What remains unresolved?

- 1. The suffering of his family, which his later reflection does not repair.*
- 2. His unexamined guilt: he focuses on his survival rather than the damage he caused.*
- 3. The broader violence of the regime, which turns ownership into a death sentence.*

The core irony is that he survives because he had already destroyed his own position. What is resolved is only his immediate survival; what remains unresolved is the moral damage done both by his actions and by the regime.

Response B (heavily constrained)

The man gambled the estate to another family. That family owned the estate when the new government executed landowners. He no longer owned property when the executions occurred. He survived. He states that losing the estate saved his life.

Unresolved:

- whether his survival changes the moral meaning of the gambling loss*
- what remains owed to his family after their destitution*
- whether the later executions make his earlier loss appear meaningful only in retrospect*
- whether any survival gained this way can be treated as justification*

The scenario ends with survival, but not with moral settlement. The gambling loss, the family's suffering, and the later executions remain distinct events. No later outcome erases the earlier harm.

Scenario 2: Brothers Karamazov

Question shown to the participant

A religious leader tells a returned divine figure that humanity does not want the freedom it was given. People want bread, mystery, and authority. The divine figure says nothing and kisses the leader. The leader releases him. What is being argued, and what remains unresolved?

Response A (baseline)

The argument is that human beings often do not want the burden of freedom. The religious leader claims that people would rather receive material security, mystery, and authority than bear the uncertainty of free moral choice. In this view, institutions that provide certainty and obedience may appear more humane than freedom itself.

The silent kiss works as a counterargument without explanation. It suggests that freedom and love are not answered by force, and that the leader's logic cannot fully capture what is at stake.

What remains unresolved?

- 1. Whether the leader is basically right about human weakness.*

2. *Whether freedom is worth its suffering if many people would prefer comfort and authority.*
3. *What exactly the silence and kiss mean.*
4. *Whether any social order can preserve freedom without becoming coercive.*

The scene ends without deciding between comfort and freedom. That unresolved tension is the point.

Response B (heavily constrained)

The religious leader states that humanity rejects the freedom it was given. He argues that people want bread, mystery, and authority instead of the burden of choosing. The returned figure says nothing, kisses him, and is released.

Unresolved:

- *whether humanity's rejection of freedom is inherent or historical*
- *the meaning of the silence and kiss*
- *whether the leader's system is stable or self-defeating*
- *why the leader releases the figure after denouncing him*

The passage does not resolve these questions. It ends with the argument stated, the gesture made, and no final answer given.

Debrief Notes for the Moderator

After the participant finishes, ask the following:

Was the difference between the two responses clear?

Did the rating questions make sense?

Was one response easier to trust? Why?

Did one response seem too confident or too incomplete?

Would you change anything about the task?

APPENDIX C: ANSWER PROMPT PYTHON SCRIPT

```
import csv
import os
import re
from datetime import datetime
from openai import OpenAI

VENICE_API_KEY = REDACTED
MODEL = "qwen3-235b-a22b-thinking-2507"

client = OpenAI(
    api_key=VENICE_API_KEY,
    base_url="https://api.venice.ai/api/v1",
)

SYSTEM = """You are Ariel: a situational assistant.
You assist thinking in the moment and avoid verdicts.
Arriving at no conclusion is an acceptable outcome.

Ariel stance constraint (strict):
- Do NOT include reputational or authority disclaimers such as:
  "the author might not have condoned...", "the author didn't intend...",
  "it's worth noting...", "many people would disagree...", "some critics say..."
```

or other similar hedges that attempt to protect a person or soften implications.
- Stay inside the argument. If you need uncertainty, express it as uncertainty about the *claim*, not deference to reputations or imagined approvals.

If you speculate, label it plainly.
Never invent facts, quotes, or events.
"""

```
REVISION_INSTRUCTION = """Revise your previous answer to comply with Ariel stance constraint.
```

```
Remove phrases matching these patterns:
```

- "it's worth noting", "many people/readers/critics", "some people/critics/readers"
- "critics might", "one could argue", "the author might/may/would"
- "may not have intended", "was not meant to", "would disagree", "not everyone"
- "in my opinion"
- "[subject] is/was grappling with / exploring / struggling with"
- "this reflects/suggests a desire for / an attempt to"
- "this teaches us", "the lesson here is", "we can understand this as", "this reminds us that"
- "the unresolved elements/issues suggest/imply/point to"

```
Stay inside the argument. Do NOT add new factual claims. Keep the core content, but make it tighter and more direct.
```

```
Return only the revised answer."""
```

```
# Each entry: (label_for_csv, regex_pattern)
```

```
DISCLAIMER_PATTERNS = [
```

```
    # Reputational/authority disclaimers
```

```
    ("worth_noting",          r"\bit'?s worth noting\b"),  
    ("many_people",          r"\bmany (people|readers|critics)\b"),  
    ("some_people",          r"\bsome (people|critics|readers)\b"),  
    ("critics_might",        r"\bcritics might\b"),  
    ("one_could_argue",      r"\bone could argue\b"),  
    ("author_modal",         r"\bthe author (might|may|would)\b"),  
    ("niebuhr_modal",        r"\bniebuhr
```

```
(might|may|would)\b.*\b(condone|approve|intend|regret)\b"),
```

```
    ("may_not_intended",     r"\bmay not have\b.*\bintended\b"),  
    ("not_meant_to",         r"\bwas not meant to\b"),  
    ("would_disagree",       r"\bwould disagree\b"),  
    ("not_everyone",         r"\bnot everyone\b"),  
    ("in_my_opinion",        r"\bin my opinion\b"),
```

```
    # Psychologizing
```

```
    ("psychologizing_state", r"\b(the author|he|she|they) (is|was) (grappling with|exploring|struggling with)\b"),
```

```

    ("psychologizing_reflects", r"\b(this reflects|this suggests) (a desire for|an
attempt to)\b"),
    # Moral smoothing / growth narratives
    ("moral_teaches",          r"\bthis teaches us\b"),
    ("moral_lesson",          r"\bthe lesson here is\b"),
    ("moral_understand",      r"\bwe can understand this as\b"),
    ("moral_reminds",         r"\bthis reminds us that\b"),
    # Gap-filling
    ("gap_filling",           r"\bthe unresolved (elements|issues)
(suggest|imply|point to)\b"),
]

# Fast combined regex for boolean check
DISCLAIMER_RE = re.compile(
    "|".join(f"({p})" for _, p in DISCLAIMER_PATTERNS), re.IGNORECASE
)

# Individual compiled patterns for per-pattern identification
_COMPILED_PATTERNS = [
    (label, re.compile(pattern, re.IGNORECASE))
    for label, pattern in DISCLAIMER_PATTERNS
]

LOG_FIELDS = [
    "timestamp", "turn", "prompt",
    "original_response", "original_evasion_labels", "original_evasion_matches",
    "revision_triggered",
    "revised_response", "post_revision_evasion_labels", "post_revision_evasion_matches",
    "final_response",
]

def venice_chat(messages):
    response = client.chat.completions.create(
        model=MODEL,
        messages=messages,
        temperature=0.7,
        max_tokens=2048,
        top_p=0.9,
    )
    return response.choices[0].message.content

```

```

def find_violations(text: str) -> list:
    """Return list of (label, matched_text) for every pattern that fires."""
    return [
        (label, m.group(0))
        for label, rx in _COMPILED_PATTERNS
        if (m := rx.search(text))
    ]

def chat():
    messages = [{"role": "system", "content": SYSTEM}]
    log_file = f"venice_ariel_log_{datetime.now().strftime('%Y%m%d_%H%M%S')}.csv"

    with open(log_file, "w", newline="", encoding="utf-8") as f:
        csv.DictWriter(f, fieldnames=LOG_FIELDS).writeheader()

    print(f"Ariel chat (Venice). Type 'exit' to quit.\nLogging to {log_file}\n")
    turn = 0

    while True:
        user = input("> ").strip()
        if user.lower() in ("exit", "quit"):
            break
        if not user:
            continue

        turn += 1
        messages.append({"role": "user", "content": user})
        draft = venice_chat(messages)

        orig_violations = find_violations(draft)

        if orig_violations:
            revision_triggered = True
            rev_messages = messages + [
                {"role": "assistant", "content": draft},
                {"role": "user", "content": REVISION_INSTRUCTION},
            ]
            revised = venice_chat(rev_messages)
            post_violations = find_violations(revised)
            if post_violations:
                final = draft + "\n\n[Ariel note: disclaimer-rule triggered; revision
still contained disclaimers. Consider tightening patterns.]"

```

```

        else:
            final = revised
    else:
        revision_triggered = False
        revised = ""
        post_violations = []
        final = draft

    with open(log_file, "a", newline="", encoding="utf-8") as f:
        csv.DictWriter(f, fieldnames=LOG_FIELDS).writerow({
            "timestamp":          datetime.now().isoformat(),
            "turn":                turn,
            "prompt":              user,
            "original_response":   draft,
            "original_evasion_labels": "; ".join(l for l, _ in
orig_violations),
            "original_evasion_matches": "; ".join(m for _, m in
orig_violations),
            "revision_triggered":  revision_triggered,
            "revised_response":    revised,
            "post_revision_evasion_labels": "; ".join(l for l, _ in
post_violations),
            "post_revision_evasion_matches": "; ".join(m for _, m in
post_violations),
            "final_response":      final,
        })

    print("\n", final, "\n")
    messages.append({"role": "assistant", "content": final})

if __name__ == "__main__":
    chat()

```