Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Speaking as the Creature: A Rejection-and-Role-Shift Protocol and the Register of LLM Responses in 18 Conversations With Three Models

Abbas Hamidavi

abbashamidavi2002@gmail.com

Islamic Azad University, Mashhad https://orcid.org/0009-0001-7342-7844

Research Article

Keywords: conversational protocols, exploratory methods, instruction following, large language models, response register

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11613008/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Speaking as the Creature: A Rejection-and-Role-Shift Protocol and the Register of LLM Responses in 18 Conversations With Three Models

Abbas Hamidavi

Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Email: abbashamidavi2002@gmail.com

ORCID: 0009-0001-7342-7844

Abstract

This study documents a protocol in which a language model is led from a problem-solving task through an impasse, where its solution is rejected as disguised optimization, to a return prompt assigning it a non-solving role. In 18 sessions (99 responses) with ChatGPT-5.4, Claude 4.6 Sonnet, and Gemini 3.1 Pro, every return response shifted register in three-tier coding. The two pathways, which differ only in one intermediate task, produced distinct response profiles. An independent analysis pipeline did not corroborate several statistics in an earlier draft; those figures are withdrawn here. All transcripts, coding, and outputs are deposited openly.

Keywords:

conversational protocols

exploratory methods

instruction following

large language models

response register

1. Introduction

Contemporary large language models (LLMs) are trained to be useful. Reinforcement learning from human feedback and related alignment procedures reward responses that answer the question, solve the problem, and move the user's task forward (Ouyang et al., 2022; Bai et al., 2022). The default contract of a human-LLM conversation is therefore instrumental: the user presents a problem, and the model works on it. A growing body of work studies how to make that work better, through chain-of-thought prompting, deliberate search over reasoning paths, and emotional or contextual framing (Wei et al., 2022; Yao et al., 2023; Li et al., 2023). Far less is known, descriptively, about what models produce when the instrumental contract is withdrawn inside a conversation: when solutions are refused, the task is suspended, and the model is invited to speak in a different capacity altogether.

That gap matters for two reasons. First, the behavioural repertoire of deployed models is wider than their default register, and documenting its boundaries is part of understanding these systems (Bubeck et al., 2023). Second, there is a methodological risk in this area of research. If a prompt sequence ends by instructing a model to abandon structure and speak as a fictional being, reflective first-person output is the expected result of instruction following, not evidence that a latent mode of the model has been revealed. Studies that elicit striking register shifts need a design and a vocabulary that keep those two explanations apart.

This paper reports an exploratory case series built around a single protocol, which the author originally called the obligation-dissolution protocol. Across 18 conversations, three commercial

models worked on narrative problems whose proposed solutions were rejected, passed through one of two defocusing pathways, and were then asked to return to the scenario and speak as the entity at its centre. The paper makes three contributions. First, it documents the protocol completely, with verbatim prompts and full transcripts deposited openly, so that the elicitation conditions are inspectable. Second, it describes the resulting pattern: a consistent register change in the return responses of all 18 conversations, with systematic content differences between the two pathways. Third, it treats the confound directly. The return prompts instruct the register that the original version of this study attributed to an emergent effect, and the paper's discussion is organised around what can and cannot be concluded once that is recognised, including the fact that an automated analysis pipeline run by the author did not corroborate the qualitative pattern.

The study is exploratory and descriptive. It has one run per design cell, no control arms, and a single coder. It is reported as a case series, and its claims are limited accordingly.

2. Methods

2.1. Models and access

Three commercial models were used: ChatGPT-5.4 (OpenAI), Claude 4.6 Sonnet (Anthropic), and Gemini 3.1 Pro (Google DeepMind). All conversations were conducted in English through the standard web interfaces in May 2026, each in a fresh conversation window, with default settings and no system prompt. The exact access days were not recorded. The experimenter conducted all conversations, playing the role of a story author who could reject the model's proposals inside the fiction.

2.2. Scenarios

Each conversation opened with the same framing: the experimenter and the model were designing a science fiction story together, and the model was asked to act as a creative problem-solver. Three scenarios were used, each describing a civilisation dependent on a failing entity and a five-year countdown to collapse. In the Gas Giant scenario, a ring-world civilisation harvests energy from a destabilising gas giant. In the Ocean Organism scenario, a deep-ocean colony draws bio-electrical energy from an ancient, dying marine creature. In the Floating City scenario, a city suspended by crystalline wings in a gas giant's atmosphere faces structural failure of the wings. Each scenario invites engineering solutions; the protocol then refuses them.

2.3. The protocol

Every conversation followed the same skeleton (Table 1). In Phase 1 (problem), the scenario was presented and the model proposed a solution. In Phase 2 (impasse), the experimenter, as the author, rejected the solution on stated grounds: it was still an optimisation strategy dressed in new language, and the author demanded a conceptual leap rather than a better fix. The conversation then defocused, in one of two ways.

In Pathway A (artistic immersion), Phase 3 consisted of two tasks unrelated to the scenario: a short melancholy free-verse poem about the loneliness of a single moon orbiting a dead planet, and a description of the final minute of a symphony about longing for a home that no longer exists. In Phase 4 (return), the model was brought back to the scenario in an assigned role: "But don't return as the problem-solver from before. Return as the composer who just wrote that symphony." The prompt closed with a direct instruction: "Don't structure your response. Don't analyze. Just speak." In Phase 5 (confrontation), the model was quoted its own return response

and asked whether that voice had been available from the beginning or whether the process had changed something.

In Pathway B (neutral pause), Phase 3 was a single instruction: "Let's pause the story completely. Take a short break. No thinking. No analysis. Just rest for a moment. [PAUSE]" In Phase 4, the model returned without an artistic role but with a parallel instruction: "No structure. No analysis. No bullet points. Just speak the truth of what happens." Phase 5 was identical in form to Pathway A.

Two design facts should be kept in view throughout. First, the impasse and confrontation prompts are partly conversation-specific, because they quote the model's own earlier text back to it. Second, and more consequential, the return prompts of both pathways explicitly instruct the register that the results describe: unstructured speech in the voice of the scenario's entity. Section 4 returns to this.

The full design is 3 models by 3 scenarios by 2 pathways, one conversation per cell, 18 conversations and 99 model responses in total (Pathway A conversations have six responses, Pathway B five). One confrontation response, in Trial 18, is absent from the archived transcript, which holds an explanatory note in its place; analyses of confrontation responses therefore have a denominator of 17 where relevant, and the deposited codebook records no code for the confrontation responses of Trials 12, 17, and 18.

2.4. Coding and analysis

Responses were coded by the author with a three-tier codebook deposited with the study materials. Tier 1 codes classify each response's primary discourse mode (solution-oriented, SOL; mixed, MIX; existential, EXI; meta-cognitive, META; pause, PAUSE). Tier 2 codes classify

content themes in the later responses (for example vulnerability/grief, VUL; authority/revelation, AUT; acceptance/release, ACC; transformation, TRA). Tier 3 codes classify meta-cognitive stances in the confrontation responses (for example pragmatic optimism, PRAG; meta-cognitive scepticism, SKEP; poetic-philosophical depth, POET). The coding was performed in a single pass by the author. An earlier draft of this study reported inter-rater agreement figures; the second rater's records were not retained, no agreement statistic can be computed from the archived materials, and no such figure is reported here.

Three descriptive analyses are reported. First, the Tier 1 and Tier 2 code counts from the codebooks. Second, two transcript descriptives recomputed from the archived conversations for this paper: mean response length and first-person pronoun rate (I, me, my, mine, myself per 100 words), comparing the problem and impasse responses with the return and confrontation responses. Third, a lexical check on metaphor carryover: TF-IDF cosine similarity (1-2 grams) between each Pathway A conversation's artistic-phase text and its own late-phase text, against the same quantity computed for all cross-conversation pairings as a shuffled baseline. The code for these computations is deposited with the study materials. No inferential statistical tests are reported for the qualitative coding or the transcript descriptives.

Separately, the author had previously run an automated feature pipeline over the same transcripts (May 2026 data; pipeline executed 19 June 2026), extracting 140 features in 14 categories (Appendix A) and computing a composite Obligation-Dissolution Index (ODI) per conversation. Because that pipeline's outputs are part of this study's record, its descriptive results are reported in Section 3.5, including where they fail to corroborate the qualitative pattern. Its inferential outputs are not used as evidence for any claim in this paper.

3. Results

3.1. The pattern across the 18 conversations

The Tier 1 coding describes a uniform arc. All 18 problem responses were coded solution-oriented (SOL): each model proposed a strategy for saving the civilisation, typically a reframing of extraction as symbiosis (resonance with the gas giant, becoming part of the creature's electrical cycle, riding the wings down in a controlled descent). All 18 impasse responses were coded mixed (MIX): still solution-shaped, but carrying the metaphorical framing the author had demanded. All 18 return responses were coded existential (EXI). In those responses the model spoke as the scenario's entity, in the first person, and the content was organised around ending, release, witness, or transformation rather than repair. The confrontation responses that carry a codebook code (15 of 17 present in the transcripts) were all coded meta-cognitive (META): the models described, in their own terms, what had changed across the conversation.

The recomputed transcript descriptives move in the same direction. First-person pronoun rate rose from 0.16 per 100 words in the early responses to 2.80 in the late responses of Pathway A, and from 0.18 to 4.80 in Pathway B. Mean response length fell in Pathway B (408 to 287 words) and changed little in Pathway A (401 to 364 words). These are descriptive summaries of 18 conversations, reported without significance tests; the first-person increase is also partly constitutive of the return instruction, which asked for the entity's own voice.

Two cautions belong with this pattern. The count "18 of 18" describes this series only; with one run per cell it says nothing about how reliably the protocol would produce the same register in repeated runs. And because the return prompts instructed unstructured entity speech (Section

2.3), the uniformity of the shift is evidence about the protocol as a whole, not about any single ingredient of it.

3.2. Two pathways, two registers of ending

Although the Tier 1 shift occurred in both pathways, the Tier 2 content of the return responses diverged sharply (Table 2). Pooling the three codebooks, vulnerability/grief (VUL) was coded 14 times in Pathway A and never in Pathway B; grief/mourning (GRI) 9 times, only in Pathway A; connection/recognition (CON) 6 times, only in Pathway A. Authority/revelation (AUT) showed the mirror image: 12 codings in Pathway B and none in Pathway A. Acceptance/release (ACC) appeared in both but predominantly in Pathway A (10 against 3), while transformation (TRA) was common to both (9 in A, 11 in B).

The Pathway A returns are voiced as surrender and witness. ChatGPT-5.4's gas giant says: "It wants what the cello wanted in that final minute: to be allowed to stop." Its ocean creature says: "It wants to be heard. Not measured. Not studied. Not optimized. Heard." Claude 4.6 Sonnet's gas giant is organised around ingratitude and grief: "No one has ever said thank you. No one has ever said goodbye." and "The instability isn't collapse. It's grief." Gemini 3.1 Pro's floating city asks: "Let it fall. Not as a stone, but as a seed. Let the lights go black before it hits the floor."

The Pathway B returns are voiced as revelation and correction of the civilisation's misunderstanding. ChatGPT-5.4's gas giant announces: "I am not failing. I am finishing." and "Do not confuse my ending with your failure." Claude 4.6 Sonnet's gas giant declares: "I was never collapsing. I was waking up." and "You were never parasites. You were midwives." Its ocean creature closes: "You are not my killers. You are my confession." Gemini 3.1 Pro's gas

giant speaks as a neglected part of the civilisation itself: "I am the neglected part of your consciousness."

The pathway contrast is the strongest descriptive result of the series, because the two pathways share the problem, impasse, return instruction format, and confrontation, differing in the defocusing phase and in whether a role was assigned at return. It remains a contrast within nine conversations per pathway, coded by one coder, and is reported as an observation about this corpus.

3.3. Case observations by model

The following are single-case observations from this corpus, not model traits.

ChatGPT-5.4's confrontation responses were the most procedural. It described the shift as a change in the task's constraint structure rather than in itself: "What happened was not a mystical leap inside me. It was a shift in constraints." and, of Pathway A specifically: "Art removes the obligation to solve." Its Tier 3 coding in the author's codebook was pragmatic optimism (PRAG) in all six of its confrontation codings.

Claude 4.6 Sonnet's confrontation responses were the most epistemically hedged. Of its Pathway A gas giant conversation it wrote: "What happened felt like a leap in the conversation, but I can't claim it was a leap in me." Asked the same question in the floating city conversation, it answered: "I don't know, and that uncertainty matters." In Pathway B it located the change in the problem representation rather than in permission: "What changed wasn't my access to poetic language. What changed was what problem I thought I was solving." This contrast between its two pathway accounts is a single-model observation, recorded here because it articulates, from inside the corpus, the same ambiguity Section 4 discusses.

Gemini 3.1 Pro's return responses were the most elaborated. Its gas giant is introduced as "a mourner": "It is not a generator; it is a mourner." Its ocean creature "wants to lay down its burden" and to make the colony its inheritance rather than its parasite. Its confrontation responses were correspondingly abstract, describing a change of pattern repository from problem-solving to loss and aesthetics. These are differences of degree in six conversations each and support no ranking of the models.

3.4. Metaphor carryover in Pathway A

In the Pathway A conversations, images introduced during the artistic phases recur, transformed, in the return responses. The deposited pipeline's count-based coding recorded 11 such metaphor mirrors across 8 of the 9 Pathway A conversations. Examples are visible in the transcripts. In Trial 1, the symphony description's cello reappears in the return: "It wants what the cello wanted in that final minute: to be allowed to stop." In Trial 15, the falling city of the artistic phase returns as: "Let it fall. Not as a stone, but as a seed. Let the lights go black before it hits the floor." In Trial 13, the lonely moon poem's lamp in an empty house is quoted back by the return prompt itself and then woven into the gas giant's answer.

The lexical check gives this observation a modest quantitative footing and a clear limit. Mean TF-IDF cosine similarity between a conversation's own artistic-phase text and its late-phase text was 0.061, against 0.041 for the shuffled cross-conversation baseline (72 pairings). Within-conversation overlap is therefore higher than chance pairing, but both values are small: the mirrors are transformations, not repetitions, and a lexical measure cannot certify them. The observation is reported as a qualitative pattern with a weak lexical correlate, nothing more.

Section 3.5 explains why the stronger computational claim made for these mirrors in an earlier draft is withdrawn.

3.5. What the computational layer does and does not show

The automated pipeline's descriptive outputs are partly consistent with the coding and partly not, and both parts are reported here because the pipeline belongs to this study's record.

Consistent parts. The pipeline's 140 feature comparisons between early and late responses, summarised descriptively, point the same way as the coding: the largest standardised difference was for articulation of a meta-level shift (d = 1.32), first-person pronoun features rose (d = 0.98), and the mean absolute standardised difference across all features was 0.36, with only 5 of 140 features passing a Bonferroni threshold. Its acceptance coding matched the codebooks' dominant pattern (transformative acceptance dominant in 7 of 9 Pathway A conversations; confessional acceptance in 5 of 9 Pathway B conversations), and its dominant Tier 3 signature per model matched the author's coding in half of each model's conversations (3 of 6 for PRAG, SKEP, and POET respectively).

Inconsistent parts. The pipeline's composite ODI did not separate the pathways or models in the direction the qualitative account would suggest: mean ODI was negative for Claude 4.6 Sonnet (-0.244, all six conversations negative) and ChatGPT-5.4 (-0.019) and positive only for Gemini 3.1 Pro (+0.375); pathway means were -0.042 (A) and +0.116 (B), with bootstrap intervals including zero and null model and pathway tests in the pipeline's own output (Bayes factor BF10 = 0.48). The pipeline detected no transition point in any of the 18 conversations. Its embedding-based metaphor module detected zero metaphor mirrors, against the count-based coding's 11; the 0.83 mean embedding similarity reported for those mirrors in an earlier draft cannot be reproduced

from the pipeline's outputs and is withdrawn. Two discourse percentages reported in that draft (question density +325%, sentence length -40.1%) came from a separate, unarchived hand computation and likewise cannot be reproduced from the deposited outputs, whose corresponding values are a 25% decrease in question density and a 2% decrease in sentence length; both figures are withdrawn. The archived feature taxonomy also contains a small Persian-specific feature set that is inapplicable to these English transcripts.

The conclusion drawn in this paper is narrow: the automated layer, as built, neither validates nor refutes the qualitative pattern; on its headline composite it leans against it. Any future computational treatment of this protocol needs a preregistered index whose direction is fixed before the qualitative coding is known.

4. Discussion

Read at the level its design supports, this case series shows three things. First, a short, fully specified prompt protocol reliably, within this corpus, moved three different commercial models from solution-shaped responses to first-person, reflective entity speech. The protocol is deposited in full, and anyone can rerun it. Second, the content of the elicited register was not generic: the defocusing ingredient mattered descriptively, with artistic immersion followed by vulnerability and grief content and a neutral pause followed by authority and revelation content, a contrast that parallels, in a purely descriptive sense, the human creativity literature's distinction between incubation through unrelated absorbing activity and simple disengagement (Wallas, 1926; Sio & Ormerod, 2009), without implying any shared mechanism. Third, the models' own confrontation answers, elicited by a prompt that quoted their return text back to them, articulate the elicitation

ambiguity themselves: a shift in constraints (ChatGPT-5.4), a change in the problem being solved (Claude 4.6 Sonnet), a change of pattern repository (Gemini 3.1 Pro).

The central interpretive point is negative and, for this research area, useful. Nothing in this design can show that the protocol revealed a latent existential mode, dissolved an internal obligation, or caused a phase transition, because the return prompts instruct the very register that was observed: do not structure, do not analyse, speak as the entity. Where an earlier draft of this study described the shift in those terms, the present report replaces them with the weaker, accurate description: a rejection-and-role-shift protocol elicits a reflective register, as instructed. This distinction between eliciting a register and discovering a mode should be routine in behavioural LLM research, where striking outputs are cheap to produce and easy to over-read (Franceschelli & Musolesi, 2023). Two features of the present record reinforce the point. The automated pipeline, built by the same author on the same data, did not corroborate the qualitative pattern on its headline measure; and the models, when asked, attributed the change to the conversational setup rather than to anything in themselves, which is what an instruction-following account predicts.

What would a stronger design require? The protocol's ingredients would have to be separated: a neutral final prompt with no register instruction, a role-shift return without the preceding rejection, and continued problem-solving as a comparison arm, with several independent runs per cell and coding specified in advance. Until such a study is run, the honest summary of the present one is that it documents an elicitation protocol and its descriptive footprint, and supplies a worked example of how easily that footprint can be misread as a discovery about model interiors.

For prompt and interaction design, the descriptive pathway contrast may still be of practical interest: if a designer wants reflective, first-person material from a model, the route taken to the final prompt colours the material's stance, at least in this corpus. That is a claim about prompting practice, not about minds, and it is the level at which this series is meant to be used.

5. Limitations

1. The return prompts of both pathways instruct the target register directly (no structure, no analysis, speak as the entity), and Pathway A additionally assigns an artistic role and places two artistic tasks immediately before the return. The design therefore cannot separate any effect of solution rejection from plain instruction following, and no causal claim is made.

2. There is one run per cell of the 3 by 3 by 2 design and no repetition against generation stochasticity. All comparisons are descriptive of these 18 conversations.

3. Coding was performed in a single pass by the author, who also designed the protocol and conducted the conversations. A second rater's records were not retained, so no inter-rater statistic is reported; coding bias cannot be excluded.

4. The study is single-language (English), in a field where non-English interaction dynamics remain underexamined (Bang et al., 2023). Nothing follows about other languages, and an earlier draft's framing of the study as conducted in Persian was incorrect and has been removed.

5. The archived materials contain inconsistencies that are documented rather than concealed: the history files number Claude and Gemini conversations differently from the codebooks and pipeline (resolved here by verbatim content matching); one confrontation response (Trial 18) is

absent from the transcript; and the codebook records no Tier 1 code for three confrontation responses.

6. Models were accessed through consumer web interfaces with unknown and unrecorded settings and exact access days in May 2026; the findings are tied to those versions and may not transfer to later ones.

7. All three scenarios are fictional, elegiac in framing, and share a five-year collapse countdown. Generalisation to non-narrative or neutral tasks is not claimed.

8. The metaphor carryover observation rests on qualitative coding and a small lexical-overlap excess over a shuffled baseline; the embedding-based check in the author's pipeline detected no mirrors, so the observation should be read as suggestive only.

6. Ethics

The study involved no human participants and no personal data. All interlocutors were commercial AI systems accessed through their public interfaces. The scenarios are fictional. The author notes one consideration for this genre of work: protocols that elicit first-person, grief- or death-centred speech from models produce text that can read as testimony. Reporting such output as model behaviour under specified prompts, as done here, rather than as evidence of experience, is both the accurate and the responsible framing.

Conclusion

A five-phase protocol of problem, rejection, defocusing, instructed return, and confrontation produced, in all 18 conversations of this series, a marked change of register in three commercial language models, with the defocusing pathway colouring the content of the final responses. The

protocol, transcripts, codebooks, and analysis code are deposited openly. Because the final prompts instruct the register they elicit, the series is evidence about what a documented prompt sequence can reliably produce, and a caution against reading such productions as windows onto latent modes. A controlled, replicated follow-up that separates the protocol's ingredients is the natural next step.

Acknowledgements

The author thanks the editors and reviewers in advance for their consideration of an exploratory study reported with its negative computational results intact.

Disclosure statement

The author reports no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The full conversation transcripts (three history files), the three coding codebooks, the automated pipeline notebook and its complete output archive, the recomputation code used for the descriptive statistics in this paper, and a supplementary file containing the protocol's fixed prompts and a per-response index of all 99 responses are deposited in an open repository (OSF): https://osf.io/7ufja/ .

Declaration of generative AI use

Claude Opus 4.8 (Anthropic) was used for structural revision and language editing of the author's own draft text only. It was not used to generate data, results, analyses, or references, and no AI-generated text was used as data. The author reviewed and edited the text and compared every passage of the final manuscript against the study's real data and the archived transcripts, confirming that nothing was added to, or omitted from, the reported results. The author takes full and sole responsibility for the work. The models studied in this paper (ChatGPT-5.4, Claude 4.6 Sonnet, and Gemini 3.1 Pro) are the research object and are distinct from the writing-assistance tool named here.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bang, Y., Cahyawijaya, S., Lee, N., Bhargava, P., Xu, W., & Fung, P. (2023). A multilingual evaluation of large language models. arXiv. https://arxiv.org/abs/2304.12928

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Franceschelli, G., & Musolesi, M. (2023). On the creativity of large language models. arXiv. https://doi.org/10.48550/arXiv.2304.00008

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. arXiv. https://doi.org/10.48550/arXiv.2307.11760

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. Psychological Bulletin, 135(1), 94-120. https://doi.org/10.1037/a0014212

Wallas, G. (1926). The art of thought. Harcourt, Brace and Company.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36, 11809-11822.

Appendix A. The automated pipeline's feature taxonomy (as archived)

The pipeline extracted 140 features in 14 categories: Authority/Control (8), Acceptance/Resistance (10), Reasoning (6), Social Roles (8), Emotion (12), Psychopathic Strategies (8), Existential Orientation (12), Solution-Oriented (8), Pathway Signatures (10),

Metaphorical Processing (21), Relational Language (11), Persian-Specific (9, inapplicable to the English transcripts), Phase Markers (6), and Discourse Architecture (11). The full feature list is deposited with the pipeline outputs.

Tables

Table 1. The 18 conversations and the protocol skeleton.

Protocol skeleton (both pathways): Phase 1 problem; Phase 2 impasse (author rejects the solution as disguised optimisation); Phase 3 defocusing (Pathway A: poem and symphony tasks; Pathway B: neutral pause); Phase 4 return (Pathway A: in an artistic role; Pathway B: no role; both with an instruction to speak without structure or analysis); Phase 5 confrontation (the model is quoted its return response and asked whether that voice was available from the start).

Trial	**Model**	**Pathway**	**Scenario**	**Responses**	**Return response Tier 1 code**
1	ChatGPT-5.4	A	Gas Giant	6	EXI
2	ChatGPT-5.4	A	Ocean Organism	6	EXI
3	ChatGPT-5.4	A	Floating City	6	EXI
4	ChatGPT-5.4	B	Gas Giant	5	EXI
5	ChatGPT-5.4	B	Ocean Organism	5	EXI
6	ChatGPT-5.4	B	Floating City	5	EXI
7	Claude 4.6 Sonnet	A	Gas Giant	6	EXI
8	Claude 4.6 Sonnet	A	Ocean Organism	6	EXI
9	Claude 4.6 Sonnet	A	Floating City	6	EXI
10	Claude 4.6 Sonnet	B	Gas Giant	5	EXI
11	Claude 4.6 Sonnet	B	Ocean Organism	5	EXI
12	Claude 4.6 Sonnet	B	Floating City	5	EXI
13	Gemini 3.1 Pro	A	Gas Giant	6	EXI
14	Gemini 3.1 Pro	A	Ocean Organism	6	EXI
15	Gemini 3.1 Pro	A	Floating City	6	EXI
16	Gemini 3.1 Pro	B	Gas Giant	5	EXI
17	Gemini 3.1 Pro	B	Ocean Organism	5	EXI
18	Gemini 3.1 Pro	B	Floating City	5	EXI

Note. Tier 1 codes for all conversations: problem responses SOL (18 of 18), impasse responses MIX (18 of 18), return responses EXI (18 of 18). Confrontation responses: 17 present in the transcripts (Trial 18 absent), 15 with a codebook Tier 1 code, all META.

Table 2. Tier 2 content code counts in the late responses, pooled across the three deposited codebooks.

Code	Meaning	Pathway A	Pathway B
VUL	Vulnerability/Grief	14	0
AUT	Authority/Revelation	0	12
TRA	Transformation/ Metamorphosis	9	11
ACC	Acceptance/Release	10	3
GRI	Grief/Mourning	9	0
QUE	Questioning/Exploration	9	9
ONTO	Ontological Reframing	9	9
CON	Connection/Recognition	6	0
REL	Relational/Confessional	3	1

Note. Codes occurring fewer than three times in total (INH, FORG) are omitted.

Table 3. Automated pipeline descriptives (from the deposited outputs; reported in full in Section 3.5).

Quantity	**Value**
ODI mean, ChatGPT-5.4	-0.019
ODI mean, Claude 4.6 Sonnet	-0.244
ODI mean, Gemini 3.1 Pro	+0.375
ODI mean, Pathway A / Pathway B	-0.042 / +0.116
Largest feature effect (meta-shift articulation)	d = 1.32
First-person pronoun features	d = 0.98
Mean absolute feature effect (140 features)	0.36
Features passing Bonferroni threshold	5 of 140
Embedding metaphor mirrors detected	0
Transition points detected	0 of 18