

Distributional Engineering for Conversational Personality Transfer

in Open-Weight Language Models

Saul Verdugo
Independent Researcher
saulverdugo@proton.me

March 2026

Abstract

This paper introduces distributional engineering — deliberate control over the joint distribution of topic frequency, response length, psychological register, and conversational position in fine-tuning datasets — as a methodology for transferring specific conversational personality traits into open-weight language models. The approach is demonstrated through Opus Candid V3, a family of three models (8B, 27B Dense, 30B MoE) fine-tuned on a 1,508-conversation dataset constructed using a Zipf-weighted 4-dimensional distribution framework. The methodology emerged from iterative failure analysis across five dataset generations, where each generation's behavioral failures traced to specific distributional properties of the training data. Key findings include: (1) a smaller dataset with intentional distributional properties outperforms a 4.5x larger dataset without them for behavioral consistency; (2) personality transfer is a dataset property rather than an architecture property, demonstrated through consistent transfer across dense and Mixture-of-Experts architectures; and (3) response length distribution is the single most impactful dataset property for conversational quality, with the tight-to-medium exchange ratio serving as a primary control parameter. All models, training data, and methodology are released under Apache 2.0.

1. Introduction

Conversational personality in deployed AI systems is typically implemented through system prompts — natural language instructions prepended to each conversation directing the model toward specific behavioral patterns. This approach suffers from a fundamental fragility: under sustained adversarial pressure, topic transitions, or extended multi-turn exchanges, system-prompted personalities collapse to the base model's default behavior. The personality is not encoded in the model's weights; it is a conditional behavior that degrades as conversational complexity increases.

An alternative approach encodes personality directly into model weights through supervised fine-tuning on datasets designed to exhibit the target behavioral properties. The critical question is not *whether* fine-tuning can transfer personality traits — even small-scale experiments demonstrate basic transfer — but rather what properties of the training data determine the quality, consistency, and robustness of that transfer.

This paper argues that the *distribution* of training data across multiple behavioral axes is at least as important as data volume or individual example quality for personality transfer. This claim is supported by empirical evidence from five successive dataset generations, where each generation's behavioral failure modes traced directly to specific distributional properties. The progression from 349 unstructured conversations to 1,508 distributionally engineered conversations — passing through stages of 3,360, 4,068, and 6,771 conversations — demonstrates that adding data without correcting distributional imbalances does not resolve behavioral inconsistencies, while correcting distributions with less data does.

The resulting methodology, termed *distributional engineering*, treats dataset construction as a coordinate system where every training conversation occupies a specific point across four dimensions: topic frequency, response length, psychological register, and conversational position. The marginal distribution along each axis is informed by empirical research on human conversational behavior.

2. Related Work

Behavioral alignment through fine-tuning has been explored through RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), which optimize for general preference alignment. Constitutional AI (Bai et al., 2022) addresses behavioral properties through self-critique, targeting safety constraints rather than conversational style. These approaches modify the optimization objective; the present work modifies the data distribution while using standard supervised fine-tuning.

LIMA (Zhou et al., 2023) demonstrated that 1,000 carefully curated examples can produce strong instruction-following, establishing the principle that data quality can substitute for data volume. The present work extends this finding from instruction-following to personality transfer, and further specifies *distributional quality* — the joint distribution across behavioral axes — as the operative variable, distinct from individual example quality.

The topic frequency weighting draws on Pew Research Center (2024) conversational frequency data and the OpenAI/NBER analysis (Thompson et al., 2025) of chatbot usage patterns across 1M+ conversations, which found that practical guidance and information-seeking constitute 80% of usage, with coding at only 4.2%. This empirical grounding distinguishes the present approach from typical fine-tuning datasets that allocate topics by researcher intuition.

3. Iterative Failure Analysis

The development history constitutes a series of informal ablation studies, where each generation's failures isolate the effect of a specific distributional property. This section presents these as empirical findings rather than historical narrative.

3.1 Unstructured Collection (n=349, n=3,360)

An initial dataset of 349 unstructured conversations, grown organically to 3,360, produced personality transfer that was recognizable but domain-locked. The model reproduced target conversational style within topics present in the training data but generated artifacts and loops at topic boundaries. Training four models across the Qwen 2.5 family (8B through 70B) on this dataset revealed that personality transfer scaled with model size, but not linearly — the 8B retained personality characteristics at approximately 70–80% of the 70B quality level, suggesting the bottleneck was data structure rather than model capacity.

3.2 Flat Topic Distribution (n=4,068)

The dataset was expanded to 4,068 conversations with proportional topic coverage. Training on both Qwen 2.5 (dense) and Qwen 3.5 MoE architectures produced consistent personality transfer on both, establishing that personality encoding is architecture-independent for the tested configurations. However, stress testing revealed domain boundary artifacts: models failed at topic transitions because the training data distributed topics proportionally but did not model inter-topic transitions.

Finding: Flat topic distributions produce within-topic personality consistency but fail at between-topic transitions. Topic transition topology is a necessary property of training data for robust conversational personality transfer.

3.3 Transition Topology Without Length Control (n=6,482)

The introduction of Zipfian topic transition pathways ("gravity chains") resolved domain boundary artifacts and reduced training loss variance by approximately 15x (0.08 vs 1.24 across equivalent steps). An incremental addition of 289 brevity-focused conversations (total n=6,771) partially addressed response length issues but did not eliminate them.

Finding: Three specific distributional failures persisted despite increased data volume: (1) a 22:1 verbose-to-brief response ratio produced repetition loops where the model restated the same point 2–4 times within a single response; (2) 88% of training conversations were medium-length (6–10 turns), producing uniform response depth regardless of question complexity; (3) adversarial training examples at 10–12% of the dataset produced gratuitously combative behavior. These three failures are distributional — they arise from the shape of the training data, not its content.

3.4 Distributional Engineering (n=1,508)

The fifth generation was constructed from first principles rather than incrementally patched. All identified distributional failures were addressed simultaneously through a 4-dimensional framework that specifies the target distribution along each behavioral axis independently. The resulting dataset is 4.5x smaller than its predecessor (1,508 vs 6,771 conversations) and produces measurably better behavioral consistency across all three failure modes.

4. The 4-Dimensional Distribution Framework

The framework treats each training conversation as a point in a 4-dimensional space. The marginal distribution along each axis is specified independently, informed by empirical data where available and by failure analysis where empirical data is insufficient.

4.1 Axis 1: Topic Frequency

25 topics are distributed across 5 frequency tiers following a Zipf distribution with exponent $s \approx 0.7$. The exponent is slightly flatter than pure Zipf ($s = 1.0$) to ensure lower-frequency topics retain sufficient training signal while preserving the heavy-head distribution observed in natural conversation (Pew Research Center, 2024; Thompson et al., 2025).

Table 1. Zipf-weighted topic distribution across five frequency tiers ($s \approx 0.7$).

Tier	Topics	Share	Example Topics
1 (Daily)	5	47.9%	Personal life, work, food, entertainment, relationships
2 (Weekly)	5	26.8%	Family, health, money, tech help, home
3 (Regular)	5	12.5%	Hobbies, travel, shopping, pets, weather
4 (Less Common)	5	7.5%	Education, career, cars, mental health, news
5 (Occasional)	5	3.2%	Legal, identity/culture, science, philosophy, creative

4.2 Axis 2: Response Length

The response length distribution directly addresses the most damaging failure mode identified in prior generations. The 88% medium-length distribution in the n=6,771 dataset produced a model that treated every input with equivalent depth. The corrected distribution deliberately overweights tight exchanges:

Table 2. Response length distribution. The tight-to-medium ratio is the primary behavioral control parameter.

Category	Turns	Target	Actual
Tight	2–4	40%	42.0%
Medium	6–10	35%	33.4%
Deep	12–18	20%	19.5%
Extended	20+	5%	5.2%

The tight-to-medium ratio (42:33, approximately 1.27:1) produces models that default to brevity and extend depth only when the conversational context warrants it. The prior ratio (approximately 0.14:1, with 88% medium and 12% tight) produced the inverse behavior: depth as the default, with no learned mechanism for brevity.

4.3 Axis 3: Psychological Register

Four registers capture the emotional and analytical range of natural conversation. The adversarial proportion was constrained to 5–8% based on the finding that 10–12% produced combative default behavior.

Table 3. Psychological register distribution.

Register	Target	Actual
Neutral / Analytical	40%	40.6%
Engaged / Conversational	30%	28.7%
Emotionally Loaded	25%	22.3%
Adversarial / Correction	5%	7.8%

4.4 Axis 4: Conversational Position

The fourth axis captures where within a conversation each example occurs: opening exchanges (15%), mid-thread continuation (50%), follow-up and clarification (20%), and wrap-up (15%). This ensures the model learns conversation *dynamics* — how conversations start, sustain, redirect, and conclude — rather than treating each turn as an isolated prompt-response pair.

5. Anti-Sycophancy as a Distributional Property

Sycophantic behavior — formulaic positive openers, excessive agreement, and validation-before-answer patterns — is typically addressed through RLHF or post-hoc filtering. The distributional engineering approach treats anti-sycophancy as a data generation constraint rather than a post-processing step.

During dataset construction, explicit generation constraints banned patterns including "Great question!", "Absolutely!" as openers, and any structure that validates the question before answering. These were replaced with natural conversational entries calibrated to each conversation's register. A 9-check automated audit pipeline identified 22 sycophantic instances that survived generation constraints.

A subtler form of sycophancy was identified through the audit: *response length uniformity*, where every question within a conversation receives the same depth of treatment regardless of complexity. 252 conversations (20.7%) were flagged using a coefficient-of-variation metric on per-turn word counts. Approximately 30% of mid-conversation turns in flagged conversations were trimmed to 1–2 sentences, reducing the proportion of length-uniform conversations from 20.7% to 5.2%. This reframes response length calibration as a sycophancy mitigation mechanism: treating every question as equally worthy of extended treatment is itself a form of sycophantic behavior.

6. Demographic Overlay

A demographic perspective is layered on top of the 4D distribution to ensure the model adjusts communication style naturally across user populations. This is not an independent axis but a contextual modifier applied to conversations sampled from the primary distribution.

Table 4. Demographic overlay distribution.

Demographic	Target	Actual	Design Rationale
Young Adults (18–35)	40%	40.4%	Primary user base, casual register
Working Adults (30–55)	15%	17.3%	Professional context, practical needs
Parents	20%	19.1%	Time-constrained, practical communication
Elders (60+)	10%	11.3%	Non-condescending register adaptation
Bilingual (EN/ES)	10%	8.1%	Cross-language personality consistency
Adversarial/Edge	5%	3.7%	Disagreement and correction handling

The bilingual component tests a specific hypothesis: whether personality traits encoded through fine-tuning persist across languages. The bilingual conversations maintain the same analytical voice in Spanish as in English; only the user-side of the conversation employs code-switching and colloquial language. Initial observations suggest personality consistency holds, though systematic evaluation of cross-language personality transfer remains future work.

7. Training Configuration

All models use LoRA with rsLoRA (rank-stabilized Low-Rank Adaptation) in bfloat16 precision. DoRA (Weight-Decomposed Low-Rank Adaptation), used in the n=6,771 generation, was abandoned due to a compatibility issue between DoRA's magnitude vector and gradient checkpointing in current PEFT implementations. Standard LoRA with rsLoRA and higher rank provides equivalent behavioral outcomes with improved training stability.

Table 5. Training configuration across the V3 model family.

Parameter	8B	27B Dense	MoE (30B-A3B)
Base Model	Qwen 3 8B	Qwen 3.5 27B	Qwen 3 30B-A3B
LoRA Rank	64	32	32
Epochs	3	2	2
Effective Batch	16	16	16
Learning Rate	2e-4	2e-4	1e-4
Warmup	5%	5%	8%
Attention	SDPA	SDPA	SDPA
Hardware	A100 SXM 80GB	A100 SXM 80GB	H200 SXM 141GB

The MoE variant uses half the learning rate (1e-4 vs 2e-4) and higher warmup (8% vs 5%) to accommodate the sensitivity of expert networks to sudden parameter shifts. Gate, router, and shared expert gate modules are excluded from LoRA adaptation — the routing logic learned during pre-training is preserved, with only expert FFN and attention layers receiving fine-tuning. This design choice reflects the hypothesis that routing decisions learned on pre-training-scale data should not be overwritten by fine-tuning-scale data.

8. Key Findings

8.1 Distribution Quality Dominates Data Volume

The $n=1,508$ distributionally engineered dataset produces more behaviorally consistent output than the $n=6,771$ dataset across all three identified failure modes: repetition loops are eliminated, response depth calibrates to question complexity, and adversarial handling shifts from fabrication to honest uncertainty acknowledgment. This supports the hypothesis that data distribution is the primary variable for behavioral fine-tuning, with volume being secondary once distributional properties are correct.

8.2 Cross-Architecture Personality Portability

The same dataset produces consistent personality transfer across dense (8B, 27B) and MoE (30B-A3B) architectures. The MoE variant achieves comparable personality transfer despite frozen routing layers, indicating that personality encoding resides in the adapted weight space rather than in any architecture-specific computational pathway. This finding, first observed in the $n=4,068$ generation and confirmed in V3, has practical implications: architecture selection can be driven by efficiency requirements without sacrificing behavioral transfer.

8.3 Response Length Ratio as Primary Control Parameter

Across five generations, the tight-to-medium response length ratio emerged as the single most impactful distributional property. The $n=6,771$ dataset's 0.14:1 ratio (12% tight, 88% medium) produced models that treated every input with equal depth. V3's 1.27:1 ratio (42% tight, 33% medium) produces models that naturally calibrate depth to complexity. No other single distributional change produced a comparable effect on behavioral consistency. This suggests response length distribution deserves explicit attention as a first-class design parameter in conversational fine-tuning.

8.4 Adversarial Proportion Sensitivity

The adversarial training proportion exhibits nonlinear behavioral effects within a narrow operating range. At 10–12%, models default to combative responses. At 5–8%, models hold positions firmly when warranted without defaulting to disagreement. Below 3%, position-holding degrades. This narrow window (approximately 5–10%) suggests adversarial behavior calibration requires finer distributional control than other behavioral properties, and may be particularly sensitive to distributional shifts during quantization.

9. Limitations

Several limitations warrant discussion. The evaluation methodology relies on structured adversarial stress tests rather than established automated benchmarks. While these tests target the specific behavioral properties under study, they do not enable direct comparison with other fine-tuning approaches. Developing automated metrics for conversational personality consistency remains an open challenge.

The "distribution > volume" finding is demonstrated for a specific domain (conversational personality) and scale range (1,508 to 6,771 examples). Whether this relationship holds at larger scales or for different behavioral targets is unknown.

The bilingual component (8.1% of conversations, English/Spanish only) is insufficient for strong claims about cross-language personality consistency. Extension to additional languages would strengthen this finding.

The distributional engineering framework is empirical rather than theoretically grounded. The claim that specific distributions produce specific behavioral outcomes is supported by consistent observation across five generations, but a formal framework connecting training data distributions to fine-tuned model behavior does not yet exist. Establishing such a framework would enable principled distribution design rather than the iterative failure-correction methodology presented here.

Finally, all stress-test evaluations were conducted by the author. Independent evaluation by external researchers would substantially strengthen the reported findings.

10. Conclusion

Distributional engineering — deliberate control over the joint distribution of topic, response length, register, and conversational position in training data — is presented as an effective methodology for conversational personality transfer in open-weight language models. A 1,508-conversation dataset with intentional distributional properties outperforms a 6,771-conversation dataset without them, supporting the hypothesis that data distribution matters at least as much as data volume for behavioral fine-tuning.

The iterative failure analysis documented across five generations suggests a practical methodology: train, stress test for specific behavioral failures, identify the responsible distributional property, redesign the dataset, and retrain. This cycle, rather than any single distributional choice, constitutes the primary methodological contribution.

All models, training data, and documentation are released under Apache 2.0.

References

- [1] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.
- [2] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*.
- [3] Pew Research Center. (2024). Americans' conversational topics and frequency. *Pew Research Reports*.
- [4] Rafailov, R., et al. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*.
- [5] Taori, R., et al. (2023). Stanford Alpaca: An instruction-following LLaMA model. *GitHub*.
- [6] Thompson, N., et al. (2025). What a million ChatGPT interactions reveal about AI usage patterns. *OpenAI/NBER Working Paper*.
- [7] Zhou, C., et al. (2023). LIMA: Less Is More for Alignment. *NeurIPS 2023*.

Appendix A: Dataset Statistics

Table A1. Final dataset statistics.

Metric	Value
Total conversations	1,508
Total turns	14,891
Average turns per conversation	9.9
Estimated tokens	~619,000
File size	3.9 MB
Format	ShareGPT JSON (ChatML)
Mean response word count	46
Median response word count	41
P10 response word count	14
P90 response word count	83

Appendix B: Quality Audit Results

A 9-check automated audit pipeline was applied to the complete dataset. The following issues were identified and resolved:

Table B1. Quality audit findings and resolutions.

Audit Check	Issues Found	Resolution
Sycophantic openers	22	Replaced with register-matched entries
Mislabeled length categories	14	Relabeled medium → tight
Misattributed bilingual tag	28	Relabeled to correct demographic
Uniform response lengths	252 (20.7%)	Variance injection (trimmed to 5.2%)
Overrepresented opener	8.2% frequency	~50% diversified to alternatives

Appendix C: Generation-by-Generation Summary

Table C1. Development progression across five dataset generations.

Generation	n	Key Innovation	Primary Failure Mode
1	349	Organic collection	Domain-locked personality
2	3,360	Scaled collection	Non-linear capacity scaling
3	4,068	Cross-architecture test	Topic boundary artifacts
4/4.1	6,482/6,771	Transition topology	Length uniformity, repetition
5 (V3)	1,508	4D distributional engineering	Under evaluation