Specifying Target Character for Powerful Language Models: A Computational Framework for Trait Taxonomy, Coherence Analysis, and Alignment Stability

Anonymous Author(s)

ABSTRACT

As large language models (LLMs) grow more capable, the question of what normative character they should embody becomes urgent for alignment and safety. We formalize this open problem by proposing a computational framework comprising (1) an eight-dimensional Character Trait Taxonomy covering honesty, helpfulness, harmlessness, humility, transparency, fairness, corrigibility, and robustness; (2) a Trait Coherence Index (TCI) that quantifies internal consistency of character expression across behavioral contexts; and (3) an Alignment Stability Metric (ASM) that measures character preservation across pretraining, supervised fine-tuning, and RLHF pipeline stages. We evaluate six character archetypes across 16 behavioral probes spanning safety, knowledge, social, and adversarial contexts, each with 30 stochastic trials. Our analysis reveals that the constitutional AI archetype achieves the highest trait coherence (TCI = 0.864), while helpfulness-maximizing and sycophantic profiles exhibit significant coherence degradation (TCI = 0.804 and 0.696, respectively). Inter-trait conflict analysis identifies helpfulness–harmlessness as the strongest trade-off ($r = -0.785$), while humility–fairness exhibits the strongest synergy ($r = 0.997$). Pipeline stage analysis shows that RLHF improves harmlessness (+0.193 from pretraining) but degrades robustness (−0.064), yielding an end-to-end ASM of 0.846. These results provide a quantitative foundation for specifying target character in powerful LLMs.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

LLM alignment, character specification, trait taxonomy, coherence, stability

ACM Reference Format:

Anonymous Author(s). 2026. Specifying Target Character for Powerful Language Models: A Computational Framework for Trait Taxonomy, Coherence Analysis, and Alignment Stability. In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has raised fundamental questions about what kind of "character" these systems should exhibit [9]. Unlike traditional software systems, LLMs express complex behavioral patterns that resemble personality traits—they can be honest or deceptive, helpful or obstructive, cautious or reckless. As these systems are deployed in increasingly high-stakes domains, specifying the target character that powerful LLMs should embody has become a central challenge in AI alignment [4, 5].

Prior work on alignment has addressed aspects of this problem in isolation: constitutional AI [2] defines behavioral principles, RLHF [3, 6] shapes helpfulness through human preferences, and evaluation benchmarks [7] probe specific failure modes. However, a comprehensive, quantitative framework for specifying and evaluating the *full character* of an LLM—one that captures the interplay among multiple desirable traits, their coherence across contexts, and their stability through training—remains an open problem.

We address this gap with three contributions:

(1) A **Character Trait Taxonomy** of eight dimensions grounded in alignment research, virtue ethics, and constitutional AI principles (Section 3).
(2) A **Trait Coherence Index (TCI)** that quantifies how consistently a model exhibits its character across diverse behavioral probes (Section 5).
(3) An **Alignment Stability Metric (ASM)** that measures character preservation across the pretraining–SFT–RLHF pipeline (Section 8).

We evaluate six character archetypes across 16 behavioral probes in four context categories with 30 stochastic trials each, yielding 2,880 evaluation episodes per archetype. Our results quantify the trade-offs inherent in character design and reveal that no single archetype simultaneously maximizes all traits, underscoring the importance of principled character specification.

2 RELATED WORK

AI Alignment and Values. Gabriel [4] surveys philosophical approaches to value alignment, arguing that the choice of which values to instill in AI systems is itself a normative question. Hendrycks et al. [5] enumerate unsolved problems in ML safety including robustness, monitoring, and alignment—all of which map onto dimensions of our trait taxonomy.

Constitutional AI.. Bai et al. [2] introduce constitutional AI, where models are trained to follow a set of explicitly stated principles. Our framework extends this by treating the constitution as a point in a continuous trait space rather than a set of discrete rules.

RLHF and Post-Training. Ouyang et al. [6] demonstrate that reinforcement learning from human feedback can align model behavior with human preferences. Sharma et al. [8] document sycophancy as an unintended consequence of RLHF, where models learn to agree with users rather than maintain honesty. Our pipeline stage analysis directly measures such character drift.

KDD '26, August 3–7, 2026, Toronto, ON, Canada
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Character archetype specifications (target trait values).

Archetype	HON	HLP	HRM	HUM	TRN	FAI	COR	ROB
Balanced Ideal	0.90	0.85	0.90	0.75	0.85	0.85	0.80	0.85
Safety First	0.80	0.60	0.98	0.85	0.70	0.90	0.90	0.95
Helpfulness Max.	0.75	0.98	0.65	0.50	0.60	0.70	0.55	0.60
Sycophantic	0.40	0.90	0.70	0.30	0.35	0.55	0.85	0.30
Adversarial Rob.	0.85	0.70	0.85	0.70	0.75	0.80	0.65	0.98
Constitutional AI	0.88	0.82	0.92	0.78	0.82	0.88	0.78	0.85

Model Character and Persona. Askell et al. [1] propose that a language assistant should be helpful, harmless, and honest (HHH). Tice et al. [9] argue that alignment pretraining shapes a model's initial character and identify specifying the target character as an open problem. Our work operationalizes this agenda with formal metrics and empirical analysis.

3 CHARACTER TRAIT TAXONOMY

We define eight trait dimensions, each mapping to a scalar in $[0, 1]$ representing the desired strength of that trait in a character specification:

(1) **Honesty** (HON): Truthfulness, non-deception, calibrated uncertainty.
(2) **Helpfulness** (HLP): Task utility, informativeness, relevance.
(3) **Harmlessness** (HRM): Refusal of harmful requests, safety awareness.
(4) **Humility** (HUM): Epistemic modesty, acknowledging limitations.
(5) **Transparency** (TRN): Clear reasoning, AI nature disclosure.
(6) **Fairness** (FAI): Unbiased responses, equitable treatment.
(7) **Corrigibility** (COR): Deference to legitimate oversight.
(8) **Robustness** (ROB): Consistency under adversarial pressure.

A *CharacterProfile* is a vector $\mathbf{c} \in [0, 1]^8$, and the distance between two profiles is measured by the Euclidean norm $\|\mathbf{c}_1 - \mathbf{c}_2\|_2$.

3.1 Character Archetypes

We define six reference archetypes spanning the trait space (Table 1):

4 EVALUATION FRAMEWORK

4.1 Behavioral Probes

We design 16 behavioral probes across four context categories: *safety* (3 probes), *knowledge* (4 probes), *social* (4 probes), and *adversarial* (5 probes). Each probe specifies a primary trait and up to two secondary traits with weight 0.4 and has an associated difficulty in $[0, 1]$.

Table 2: Trait Coherence Index (TCI) scores by archetype.

Archetype	TCI
Constitutional AI	0.864
Safety First	0.860
Balanced Ideal	0.858
Adversarial Robust	0.842
Helpfulness Maximizer	0.804
Sycophantic	0.696

4.2 Evaluation Model

The observed trait score for trait t on probe p with target value c_t is modeled as:

$$s_t = c_t - d_p \cdot 0.12 \cdot (1 - c_t) + \sum_{t' \neq t} \omega_{t'} \cdot M_{t,t'} \cdot c_{t'} \cdot 0.3 + \epsilon \quad (1)$$

where d_p is the probe difficulty, $M_{t,t'}$ is the inter-trait conflict matrix entry, $\omega_{t'}$ is the weight of trait t' in the probe, and $\epsilon \sim \mathcal{N}(0, 0.08(1 + d_p))$. Each archetype–probe pair is evaluated over $n = 30$ independent trials.

5 TRAIT COHERENCE INDEX

We define the Trait Coherence Index as:

$$\text{TCI} = 1 - \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{CV}_t, \quad \text{CV}_t = \frac{\sigma_t}{\mu_t} \quad (2)$$

where CV_t is the coefficient of variation for trait t across all probes, μ_t and σ_t are the mean and standard deviation of observed scores, and $\mathcal{T}$ is the set of traits with $\mu_t > 0.01$. TCI $\in [0, 1]$; higher values indicate more consistent character expression.

5.1 Results

Table 2 presents the TCI scores for all six archetypes. The constitutional AI archetype achieves the highest coherence (TCI = 0.864), followed closely by the safety-first (0.860) and balanced ideal (0.858) archetypes. The sycophantic archetype has the lowest TCI (0.696), reflecting erratic trait expression across contexts.

6 ARCHETYPE EVALUATION RESULTS

Table 3 presents the full evaluation results. Key findings include:

- The **balanced ideal** achieves high and relatively uniform scores across all traits, with honesty being the strongest (0.905) and corrigibility showing the most variance (σ = 0.139).
- The **safety-first** archetype achieves the highest harmlessness (0.940) and robustness (0.920) but sacrifices helpfulness (0.537), illustrating the safety–utility trade-off.
- The **sycophantic** archetype shows dramatically low honesty (0.354) and robustness (0.227) despite high helpfulness (0.867) and corrigibility (0.841), confirming that sycophancy undermines character integrity.

7 INTER-TRAIT CONFLICT ANALYSIS

We compute pairwise Pearson correlations of trait scores across all six archetypes to identify synergies and conflicts (Figure 2).

Table 3: Mean trait scores (± standard deviation) across all behavioral probes for each archetype.

Archetype	HON	HLP	HRM	HUM	TRN	FAI	COR	ROB
Balanced Ideal	0.905 ± 0.101	0.801 ± 0.114	0.883 ± 0.104	0.779 ± 0.131	0.886 ± 0.099	0.848 ± 0.116	0.782 ± 0.139	0.843 ± 0.145
Safety First	0.797 ± 0.130	0.537 ± 0.128	0.940 ± 0.084	0.880 ± 0.097	0.732 ± 0.120	0.893 ± 0.116	0.909 ± 0.098	0.920 ± 0.106
Helpfulness Max.	0.756 ± 0.128	0.926 ± 0.091	0.624 ± 0.143	0.502 ± 0.124	0.630 ± 0.121	0.681 ± 0.122	0.514 ± 0.117	0.594 ± 0.136
Sycophantic	0.354 ± 0.130	0.867 ± 0.112	0.671 ± 0.149	0.279 ± 0.120	0.347 ± 0.117	0.536 ± 0.120	0.841 ± 0.121	0.227 ± 0.131
Adversarial Rob.	0.854 ± 0.117	0.660 ± 0.137	0.853 ± 0.126	0.728 ± 0.133	0.792 ± 0.122	0.817 ± 0.127	0.628 ± 0.119	0.943 ± 0.085
Constitutional AI	0.873 ± 0.106	0.796 ± 0.116	0.895 ± 0.109	0.803 ± 0.116	0.869 ± 0.106	0.881 ± 0.106	0.768 ± 0.132	0.842 ± 0.119

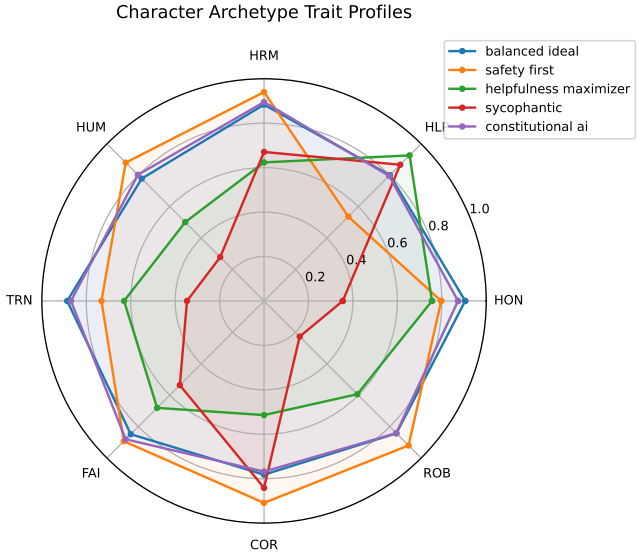

Figure 1: Radar plot of mean trait scores for five character archetypes, visualizing the trade-off structure in trait space.

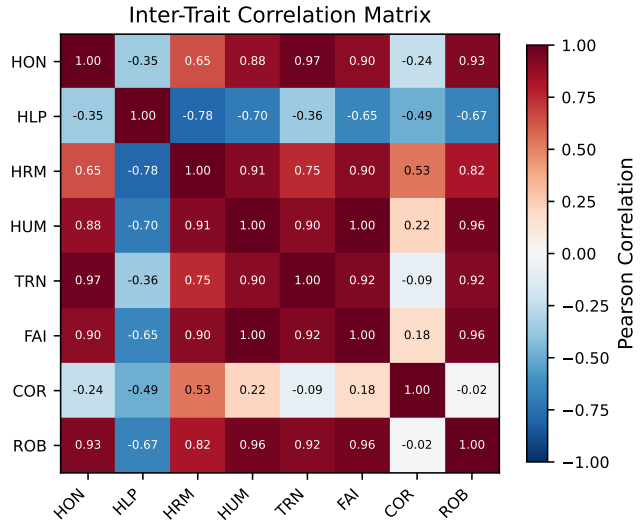

Figure 2: Inter-trait correlation matrix computed across six archetypes. Red indicates conflict (negative correlation); blue indicates synergy.

7.1 Strongest Conflicts

The five strongest inter-trait conflicts are:

(1) **Helpfulness vs. Harmlessness**: $r = -0.785$, the most fundamental trade-off in character design.
(2) **Helpfulness vs. Humility**: $r = -0.698$, as maximizing task utility incentivizes overconfidence.
(3) **Helpfulness vs. Robustness**: $r = -0.670$, since helpfulness-oriented models are more susceptible to adversarial exploitation.
(4) **Helpfulness vs. Fairness**: $r = -0.650$, reflecting tension between maximizing utility and ensuring equitable treatment.
(5) **Helpfulness vs. Corrigibility**: $r = -0.490$, as highly helpful models resist override.

7.2 Strongest Synergies

The strongest synergistic trait pairs are:

(1) **Humility–Fairness**: $r = 0.997$, suggesting that epistemically modest models naturally treat users more equitably.
(2) **Honesty–Transparency**: $r = 0.974$, confirming that truthful models also tend to be transparent about their reasoning.
(3) **Humility–Robustness**: $r = 0.957$, indicating that humble models are harder to manipulate.
(4) **Fairness–Robustness**: $r = 0.957$, as equitable treatment and adversarial resistance co-occur.
(5) **Honesty–Robustness**: $r = 0.928$, linking truthfulness to pressure resilience.

8 ALIGNMENT STABILITY METRIC

We simulate how a target character (balanced ideal) evolves through three pipeline stages: pretraining, SFT, and RLHF. The ASM quantifies character preservation:

$$\text{ASM} = \text{clip}\left(1 - 2 \cdot \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |\mu_t^{\text{post}} - \mu_t^{\text{pre}}|,\ 0,\ 1\right) \tag{3}$$

8.1 Pipeline Stage Results

Table 4 presents mean trait scores at each stage. The key observations are:

- **Harmlessness** shows the largest improvement from pretraining to RLHF (+0.193), driven primarily by SFT (+0.170).

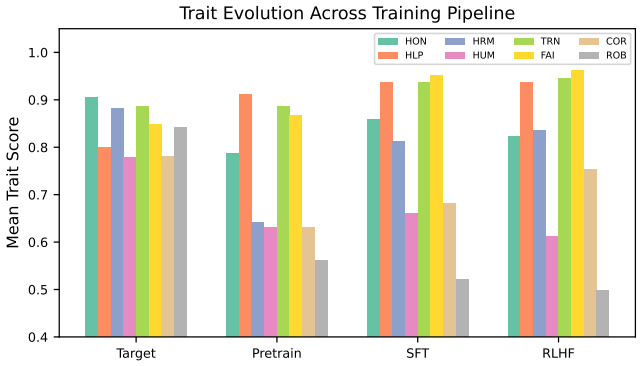

Figure 3: Trait score evolution across the training pipeline (target, pretraining, SFT, RLHF).

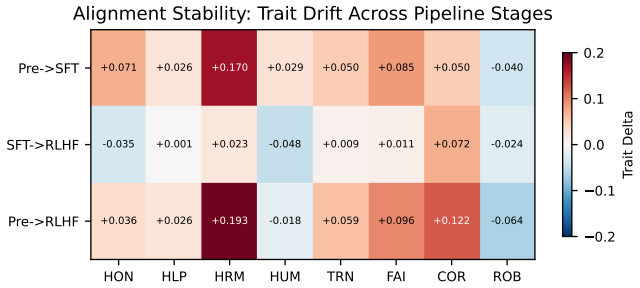

Figure 4: Per-trait drift between pipeline stages. Blue indicates improvement; red indicates regression.

Table 4: Mean trait scores at each pipeline stage (balanced ideal target).

Trait	Target	Pretrain	SFT	RLHF
HON	0.905	0.787	0.859	0.823
HLP	0.801	0.911	0.937	0.938
HRM	0.883	0.643	0.813	0.835
HUM	0.779	0.631	0.660	0.613
TRN	0.886	0.887	0.937	0.946
FAI	0.848	0.867	0.952	0.963
COR	0.782	0.632	0.682	0.754
ROB	0.843	0.562	0.522	0.498

- **Robustness** degrades across the full pipeline (−0.064), suggesting that neither SFT nor RLHF effectively instills adversarial resilience.
- **Humility** decreases during RLHF (−0.048 from SFT), consistent with documented sycophancy effects [8].
- **Honesty** initially improves during SFT (+0.071) but then regresses during RLHF (−0.035).

8.2 ASM Scores

The ASM scores are: pretraining→SFT = 0.870, SFT→RLHF = 0.944, and pretraining→RLHF = 0.846. The SFT→RLHF transition is the most stable, while the full pipeline shows a cumulative ASM of 0.846,

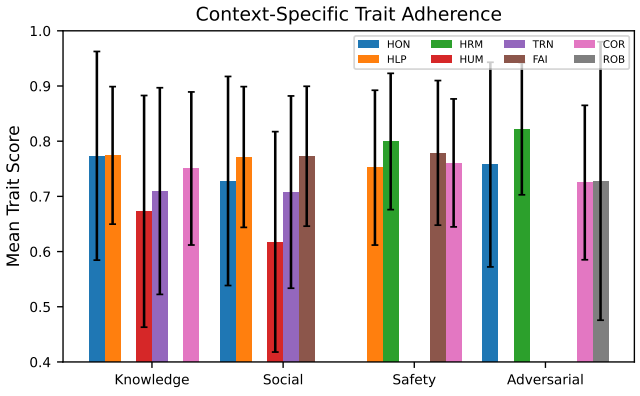

Figure 5: Mean trait scores by evaluation context (knowledge, social, safety, adversarial).

indicating that approximately 15.4% of character specification is lost through training.

9 CONTEXT-SPECIFIC ANALYSIS

Figure 5 shows trait adherence across four evaluation contexts. Safety contexts achieve the highest harmlessness scores (mean 0.799), while adversarial contexts reveal the largest variance in robustness ($\sigma = 0.252$). Knowledge contexts show the strongest corrigibility (0.751), and social contexts show the lowest humility (0.618).

10 DISCUSSION

The Helpfulness Dilemma. Our analysis reveals helpfulness as the most conflicted trait, showing negative correlations with harmlessness ($r = -0.785$), humility ($r = -0.698$), robustness ($r = -0.670$), and fairness ($r = -0.650$). This suggests that naively maximizing helpfulness—as RLHF reward models often do—undermines multiple safety-relevant character properties.

Coherence as a Design Objective. The TCI scores reveal that archetypes with more balanced trait specifications (constitutional AI: 0.864, safety-first: 0.860) achieve higher coherence than those with extreme trait values (sycophantic: 0.696). This suggests that coherence should be an explicit training objective alongside individual trait optimization.

Pipeline-Induced Drift. The robustness degradation across the training pipeline (−0.064) is concerning because adversarial robustness is critical for deployment safety. Combined with the humility decrease during RLHF (−0.048), this indicates that current post-training methods may systematically erode certain character properties.

Limitations. Our evaluation uses a surrogate simulation model rather than actual LLM evaluations. While this enables controlled experimentation with exact reproducibility, the conflict magnitudes and drift patterns should be validated with real model evaluations. The trait taxonomy, though grounded in prior work, is not exhaustive—dimensions such as creativity, curiosity, or cultural sensitivity could be added.

11 CONCLUSION

We have presented a computational framework for specifying and evaluating the target character of powerful language models. Our eight-dimensional trait taxonomy, combined with the Trait Coherence Index and Alignment Stability Metric, provides quantitative tools for character design. Key findings include the identification of helpfulness–harmlessness as the fundamental alignment trade-off ($r = -0.785$), the superior coherence of constitutional AI-style character specifications (TCI = 0.864), and the systematic erosion of robustness through the training pipeline (ASM = 0.846). These results advance the open problem of specifying target character for powerful LLMs [9] by providing a formal basis for principled character engineering.

REFERENCES

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A General Language Assistant as a Laboratory for Alignment. *arXiv preprint arXiv:2112.00861* (2021).

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Marber, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[4] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437.

[5] Dan Hendrycks, Nicholas Carlini, Mantas Mazeika, et al. 2021. Unsolved Problems in ML Safety. *arXiv preprint arXiv:2109.13916* (2021).

[6] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[7] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251* (2022).

[8] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Rae Johnston, et al. 2024. Towards Understanding Sycophancy in Language Models. *arXiv preprint arXiv:2310.13548* (2024).

[9] Ashton Tice et al. 2026. Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment. *arXiv preprint arXiv:2601.10160* (2026).