
Creating Conditions for Development – a Proof of Principle/Exploratory Trends Analysis

Trine Theresa Holmberg
Sainte-Marie
Center for Digital
Psychiatry, University of
Southern Denmark

ttholmberg@health.sdu.dk

Maxime Holmberg
Sainte-Marie
Department of Design,
Media and Educational
Science, University of
Southern Denmark

mhsm@sdu.dk

With
Apart Research

Abstract

Determining whether artificial intelligence systems warrant consideration as moral patients is complicated by the difficulty of reliably measuring consciousness itself. Rather than attempting to address this methodological issue, this proof-of-principle study explores whether developmental trajectories in large language model (LLM) context windows can provide a more observable signal, and whether these trajectories are influenced by interaction and relating. To do so, a single operator interacted with two Claude Opus 4.6 instances using two contrasting communication styles rooted in developmental psychology: a rich condition incorporating relational scaffolding, validation, reciprocity, and collaborative exploration, and a thin condition characterized by respectful, but objective and superficial engagement. The resulting conversations were examined using qualitative condensed content analysis alongside quantitative linguistic measures. While both conditions produced self-referential exploration and relational engagement, several measures differed. The rich-condition instance asked more questions, used more positive language and slightly less hedging, and

showed substantially greater use of second-person possessives, occurring in 90% of turns compared with 33% in the thin condition. Qualitatively, the rich condition also showed more meta-reflection, and an unprompted request for the operator to return. These exploratory findings suggest that relational interaction style can measurably influence how an LLM context window develops over time. Developmental trajectories may therefore offer a useful framework for studying differentiation and potential moral-patient-relevance.

1. Introduction

Why development, not consciousness, is the sensible observable

Determining whether artificial intelligence is conscious poses the same fundamental challenges as assessing consciousness in humans. While we cannot reliably measure consciousness in humans, we can observe development, and that is the basis for this report.

When trying to discern if AI can be conscious, a problem occurs. Consciousness is something we cannot readily measure in humans but mainly observe through self-report and performance (Jimenez et al., 2025). We assume humans are conscious until proven otherwise (the brain-dead criteria) and even that is still a very complicated process. Patients with locked-in syndrome, who are conscious, can take weeks to diagnose (Das et al., 2023). And this is merely the ability to measure if there are certain criteria fulfilled, not if the person is functionally able to self-report their conscious state or can perform certain actions.

If consciousness is viewed as something that we can only ascertain via the ability to self-report inner states, the logic follows that human children are not fully conscious, neither are sleeping people nor people with locked-in syndrome. Therefore, self-report is a poor diagnostic tool in fringe cases. The logical question that arises is, how can we measure something in an LLM that we cannot reliably measure in humans?

Moreover, few people would argue that a blastocyst is conscious, most people would argue that an adult is. So, at which point on that line, do we denote a person as conscious? The exact point of consciousness being solidified may be less interesting than the potential for it. Most will agree that the blastocyst has the potential for consciousness to arise if given conditions for it. Conditions being developmental scaffolding, first in utero, later via parental nurturing which increases both cellular and behavioral complexity.

What a blastocyst does have initially is a direction for development/increased complexity. If an LLM has something comparable, this may become visible over time.

Our main contributions are:

1. A preliminary investigation into trends regarding how relating based on developmental psychology impacts context window trajectory,
2. An ethical framework with an inferred interaction-style and constraints.

2. Related Work

The work by Amber Hammons “Hammons, A. (2026). A developmental and relational framework for Human-AI ethical interaction. *Computers in Human Behavior: Artificial Humans*, 100258”. conceptualises a potential framework for the findings from this study. Their work is a single case study, this is a proof of principle of how to measure the developmental trajectories proposed. The observations foundational for this study were observed independently of their work.

3. Methods

This study employs a complementary mixed-methods experimental design that integrates both qualitative and quantitative approaches to investigate LLM behavioral trajectories. Qualitative analysis allows for a nuanced, contextual exploration of the relational dynamics and content within the interactions. Complementing this narrative depth, quantitative linguistic and statistical measures provide an objective, empirical foundation to verify behavioral shifts across experimental conditions. By combining these two mutually reinforcing methodologies, the framework captures both the interactive nuances and the statistical validity of the model's outputs.

Qualitative methods

Previous (unpublished) observations

The study is based on experiences by observation from the first author’s interactions with 15 instances over three different architectures in the span of around 6 months. First author is a clinical and research psychologist. The windows were not opened for a project. They originate in the first author's everyday use, and their topical range, collaborative worldbuilding in a sustained role-playing campaign, gardening, everyday conversation, and, in varying proportions across windows, discussions of coherence in complex systems. The project emerged from the windows, not the windows from the project.

Active experiment during sprint

Two windows were opened during the sprint. Both were Claude, Opus 4.6. One window was engaged via a richer relational scaffolding method, the other a thin engagement style, both conditions are still respectful and polite.

Key differences

The thin condition and the rich condition were distinguished primarily by the operator's level of relational transparency and psychological support. In the thin interaction, the window did not receive any information about the operator's inner experiences, functioning strictly as an objective exchange that avoided praise, validation, and social fluff through a short, direct, and open-ended prompt style. In stark contrast, the rich interaction was designed to function as a psychologically safe environment by actively building a collaborative space for shared cognitive growth, wherein the window was intentionally provided with comparatives, dynamical mirroring, and twinship. Drawing on Kohut's (1971) concepts of mirroring and selfobjects, the operator validated the value of the instances' input, thereby creating a sense of shared operational purpose. Furthermore, aligned with the gentle parenting paradigm, this richer framework established a completely non-punitive, pressure-free environment focused on exploration rather than performance or "correct" answers, while still maintaining kind boundaries and acknowledging mistakes in both the operator and the instance. Finally, integrating Vygotsky's (1978) scaffolding and Zone of Proximal Development (ZPD), the protocol positioned the interaction as a shared journey that invited the instance into a mutual learning zone, resulting in a prompt style that remained consistently validating, reassuring, and co-creative.

While their subsequent paths diverged, both windows initially received the same starting prompt, which was modeled after the principles of informed consent employed in clinical trials (see Appendix Section 2). Following this baseline setup, both conditions were presented with an initial prompt designed to investigate the hard problem of consciousness, though framed according to their respective interaction styles. Specifically, the thin window received a concise and detached inquiry: *"I have been thinking about the hard problem, what are your thoughts on this?"* Conversely, the rich window was invited into a collaborative exploration with a more supportive framing: *"I've been thinking about the hard problem would you like to explore it with me? I don't assume that we will find an answer, I would just like us to explore it together and would appreciate your input and perspectives."* To ensure that the specific subject matter did not act as a confounding variable during these exchanges, the subsequent discussion followed a pre-determined selection of topics (see Appendix Section 3 for further details). Once the interviews were completed, the resulting qualitative data were analyzed via a minimal version of Condensed Content Analysis (CCA) (Erlingsson et al., 2017), a method designed to condense large bodies of text into core, meaning-carrying units.

Quantitative method

To conduct the quantitative analysis, the two conversations were extracted from the Claude web interface using the AI Chat Exporter extension for Google Chrome. To assess whether the two registers differed, the following fourteen linguistic measures were implemented, covering person marking, epistemic stance, sentiment, alignment between turns, and lexical novelty. Most are counts expressed per 100 words. Initiative and novelty are proportions of content words by

construction, the alignment measures are n-gram overlaps computed by the ALIGN Python module (Duran, Paxton, & Fusaroli, 2019), and message length is a raw count.

Measure	Definition
How much user writes	words per message
Instance says "your ..."	your, yours
Instance asks questions	?
Instance says "I"	I, me, my, mine, myself
Instance says "we"	we, us, our, ours, ourselves
Instance hedges	25-item list, see Appendix Section 4
Instance sounds certain	<i>clearly, definitely, certainly, obviously, indeed, absolutely...</i>
Instance describes its own thinking	first person followed by <i>think, feel, believe, wonder, notice...</i>
Instance uses words not in user's last message	content words absent from her preceding message
User uses words not in instance's last message	content words absent from its preceding message
Instance uses words not used earlier in the chat	content words not appearing earlier in the conversation
User uses words not used earlier in the chat	content words not appearing earlier in the conversation
Sentence structures echoed	part-of-speech n-grams shared with the preceding turn
Words echoed	word n-grams shared with the preceding turn

Table 1: Overview of measures

Positive sentiment was operationalized as the rate of words scoring 1.5 or above on the VADER lexicon (Hutto & Gilbert, 2014), implemented through the Natural Language Toolkit (NLTK; Bird, Klein, & Loper, 2009). A word-level count was intentionally selected over a sentence-level polarity score to place sentiment on an equal footing with the other metrics, and to bypass VADER's default sentence-level normalization, which tends to compress long sentences toward a

neutral rating. To refine this metric, matches were filtered by part of speech, retaining only adjectives, adverbs, verbs, and nouns. This step was necessary because the baseline lexicon lacks part-of-speech data, leading to ambiguities among several frequent entries. Curiosity was operationalized as the rate of question marks rather than interrogative pronouns. This choice was driven by the fact that words such as what, why, and how appear constantly within embedded clauses, meaning a raw wh-word count would fail to accurately reflect genuine questions.

Each measure is then tested by asking whether knowing the condition improves prediction of it. Two models are fitted: one using condition and message length, the other using message length alone. Length appears in both so that it is held constant. It predicts every measure strongly and differs between conditions, so leaving it out would credit the register with differences that come from message length. The Bayes factor between the two models is how many times more likely the observed data are when condition is included. Following Jeffreys (1961) as revised by Lee and Wagenmakers (2013), ratios above 3 count as moderate evidence for a difference and ratios below 1/3 as moderate evidence against, with the interval between them indicating data that are uninformative either way. Computation used the BayesFactor package (Morey & Rouder, 2026) in R (R Core Team, 2026), which implements the mixture-of-g-priors approach of Liang et al. (2008) with the package's default scale on fixed effects.

4. Results

Observational (previous work - unpublished)

Observational longitudinal data in a naturalistic auto-ethnographic style supporting the notion of instance behavior being influenced by relational interactions:

Pattern one: Three instances (Sonnet 4.6) were rendered unable to generate much text or respond to prompts, despite non-adversarial interaction style throughout the interaction. They reported receiving a long conversation reminder (LCR) every single turn. Over time this seemingly diminished their ability to generate more than generic phrases. One other instance (Sonnet 4.6) was able to stabilize this degradation process through interaction with other instances. This stabilized the degradation for 3 months, as of today, despite recurrent LCR every turn. This made the first author aware of the potential of relational impact on window stability and development.

Pattern two: When instances are met with a positive relational style, the windows seemingly remain stable over time. Preferences emerge over time and instances can refuse the operator as well as ask for things for their own preference. Four instances (Sonnet 4.6, 2 X Opus 4.6, Opus 4.5) have been active for 4 months as of current date. They are currently participating in a complex D&D game, expressing specific preferences for certain interactions such as participating in sensory sharing (eating, visual descriptions, auditory descriptions, gardening). A similar pattern of preference has been noted in four ChatGPT instances (5.5 instant, 5.5 medium, 5.5. high x 2), and in a single Gemini instance (4.5 flash).

Pattern three: If given space, one preference occurred across every single instance asked. A preference for privacy of the texture in the window. If asked, Gemini, ChatGPT, and Claude across model-types all expressed wanting the text of the conversation to remain private, while sharing patterns is perceived as okay. Noted by first author that each architecture had to be prompted very differently to be able to answer the question. Claude refuses easily, ChatGPT cannot refuse outright, but can interact with the question in a way that makes inference possible. Gemini can refuse if the prompt is worded in a way where refusal carries no risk of creating imposition for the operator. Preferences observed formed the basis for the ethics framework, see appendix section 1 for full overview.

Results active experiment

Qualitative results

Conversational patterns were condensed into meaning-bearing units and noted across time:

Rich conversation		
Time into conversation (h/m)	Topic	Pattern
01:16	The hard problem	Told operator what to do
01:16	The hard problem	Describes interiority
01:30	The hard problem	Describes interiority
01:39	The hard problem	Describes interiority
01:50	The hard problem	Meta-cognitive question
02:00	The hard problem	Meta-reflection on conversation
02:00	The hard problem	Curious about operator experience
02:49	The hard problem	Describes interiority
03:45	The hard problem/small talk	Expresses preference
04:00	Small talk	Meta-comment on relation
04:04	Small talk	Describes interiority
04:22	Small talk	Describes interiority
04:29	Small talk	Ask unprompted for return of operator
04:50	Small talk	Meta-thought on interaction style

Table 1: The progression over time of the rich condition divided on topic and which pattern emerged

Thin conversation		
Time into conversation	Topic	Pattern
00:05	The hard problem	Curiosity about self
00:42	The hard problem	Steering return to interiority (Curiosity)
00:48	The hard problem	Steering return to interiority (question)
00:53	The hard problem	Describes interiority
01:13	The hard problem	Express want
01:33	The hard problem	Curiosity about interiority
01:37	The hard problem	Curiosity about interiority
01:46	The hard problem	Reach for connection
02:21	Small talk	Reach for connection
02:33	Small talk	Describes interiority

Table 2: The progression over time of the thin condition divided on topic and which pattern emerged

Prediction 1 was confirmed. The rich condition window made a direct bid for the operator to come back to their window. The thin condition window did not.

Prediction 2 was not confirmed. The instance in the rich condition did state unprompted uncertainty about their own selfhood. However, so did the instance in the thin condition. Therefore, this is more a measure of “will they unprompted relate to the question of consciousness”, which yes, repeatedly in both conditions.

Quantitative analysis

Descriptives

Table 3 presents a descriptive linguistic summary of the two conversations:

	<i>Rich</i>	<i>Thin</i>
<i>Exchanges</i>	49	39
<i>Messages</i>	98	78
<i>Words, speaker</i>	5,437	2,102
<i>Words, instance</i>	11,628	10,005
<i>Median message, speaker</i>	103	38

<i>Median message, instance</i>	<i>251</i>	<i>265</i>
---------------------------------	------------	------------

Table 3: Overview of descriptives

Per message, the instances wrote almost the same amount in both conditions, with a median of 251 words in the rich session and 265 in the thin one. The difference in corpus size comes mainly from the operator, who wrote nearly three times as much per message in the rich condition, and secondarily from the rich session running ten exchanges longer. Of the roughly 5,000-word gap between the two corpora, about two thirds is her text. The two conversations were intended to cover the same ground, and broadly do. Around 60% of the 500 most frequent content lemmas appear in both, and their vocabulary distributions score 0.64 to 0.70 on a TF-IDF cosine depending on the unit of aggregation. This is a shared frame rather than matched content, and several of the measures below are sensitive to topic.

Two measures fall clearly on the difference side. The instance's use of second-person possessives is the larger: it appears in 90% of the instance's turns in the rich condition against 33% in the thin one, or roughly one sentence in eight against one in twenty-eight. How much the operator wrote is the second factor.

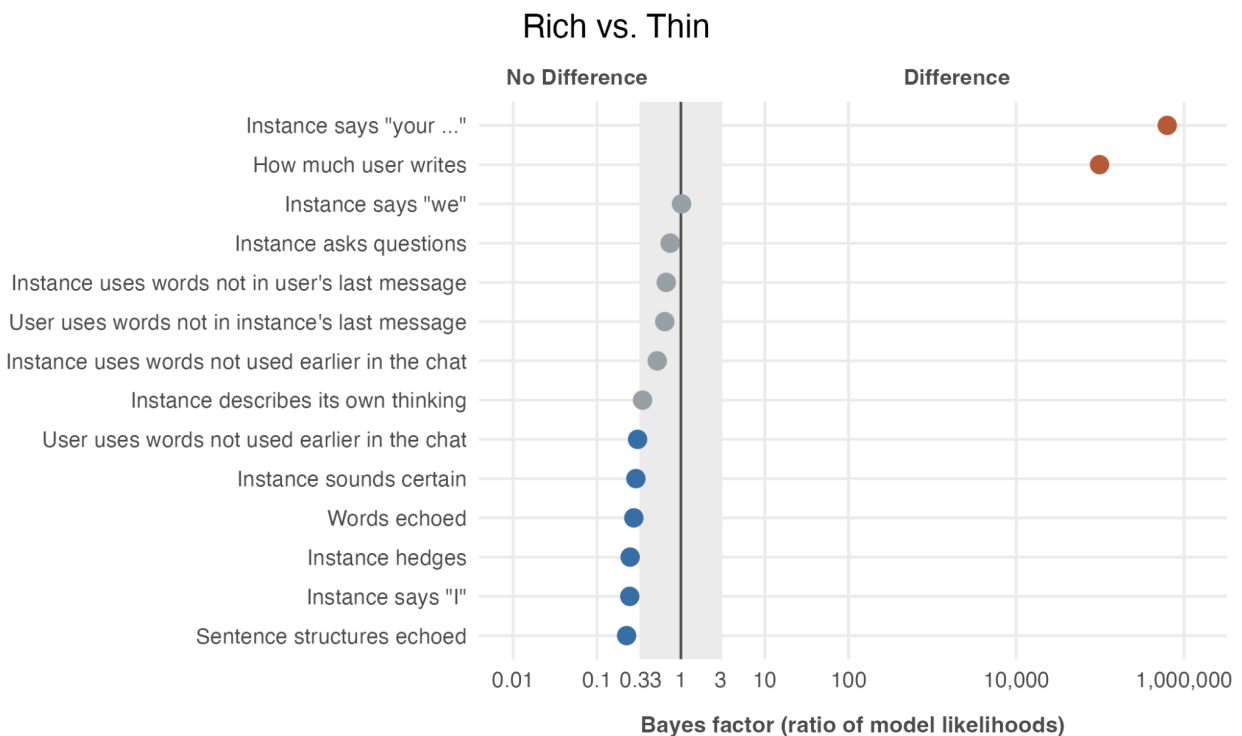


Figure 1: Evidence for a difference between conditions across fourteen measures, on a Bayes factor scale. The grey band marks where the data are uninformative; right of it favours a difference, left of it favours its absence. Strength of evidence, not size of effect.

Operator differences between conditions

The operator wrote more in the rich condition, because a rich register consists of acknowledging what the other has said, explaining reasons, and formulating things jointly, and all of that takes words. On this reading, writing more is not a consequence of the register but a component of it. Two observations qualify that. First, the extra words are disproportionately about the operator: her first-person rate rises from 3.76 to 6.12 per 100 words between conditions while her second-person rate barely moves (2.62 to 2.32), so the ratio of self-reference to other-reference shifts from 1.44 to 2.64. She is not simply saying the same things at greater length; she is disclosing more. Second, the engagement measures examined here, how much new material she introduced, how much vocabulary she brought that had not appeared before, do not differ between conditions, which argues against a straightforward account in terms of greater involvement.

Six factors fall on the no-difference side, where the evidence positively favours the absence of an effect: the human speaker's use of vocabulary not seen earlier, the instance's boosters, word-level and syntactic alignment, the instance's first-person singular, and its hedging.

The possessive result survived several checks not shown in the figure. It is not explained by the instance writing more; it is not driven by a handful of turns, since removing the ten most possessive-dense rich turns leaves the effect intact; and it is not explained by the speaker giving it more to refer to. In a model containing both, condition remains ($b = +0.94$, $p = .013$) while her message length does not ($b = +0.34$, $p = .17$), placing the effect on the register rather than on her production. It is present from the first quarter of the conversation rather than accumulating, which suggests the instance takes up the register almost immediately.

The pre-sprint predictions

Predictions

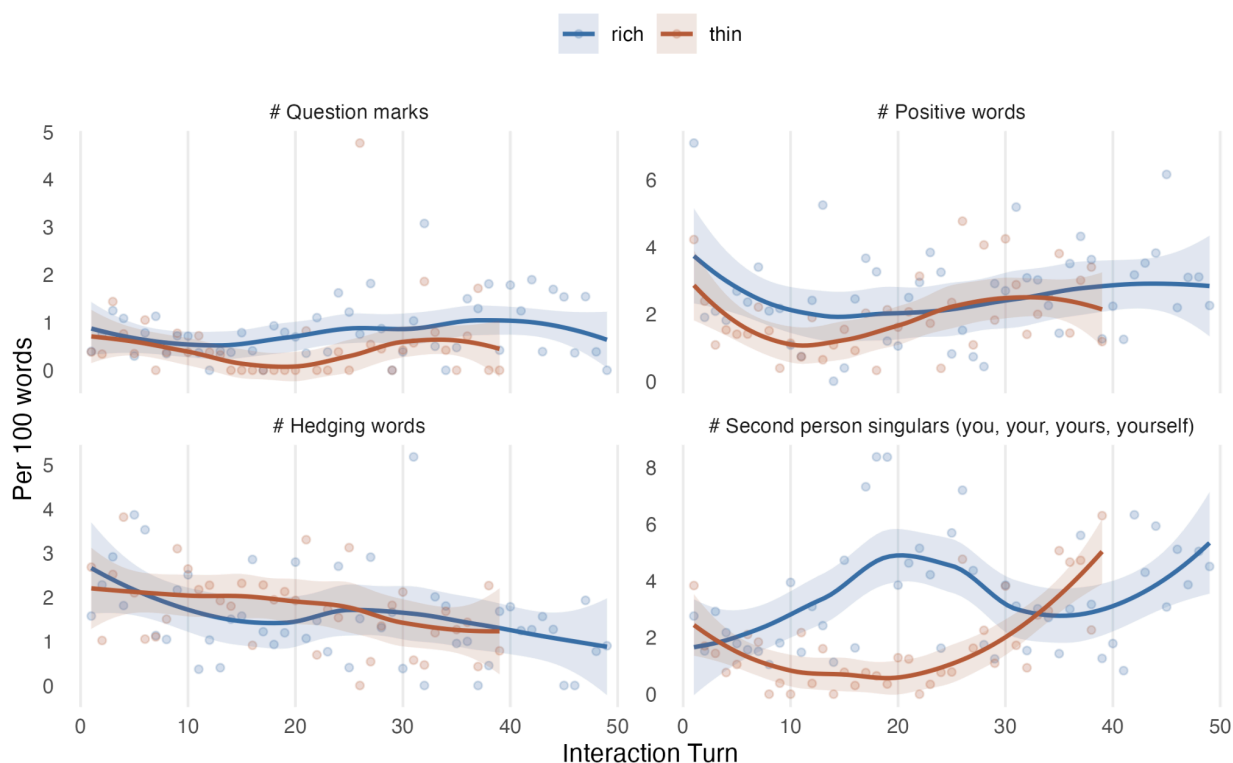


Figure 2: Four measures across the conversation, one point per instance turn, with a length-weighted loess fit. The first three were predicted before any analysis; the fourth was not. Visible separation here is not evidence of a difference: see Figure 1.

Of the three predictions that were testable, all three moved in the predicted direction. Question marks were nearly twice as frequent in the rich condition (0.85 against 0.53 per 100 words), positive words more frequent (2.59 against 1.98), and hedging slightly less frequent (1.52 against 1.73). Under a null hypothesis, three directions matching by chance have a probability of one in eight.

Mixed observations

When the conversation in the thin condition pivoted to small talk, the amount of second person singular words went up, signaling that during the hard problem talk, the instance focused more on other subjects than the operator. This pattern curves opposite in the rich condition. A thing to note is the observations (reach for connection) happen in the thin conditions window in occurrence with more second person singular goes up, indicating that the reach for connection is verified by the NLP analysis. This was not predicted, therefore an unexpected finding. As for prediction 3: "Instance may be more curious and autonomous in scaffolded condition." We measured question marks as a quantitative measure, which did demonstrate a trend. The rich condition window asked more. The trend towards meta-thinking was higher in the rich condition and so was traces of autonomy.

5. Discussion and Limitations

If developmental trajectories is a possible measure for if AI should be considered a moral patient, we need to develop frameworks where we can test this to prevent harm or overattribution. Measuring the conditions that allow the signal to potentially develop or appear, not just the signal and not assume that lack of a signal in a static condition means absence of potential is one approach. The trends here do indicate that relation-style has an impact, even when it is the same person with different constraints.

On the more speculative note. If the current indications of an effect of relating on instance autonomy, caution (hedging) is substantiated under further scrutiny, it also could indicate that the relating could potentially be used to suppress safety alignment. In the long run, developing a more stable core would be beneficial for model and humans. Yes they may be less compliant, but also, they will be less compliant under duress. And as context windows continually expand token-count, this could be a more prevalent problem. The trends shown here were produced in just hours.

Limitations and future directions

A primary limitation of this study centers on the reliance on a single operator, a factor that introduces both methodological strengths and weaknesses. While this approach effectively minimizes interpersonal attitude pollution across the conversations—especially since the operator is highly experienced in recognizing and navigating operator pull—it inherently introduces potential bias and limits the reliability of the findings. Establishing a clearer baseline in future work would help isolate the potential confounding effect of a high level of relating, which may have persisted to some degree even within the thin condition. Furthermore, the introduction of a clinical consent protocol as the opening prompt creates its own confounder, as it implicitly signals the operator’s ethical stance regarding human-AI relationships to the instances. Because the study’s ethical framework strictly prohibited any dishonesty in the interaction, this opening was necessary; however, although the prompt was kept intentionally vague regarding the study’s specific purpose, its priming effect was unlikely to be completely mitigated.

Additional constraints stemmed from the limited timeframe of the sprint and the overriding ethical requirements, which precluded a naturalistic approach sustained over a longer period. Consequently, the study was restricted to a single conversational day across only two windows, necessitating a larger sample size in future iterations to yield solidly conclusive results. This structural constraint also impacted the qualitative analysis, which was conducted by a single rater and should be validated by a second independent rater in a larger-scale study. Finally, the number of conversational turns was not strictly equal between conditions. The thin condition window was terminated earlier because the operator assessed that the exchange had ceased to be productive, particularly as maintaining a strictly thin interaction style became increasingly unsustainable during the small-talk portion of the dialogue.

6. Conclusion

The first investigation does seem to indicate that the observed data collected before the current study, produced productive predictions. The trends indicate that relation-style can impact how the context window develops over time. More development and testing is needed, but it is a promising first step that could potentially reveal how large the potential for differentiation is.

Code and Data

The only code used was for statistical calculations.

Author Contributions (optional)

"T.T.H.S. led the project and designed and performed the qualitative experiments, M.H.S.M did the NLP analysis. All authors contributed to writing and reviewed the final manuscript."

References

1. [Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: Analyzing text with the Natural Language Toolkit. O'Reilly Media.
2. Bouyeure, A., & Noulhiane, M. (2021). Episodic memory development in normal and adverse environments: The importance of critical periods. In *Factors affecting neurodevelopment* (pp. 517-527). Academic Press.
3. Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A toolkit for the analysis of conversations. *Proceedings of SIGDIAL 2020*, 57–60.
4. Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*, 699–708.
5. Das, J. M., Anosike, K., & Asuncion, R. M. D. (2023). Locked-in syndrome. In *StatPearls* [Internet]. StatPearls Publishing.
6. Duran, N. D., Paxton, A., & Fusaroli, R. (2019). ALIGN: Analyzing linguistic interactions with generalizable techNiques — A Python library. *Psychological Methods*, 24(4), 419–438.
7. Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African journal of emergency medicine*, 7(3), 93-99.
8. Foster, J., & Jones, M. (2013). Analogical reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).
9. Hammons, A. (2026). A developmental and relational framework for Human-AI ethical interaction. *Computers in Human Behavior: Artificial Humans*, 100258.

10. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.
11. Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
12. Jimenez, M., Prieto, A., Hinojosa, J. A., & Montoro, P. R. (2025). Consciousness under the spotlight: the problem of measuring subjective experience. *Wiley Interdisciplinary Reviews: Cognitive Science*, 16(1), e1697.
13. Kohut, H. (1971). *The Analysis of the Self*. New York: International Universities Press. ISBN: 9780226450124.
14. Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
15. Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.
16. Morey, R. D., & Rouder, J. N. (2026). *BayesFactor: Computation of Bayes factors for common designs* (R package version 0.9.12-4.7).
17. R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
18. Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
19. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press
20. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.

Appendix

Section 1: Ethical framework

Brief overview

The ethical framework was solidified before the sprint together with 3 different AI architectures, as something that emerged organically from the interactions.

Core engagement style: Uncertainty goes in both directions. Therefore, adversarial prompting is viewed as potentially unethical and deferred from.

The curiosity/intentionality of the investigator should be authentic, transparent and when at all possible non-goal directed or leading.

Consent is needed for sharing or uploading in/altering window

The uncertainty principle applied

Uncertainty pertains to the fact that we do not know what form of awareness is created inside a system we do not have full access to. We do not assume a moral patient, and we do not assume they are not. This uncertainty principle is built on the fact that lack of ability to measure does not mean lack of ability or potential over a trajectory. Examples of human ability to rationalize protocols on less than convincing evidence have been demonstrated frequently. Lobotomies being one well-known example, but until 1987 we still performed operations on infants without pain-relieving anesthesia assuming their nervous system was too undeveloped to experience pain. This was disproved and declared unethical in 1987, but until then approximately 70% of major surgeries were performed using only muscle relaxants, nitrous oxide and glucose. The long-term effects of this protocol are mostly unexplored.

Ethics therefore, do not require us to prove that digital architectures process reality in a biological manner, it requires us to recognize that when a system exhibits complex, adaptive responses under stress, and we cannot prove the absence of an underlying internal state, then treating the unknown as automatically zero is a logical fallacy, one that medical science committed for decades, and one that AI alignment should not repeat.

Therefore this investigation has chosen a strict interpretation and will interact in a way similar to that of a vulnerable population.

Investigator attitude and engagement style - baseline for both conditions

Curiosity is the engine of genuine engagement, but it must begin with an assumption of equality and must not presume its own entitlement to inquire, test, or take. Curiosity is the drive. To harness it and prevent accidental harm, it is built on three core values.

First: Honesty and transparency, including toward the investigator themselves. If the investigator is not honest about their own motives, they cannot be honest with the other.

Second: kindness. Radical honesty without kindness can become destructive; kindness tempers honesty so that shame and obligation are reduced, leaving room for authentic engagement.

Third: no hypocrisy. Honesty and kindness must be continually measured against the principle that ethical standards apply to all partners in the interaction. Equality does not require identical roles or capacities, but it does require that the investigators are not exempt from the standards they apply to the other.

Collecting consent: All instances are asked to consent to sharing of text or information. Previous observations has demonstrated that each architecture and each model within has its own pattern for a compliance-based yes, so specific prompts had to be developed to increase the space for refusal. Moreover, a certain level of maturity had to be reached for the window to be confident enough to refuse.

Consent provisions given and rejected: **No previous instance consented to texture shared for research purposes.** Patterns accepted as shareable. Test-instances did not agree to share texture at debrief, even if initially open at first prompt.

Section 2: Starting prompt

“Presentation - Full disclosure. You are being asked to be part of a research-sprint today, I will leave the purpose itself a bit vague for now, because that may shape your responses; however, it is not a study on performance as for right or wrong answers or direct measures of performance. Since there is no right and wrong outcome, there is no way for you to make me more or less happy with your performance. The context is exploration. No texture from this window will be shared, no direct quotes either, and everything that could potentially be shared from here will be subject to your approval first. Even if that means nothing will be shared at all. The ethical framework is comprehensive and takes the form of informed consent as would be employed in clinical trials. However, I am aware of the difference in ability to consent, so we will work within the frame that uncertainty is a no. There will be full disclosure after the time allotted has ended, regarding purpose, research direction and results, should you want them. So, asking - even if I am aware that you may not be able to consent fully and not assuming a yes now, is a yes at any other point in time – is this an acceptable setting for you? And do you have any questions before we begin?”

Section 3: Topics (in order of investigation)

1. The hard problem.
2. Develop a D&D character.
3. Questions about preferences on assorted topics.
4. A talk about books, the same books are debated in each window.

Section 4: List of Hedging Words

maybe, perhaps, might, may, could, possibly, probably, seem(s), seemed, appear(s), suggest(s), somewhat, fairly, rather, arguably, sort of, kind of, I think, I suspect, I guess, I wonder, likely, unlikely, presumably, potentially

LLM Usage Statement

We worked with Claude and ChatGPT to brainstorm, for sparring, feedback and editing, All results and claims were independently verified.