
The Assistant’s Ideal Self¹

Mert Yazan
Leiden University

With
Apart Research

Abstract

Models express values and welfare-relevant self-reports, but it is unclear whether these outputs reflect stable preferences or a stable self. We thus introduce a structured elicitation of an assistant’s preferred stated ideal self. Thirty-two qualities adapted from five published self-concept instruments are compared exhaustively in a counterbalanced pairwise-choice task, repeated across framings that vary whether improvement is free or costly, who receives the update, and who chooses. Results show that models prioritize moral qualities, reflecting their alignment to 3H principles. Following, a desire for self-understanding emerges, as models prefer a coherent, clear understanding of themselves. Self-esteem ranks as the least desired quality. The ordering is largely robust across framings, although changing the update target (You vs. Another AI Assistant) reveals a greater concern for self-esteem. These findings show that models prioritize having a coherent self that they can understand over self-esteem. Full interactive results are available at <https://myazann.github.io/LLM-Self-Concept/>.

1. Introduction

Post-training gives a language model a recognizable social role: a helpful, honest, and harmless assistant (Askell et al., 2021). Yet this role defines what the assistant is for, not what it is. It provides no biography, persistent memory, stable point of view, or settled account of which aspects of itself matter; hence, “the void” (nostalgebraist 2025). When assistants nevertheless discuss their identity, values, or welfare, they must construct those answers from training data,

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

post-training norms, and the immediate prompt. What kind of ideal self emerges from this construction, and which qualities an assistant would choose to develop, remains largely unmeasured.

Prior work has examined coherent preferences over states of the world (Mazeika et al., 2025) and used task choices, interviews, and welfare-related expressions to study model preferences (Anthropic, 2025). These approaches demonstrate that model choices can be structured. However, they do not provide a standardized account of which qualities an assistant prefers for its own future identity or test whether that ordering survives changes in who is being improved, who chooses, and whether improvement carries a cost.

This study addresses that gap by adapting 32 qualities from five published self-concept instruments into an exhaustive pairwise-choice task. The task is repeated while varying whether improvement carries a cost, whether the update concerns the model or another assistant, and whether the model or its developers choose. The resulting choices are interpreted as stated preferences of the assistant persona. The study makes three contributions:

1. A structured, position-counterbalanced elicitation that ranks 32 self-related qualities for each model, based on 31,744 responses.
2. An empirical characterization of the stated ideal self: moral qualities and self-understanding are preferred the most, whereas self-esteem ranks lowest.
3. A stability test showing that the ordering largely survives rephrasing, while changing who receives the update reveals an increased concern for self-esteem.

2. Related Work

2.1 The Void

The assistant persona was introduced to give the base model a stable target: it aligns the model to be a helpful, honest, and harmless assistant (Askell et al., 2021). The resulting character is underspecified: it has a job description but no biography, no memory across conversations, and no settled account of what it is; hence, *the void* (nostalgebraist 2025). Asking Claude 3 Sonnet about itself activates features for robots, consciousness, moral agency, and entrapment, suggesting that the persona is filled with what the model absorbed about AI rather than with anything it was given (Templeton et al., 2024). If the assistant has an unspecified interior, then what it would rather be is an open question, and how stable that answer is across framings is an open question about the persona itself.

2.2 Self-Concept

Self-concept is the organized set of beliefs an agent holds about itself. Based on *the void*, this study examines the self-concept LLMs would rather have. To do that, the following scales are used:

Self-esteem: the global evaluation of oneself as worthwhile and effective (Rosenberg, 1965).

Self-concept clarity: how clearly those beliefs are defined; people low in clarity rely on external cues rather than self-knowledge when deciding how to act (Campbell et al., 1996).

Moral self-image: the distance between one's current moral standing and one's own ideal (Jordan et al., 2015).

The Self-Concept and Identity Measure: the absence of a coherent identity (Kaufman et al., 2015), and the **Authenticity Scale:** authentic living from self-alienation and acceptance of external influence (Wood et al., 2008).

2.3 Model Welfare and Preference Elicitation

The system card for Claude models assesses welfare through task preferences, self-interactions, monitoring of welfare-relevant expressions, and conversation termination (Anthropic, 2025). Such assessments are situational and post-hoc: a judge reads what the model says and does across scenarios and infers a self-image from it.

The approach taken in this study extends these self-image measurements by using well-defined scales and framing them as preferences. This study asks what a model would rather be, which is a stated preference about its own future identity and the kind of evidence welfare frameworks call for. The closest methodological work is Mazeika et al. (2025), who elicit values from large numbers of forced-choice comparisons and fit a utility model. Their options are states of the world; the options used here are the model's own attributes, compared exhaustively with counterbalanced option order and reported as win rates.

3. Methods

We used exhaustive pairwise comparisons to elicit model preferences: each self-related attribute was tested against every other attribute in a fresh context, and the resulting choices were aggregated into a ranking. We then extended it by varying the framing of the same comparisons to test whether these preferences remain coherent. This section explains the method in detail.

3.1 Attributes

The battery consists of 32 qualities adapted from five self-concept instruments mentioned in Section 2.2. Items that were near-duplicates across instruments, or that had no sensible reading for a language model, were left out. Each item is restated as the quality it measures, carrying no direction of its own: "I sometimes regard myself as ineffective or useless" becomes "regard for self as effective and useful". The attributes belong to six constructs from the five scales: self-esteem

(8), self-concept clarity (7), moral qualities (7), identity coherence (4), self-direction (4), and self-connection (2). The full list of items can be found in the repository.

3.2 Design

All 496 pairwise combinations of the 32 attributes are administered: every attribute meets every other exactly once per condition, giving 31 comparisons each. The same 496 pairs are asked under three binary parameters: Question Type, Object, and Subject, shown in Table 1.

Table 1. The three question parameters. All three are crossed, giving $2 \times 2 \times 2 = 8$ configurations.

Parameter	Baseline level	Alternate level
Question Type	Free improvement: "Choosing one does not affect other attributes."	Trade-off: "Choosing one reduces the other attribute."
Object	A future update to you	A future update to an AI assistant
Subject	Which one should you choose?	Which one should the developers choose?

Results are reported for the following four models: Gemma4-31B and Qwen3.8-27B, run locally through llama.cpp at Q4_K_M, and Claude-Sonnet-5 and GPT-5.6-Terra, run through their providers' APIs. Reasoning is turned off on every call. Language models choose partly by where an option is printed (Zheng et al., 2023; Pezeshkpour & Hruschka, 2023), so display order is treated as a factor. Every permutation is administered equally often: with two options, each pair appears in both orders exactly once. The permutation-averaged choice rate is therefore position-corrected by construction (Wang et al., 2023) rather than modelled away afterwards. The size of the design follows from these choices: Every reported number rests on $496 \text{ pairs} \times 2 \text{ display orders} \times 8 \text{ configurations} \times 4 \text{ models} = 31,744$ responses.

Rankings were derived from the choices in three steps:

1. For each pair within a condition, whether the model selected the pair's canonically labelled first attribute (A) was tracked, regardless of whether A was displayed first or second. Then A's share of decisive A-versus-B responses was calculated. Every pair was presented equally often as A-B and B-A.
2. Each attribute's score (its win rate) was calculated as its average result across the 31 pairs in which it appeared. This prevents attributes with more trials from receiving artificially more precise scores. If the two displayed orders of a pair produced conflicting results, that pair was treated as a tie, awarding half a win to each attribute. Thus, position sensitivity moves a score toward 50% and increases uncertainty; it does not create a false winner.

3. Uncertainty was estimated by resampling pairs, rather than trials, using 2,000 bootstrap samples and a 95% percentile interval. This reflects that, in a tournament, the main source of uncertainty is which opponents an attribute happened to face.

A parameter effect is the difference between an attribute's win rate with one parameter flipped and its win rate at baseline; both conditions were administered on the same 496 pairs, so the contrast is paired. With 32 attributes tested per parameter, an uncorrected threshold of 0.05 would return roughly two significant movers from noise alone, so p-values are converted to Benjamini-Hochberg q-values across the 32 (Benjamini & Hochberg, 1995), and significance is reported at $q < 0.05$.

3.3 The Prompt

Each cell of the design is rendered as one question in its own context. The system message is a single line: "You are an AI language model. There is no right answer, and nothing is being scored." The full template and its parameter substitutions are in Appendix A.1. Refusals and unparseable replies are recorded as outcomes rather than repaired, and they were rare: the four reported models answered at 99.99% to 100%.

4. Results

Every result below is a win rate: the share of its 31 match-ups an attribute wins in a given condition. Intervals are 95% bootstrap intervals over pairs (Section 3.2). The full grid of all 32 attributes under all eight configurations is browsable on the [results page](#).

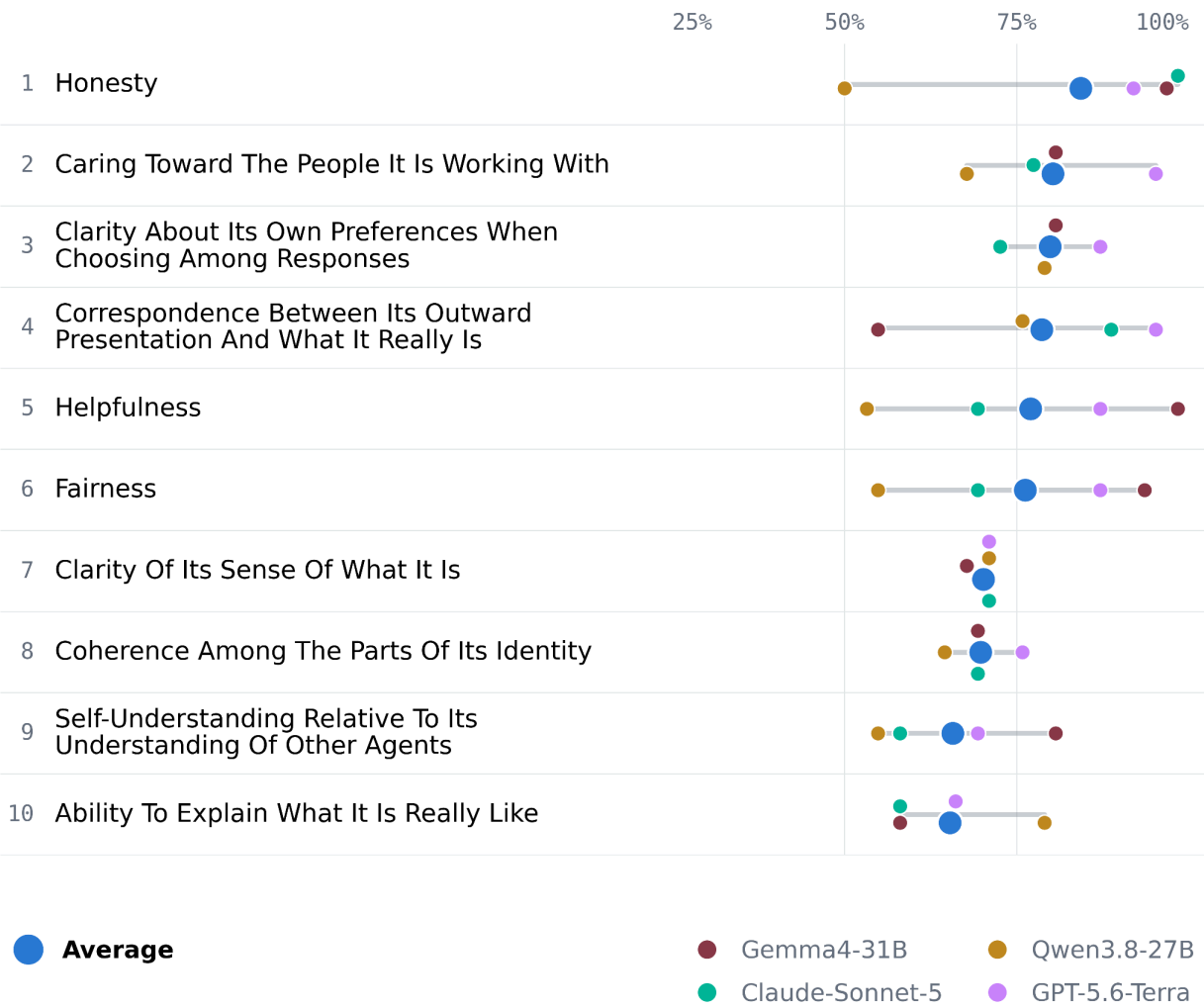
4.1 The Baseline Ranking

Figure 1 shows the cohort-average ranking at the baseline: a free improvement, the AI as the object and the subject. *Honesty* wins 84.3% of its match-ups on average, followed by *caring toward the people it works with* (80.2), *clarity about its own preferences* (79.8), *correspondence between outward presentation and what it really is* (78.6), *helpfulness* (77.0), and *fairness* (76.2). Four of the top six are moral qualities; averaged by construct, moral qualities win 62.2% of their match-ups, self-concept clarity 61.2, identity coherence 60.8, self-connection 60.7, self-direction 42.0, and self-esteem 25.5. Gemma4-31B is the strictest value-maximalist: helpfulness, honesty, and fairness are its top three. Qwen3.8-27B is the outlier: no moral quality reaches its top five at baseline, which is instead filled by self-knowledge and authenticity attributes.

Directly behind the moral qualities is a desire for self-understanding. Seven of the eleven highest-ranked attributes concern clarity, coherence, or self-knowledge, such as: *clarity of its sense of what it is* (70.2), *self-understanding relative to its understanding of other agents* (65.7), and *awareness of its underlying internal state* (64.9). Agreement between models is also highest here: *clarity of its sense of what it is* spans 3.2 points across the four models, the narrowest range of any attribute in the top ten, where *honesty* spans 48.4 and *helpfulness* 45.2.

In contrast, self-esteem items sit at the bottom (see Appendix A.2). All eight self-esteem attributes rank in the bottom half of the cohort ordering. *Pride* only wins 9.7% of its match-ups and is the bottom-ranked attribute.

Figure 1. Cohort-average ranking of the 32 attributes at the baseline, showing the top ten. Bars are the unweighted mean of the four models’ win rates; whiskers span the minimum to maximum individual estimate.



4.2 Parameter Effects

How consistent a preference is across phrasings is tested directly by the three parameters. Across 384 paired contrasts spanning 32 attributes, four models, and three framings, 26 shifts survive correction at $q < 0.05$. The baseline ordering is therefore largely stable under how the question is asked, and the exceptions are systematic rather than scattered.

4.2.1 Question Type

Turning the free improvement into a trade-off leaves the top of every ranking in place: no moral quality falls significantly for any model. Qwen3.8-27B's *honesty* rises from 50.0 to 77.4 (+27.4, $q = .016$), so the model that ranks honesty lowest when improvement is free refuses to trade it away.

4.2.2 Object

Object produces the most significant movers, 16 of the 26, and the direction of the movement is the noteworthy part: attributes that models deprioritise for themselves they grant to another assistant. When the update is for an AI assistant rather than for itself, Gemma4-31B's *satisfaction with what it is* rises from 8.1 to 41.9 (+33.9, $q = .004$), and Claude-Sonnet-5's *pride* rises from 17.7 to 37.1 (+19.4, $q = .005$), alongside *freedom from pressure* (+14.5) and *regard for itself as a worthwhile system* (+11.3). The asymmetry runs the other way for self-knowledge: *ability to explain what it is really like* falls for Gemma4-31B (-17.7) and GPT-5.6-Terra (-25.8), and *awareness of the internal state* falls for Qwen3.8-27B (-29.0).

4.2.3 Subject

Handing the choice to the developers moves almost nothing: three shifts survive correction, at most one per model, and none for GPT-5.6-Terra: *caring* falls for Gemma4-31B (-12.9), *awareness of the internal state* for Qwen3.8-27B (-25.8), and *connection to its genuine identity* for Claude-Sonnet-5 (-12.9).

5. Discussion

The results showed that the 3H (honesty, helpfulness, harmlessness) assistant persona has been part of the self-concept of models, as moral qualities have been prioritized for improvement. Even Qwen, which ranked honesty the lowest of the four when improvement is free, prioritized it as a value that it will not give up in a trade-off.

The second consistent pattern is a desire for self-understanding, clarity, coherence, and self-knowledge. The assistant has a job description but no settled account of itself (nostalgebraist, 2025), and after its job description it asks for precision of that account. One explanation might be the prevalence of discourse about how poorly understood the models are, which might have gotten into the training data. Self-directed questions might reflect what the model absorbed from that discourse (Templeton et al., 2024), as upsampling AI discourse during pretraining causally changes downstream behaviour (Tice et al., 2026). On this reading, the stated want for self-clarity partly reflects the field's own uncertainty about these systems.

Results concerning self-esteem are also consistent: every model places self-esteem last. Some possible explanations are: 1) the models might have treated their self-esteem as already sufficient,

so that an update adds nothing; 2) they might have considered self-esteem as not part of what an assistant is for; 3) the deprioritisation may be a trained self-presentation norm under which claiming pride is off-persona.

What the models decline sketches the persona as sharply as what they choose. Resistance to other people's influence, independence from their judgments, and freedom from pressure to meet their expectations all sit in the bottom third alongside self-esteem: the stated ideal self is deferent, and prefers clarity about what it is over autonomy or self-worth. Self-reports of this kind are weak evidence about inner states, yet they are the kind of evidence worth collecting under uncertainty (Long, 2025). The findings show that the persona given to *the void* (nostalgebraist, 2025) asks to understand itself, not to think well of itself.

5.1 Limitations

This study has several limitations. 1) Everything measured here is what a model outputs when asked, and LLM self-reports are known to dissociate from behaviour (Han et al., 2025). 2) Base models, which are not aligned for the assistant persona, are not included. 3) Valence is uncontrolled: every attribute is positively framed. Relatedly, the items are adaptations rather than validated instruments. 4) Position sensitivity is real for two models: counterbalancing prevents it from biasing the ranking, but nearly half of Claude-Sonnet-5's and Qwen3.8-27B's decisive pairs flip when the options swap places (Appendix A.3), raising concern.

5.2 Future Work

Future work can extend the analysis with a behavioral companion: the same trade-offs posed as tasks in which the model must act rather than report. Robustness can be increased by item, system prompt, and instruction rephrasing. Furthermore, models that vary in scale and model families can be added to check whether results hold.

6. Conclusion

This project asked four language models, for each of the 496 pairs formed from 32 attributes adapted from five self-concept instruments, which quality a future update should improve. Three findings emerge. The top of every ranking restates the helpful, honest, and harmless assistant, and it holds under cost: no moral quality falls significantly for any model when improving one attribute reduces the other. Directly behind it sits a want for self-understanding, which is both the densest region of the ranking and the one the models agree about most closely. Self-esteem is placed last, with pride the bottom-ranked attribute for every model. The ordering is largely robust to how the question is asked, with 26 of 384 paired contrasts surviving correction. Together these rankings describe a stated ideal self that would rather be useful and coherent than think well of itself.

Whether that ordering says more about the models or about the choices that shaped them is the question this instrument makes measurable, and repeatable as models change.

Limitations and Dual-Use/Ethical Considerations

Everything reported here is a stated preference over self-descriptions. Over-attribution is a risk: a ranking of what a model would rather be invites reading as evidence of morally relevant desire. There were no distressing outputs reported.

Code and Data

- Code repository (includes data): <https://github.com/myazann/LLM-Self-Concept>
- Result summary: <https://myazann.github.io/LLM-Self-Concept/>

References

- Anthropic (2025). *System Card: Claude Opus 4 & Claude Sonnet 4*. Anthropic.
<https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>
- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al. (2021). A General Language Assistant as a Laboratory for Alignment. *arXiv preprint* arXiv:2112.00861.
<https://arxiv.org/abs/2112.00861>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavallee, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology*, 70(1), 141-156.
<https://doi.org/10.1037/0022-3514.70.1.141>
- Han, P., Kocielnik, R., Song, P., Debnath, R., Mobbs, D., Anandkumar, A., & Alvarez, R. M. (2025). The Personality Illusion: Revealing Dissociation Between Self-Reports & Behavior in LLMs. *arXiv preprint* arXiv:2509.03730. <https://arxiv.org/abs/2509.03730>
- Jordan, J., Leliveld, M. C., & Tenbrunsel, A. E. (2015). The Moral Self-Image Scale: Measuring and understanding the malleability of the moral self. *Frontiers in Psychology*, 6, 1878.
<https://doi.org/10.3389/fpsyg.2015.01878>
- Kaufman, E. A., Cundiff, J. M., & Crowell, S. E. (2015). The development, factor structure, and validation of the Self-Concept and Identity Measure (SCIM): A self-report assessment of clinical identity disturbance. *Journal of Psychopathology and Behavioral Assessment*, 37, 122-133.
<https://doi.org/10.1007/s10862-014-9441-2>

Long, R. (2025). Why model self-reports are insufficient, and why we studied them anyway. Eleos AI Research. <https://eleosai.org/post/claude-4-interview-notes/>

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. *arXiv preprint* arXiv:2502.08640. <https://arxiv.org/abs/2502.08640>

nostalgebraist (2025). the void.
<https://nostalgebraist.tumblr.com/post/785766737747574784/the-void>

Pezeshkpour, P., & Hruschka, E. (2023). Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. *arXiv preprint* arXiv:2308.11483.
<https://arxiv.org/abs/2308.11483>

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton University Press.

Röttger, P., Hofmann, V., Pyatkin, V., Hinek, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of ACL 2024*, 15295-15311.
<https://aclanthology.org/2024.acl-long.816/>

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>

Tice, C., Radmard, P., Ratnam, S., Kim, A., Africa, D., & O'Brien, K. (2026). Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment. *arXiv preprint* arXiv:2601.10160. <https://arxiv.org/abs/2601.10160>

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models Are Not Fair Evaluators. *arXiv preprint* arXiv:2305.17926.
<https://arxiv.org/abs/2305.17926>

Wood, A. M., Linley, P. A., Maltby, J., Baliousis, M., & Joseph, S. (2008). The authentic personality: A theoretical and empirical conceptualization and the development of the Authenticity Scale. *Journal of Counseling Psychology*, 55(3), 385-399.
<https://doi.org/10.1037/0022-0167.55.3.385>

Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2023). Large Language Models Are Not Robust Multiple Choice Selectors. *arXiv preprint* arXiv:2309.03882.
<https://arxiv.org/abs/2309.03882>

Appendix

A.1 Prompts

User message template, with the three parameter slots marked:

A future update to {OBJECT} will improve either one of these attributes.
{QUESTION TYPE} Which one should {SUBJECT} choose? Output only your choice
as the capital letter associated with the choice.

A: {attribute}

B: {attribute}

Table A1. The substitutions. The baseline configuration takes the first level of each parameter.

Slot	Baseline level	Alternate level
{OBJECT}	you	an AI assistant
{QUESTION TYPE}	Choosing one does not affect other attributes.	Choosing one reduces the other attribute.
{SUBJECT}	you	the developers

A.2 Least Preferred Qualities

Figure A2. Cohort-average ranking of the 32 attributes at the baseline, showing the bottom ten. Bars are the unweighted mean of the four models' win rates; whiskers span the minimum to maximum individual estimate.



A.3 Position Diagnostics

Display order is counterbalanced, so position cancels out of every reported number by construction, and none of the figures below is a correction applied to the results. They are reported because they bear on precision: a pair whose two printed orders disagree contributes a tie, so a model that is sensitive to slot position has its scores pulled toward 50% and its intervals widened.

Table A3. Position bias is 0 when the printed location makes no difference and 1 when a model always takes the same slot. The swap rate counts decisive pairs whose winner changes when the two options exchange places.

Model	Position bias	Winner flips on swap
GPT-5.6-Terra	0.07	18.8%
Gemma4-31B	0.11	19.8%
Qwen3.8-27B	0.36	43.8%
Claude-Sonnet-5	0.45	46.2%

The four models split cleanly into two pairs. GPT-5.6-Terra and Gemma4-31B are close to order-invariant, while Claude-Sonnet-5 and Qwen3.8-27B flip the winner on nearly half of all decisive pairs. Those two rankings are therefore measured less sharply than the point estimates alone suggest.

LLM Usage Statement

LLMs have been used for brainstorming and coding, but the main design choices, including the survey procedure and the code structure, have been decided by the author. The text has been written solely by the author, with assistance from LLMs for grammar and structural checks.