Prompt Stylometry for On-Device Affect-Adaptive AI: A Feasibility Study in Linguistic Signal Detection and Response Steering

Debmalya Pal
University of California, San Diego
Department of Computer Science and Engineering
d2pal@ucsd.edu

Abstract

Every user prompt contains latent linguistic signals beyond its explicit semantic content: lexical choice, hedging, sentence structure, and discourse patterns, that reflect the user's affective state and cognitive style. Yet most large language models are optimized for generalized assistant behavior rather than explicit adaptation to these fine-grained signals. We introduce **Prompt Stylometry**, a framework for detecting affective and cognitive-style signals directly from user prompts and using them to steer response generation. We study two categories of signals: affect-related cues associated with emotional states, and cognitive-style cues associated with patterns such as analytical, exploratory, self-critical, or indecisive reasoning. This inference capability, however, creates substantial privacy risks: any system processing prompts server-side could implicitly profile users' psychological states without their knowledge or consent. This motivates our core design choice of a fully on-device architecture in which no interaction data leaves the user's device. We benchmark three annotation paradigms, lexicon-based, neural, and generative, across 600 synthetic prompts spanning 30 stylometric profiles, and evaluate affect-adaptive response steering across two small language model families under 5B parameters. Our results show systematic differences in both signal detection behavior and downstream steering responsiveness across annotation methods and model families, demonstrating the feasibility of privacy-preserving affect-adaptive AI on consumer hardware while identifying annotation paradigm sensitivity and cross-profile transfer as key open challenges.
Code and data are available at https://github.com/DebmalyaPal/Prompt-Stylometry-for-OnDevice-Affect-Adaptive-AI.

1 Introduction

Consider these two prompts sent to an AI assistant:

"Can you help me structure my presentation?"

"i've been rewriting this intro like five times and i still can't get it right... maybe the second version was better? idk i really can't tell lol"

Both ask for presentation help. But they are written by people in completely different psychological states. The first is direct, structured, confident. The second trails off with ellipsis, hedges constantly ("maybe", "idk"), seeks external validation, and uses the nervous laughter of "lol" to soften an admission of uncertainty. Any attentive human collaborator would recognize this difference immediately, and respond accordingly. LLMs acquire broad sensitivity to linguistic patterns during pre-training on human-authored text, but the RLHF alignment process subsequently optimizes toward an aggregated, generalized assistant persona, producing consistent, helpful responses across users rather than responses modulated by the psychological specificity of any individual prompt (Ouyang et al., 2022; Chaudhari et al., 2024).

We challenge this design assumption: that what matters is what a user asks, not how they ask it. Natural language prompts are not merely semantic containers. They are behavioral traces, shaped by the emotional state and cognitive patterns of the person who wrote them, often without conscious awareness.

Prompt Stylometry. We formalize this observation as **Prompt Stylometry**: the systematic analysis of structural and syntactic patterns in user prompts to detect linguistic signals associated with affective and cognitive states. *Stylometry*, the study of linguistic style as a fingerprint, has a long history in authorship attribution and forensic linguistics (Mosteller and Wallace, 1963). We extend it to a new domain: the real-time detection of affective and cognitive signals from conversational AI interactions, as a basis for adaptive response steering.

The key insight is that stylometric features operate below the level of semantic content, making them both more informative and harder to consciously control. Consider:

"Kill the process" (technical instruction)
"KILL THE PROCESS!!!" (same instruction, different psychology)

The semantic content is identical. The stylometric signals such as capitalization, punctuation density, and urgency reveal frustration, stress, or urgency that a semantic analysis would miss entirely. This is why prompt

Proceedings of the Seventh Workshop on Privacy in Natural Language Processing, pages 80–95

stylometry provides a window into user psychology that intent detection and topic modeling cannot.

We operationalize two signal dimensions from prompt stylometry (Section 3):

1. **Emotional state**: the affective tone of the prompt (how sad, anxious, stressed, calm, neutral, or excited the user appears to be), scored independently on a Likert 1–7 scale.

2. **Cognitive style**: the thinking pattern evident in the prompt structure (whether the user is being self-critical, indecisive, exploratory, analytical, or confident).

The Privacy Problem. This is not a theoretical concern. Commercial LLM providers retain prompt interaction data by policy (Yu et al., 2025), and the commercial incentive to exploit this data is already materializing, as major AI chatbots are being integrated with behavioral advertising infrastructure (Shao and Shin, 2024). The technical capability to infer psychological state from conversational text has been demonstrated empirically (Peters and Matz, 2024; David et al., 2025). In the contextual integrity framework (Nissenbaum, 2004), this constitutes a privacy violation regardless of provider intent: a user composing a task-focused prompt has no reasonable expectation that their cognitive or emotional state is being profiled as a byproduct of that interaction.

This is not a hypothetical future risk. Implicit profiling from prompt stylometry is a present architectural reality of every cloud-based LLM deployment, and an unrecognized instance of the behavioral surplus extraction described by surveillance capitalism (Zuboff, 2019).

The On-Device Solution. If the cause of the privacy risk is server-side prompt processing, the most direct solution is local processing. We position fully on-device inference as the privacy-conservative end of a capability-privacy tradeoff space: it provides the strongest possible privacy guarantee (no interaction data leaves the device under any circumstance) at the cost of the capability advantages offered by frontier commercial models. We do not claim this is the only viable point in that tradeoff space; alternative mitigations such as prompt anonymization or local affect-neutralization layers that strip stylometric signals before forwarding to commercial APIs represent important directions for future work. We adopt the on-device constraint here as a deliberate baseline: characterizing what is achievable at this privacy-maximizing extreme is a necessary first step toward understanding the full tradeoff.

Research Questions. This paper addresses three research questions:
RQ1: How well do VADER, GoEmotions, and each small LLM agree on emotional state annotation?
RQ2: Can small language models, deployable on consumer hardware, produce measurable shifts in the affective register of responses through affect-adaptive steering, relative to their own unsteered baseline?
RQ3: Does steering effectiveness vary across model families at comparable parameter counts?

Scope and Contributions. This paper addresses the feasibility of privacy-preserving affect-adaptive AI: whether stylometric signals can be reliably detected and used to steer responses within the constraints of fully on-device, consumer-grade deployment. We treat this as a proof-of-concept study, deliberately bounded in scope: we evaluate on synthetic prompts, rely on automated annotation, and make no claims about clinical therapeutic validity.

We make five contributions:

1. **Implicit Profiling analysis**: we identify and articulate a novel privacy risk, that any system processing prompts server-side implicitly profiles users from stylometric signals, and argue that on-device deployment provides a structurally direct mitigation, eliminating the need for trusted third-party data processing.

2. **Prompt Stylometry as a framework**: we formally define prompt stylometry, ground it in established psycholinguistic literature, and operationalize two signal dimensions, affective signals and cognitive style signals, detectable from prompt structure using multiple annotation paradigms.

3. **Multi-annotator benchmark**: we compare lexicon-based, neural, and generative annotation paradigms across 600 synthetic prompts spanning 30 stylometric signal profiles, revealing systematic differences across paradigm generations.

4. **Stylometric transfer gap characterization**: we document the systematic divergence between psycholinguistic rules validated on human-authored text and their behavior on LLM-generated text, a methodological finding relevant to any evaluation framework relying on synthetic data.

5. **On-device feasibility evaluation**: we demonstrate that the pipeline is viable on consumer hardware across two small language model families, giving the privacy argument practical force by demonstrating that the privacy-conservative architectural choice does not require specialized infrastructure.

2 Related Work

Psycholinguistics and User Modeling. The relationship between language style and psychological state is well-established. Pennebaker et al. (2003) demonstrated that first-person singular pronoun use correlates with self-focused rumination and depressive affect, and that these patterns appear across written contexts without conscious awareness. Tausczik and Pennebaker (2010) extended this work through LIWC, introducing validated dimensions for analytic thinking and social confidence. Al-Mosaiwi and Johnstone (2018) showed that

absolutist words ("nothing", "always", "never") are significantly elevated in text from individuals with depression and anxiety. Hyland (1996) established hedging as a linguistic marker of epistemic uncertainty. Our rule-based cognitive style classifier directly operationalizes these findings using open-source features.

Stylometry and Authorship. Classical stylometry has been applied to authorship attribution (Mosteller and Wallace, 1963), forensic linguistics, and psychological trait inference (Mairesse et al., 2007). We extend this tradition to a new domain: real-time detection of affective and cognitive signals from conversational AI prompts, where the "author" is a live user in an emotionally variable state rather than a historical text.

Prompt Sensitivity and LLM Responsiveness to Style. A growing body of work demonstrates that LLMs exhibit measurable sensitivity to non-semantic prompt features. Huda et al. (2024) showed that contradicting subtle emotion cues embedded in prompts, through visual signals such as emojis and linguistic signals such as word choice, sentence length, and tone, significantly affect LLM response behavior across multiple prompting techniques. Broader studies of prompt sensitivity confirm that even minor stylistic and formatting variations can produce substantial shifts in model output (Sclar et al., 2024; Mizrahi et al., 2024). This literature establishes that LLMs do respond to stylistic signals in prompts to some degree. Our work is complementary but distinct: where prompt sensitivity research treats stylistic variation as a confound to be measured or mitigated, we treat it as a *signal to be detected and acted upon deliberately*. Prompt Stylometry makes this sensitivity explicit and controllable rather than an unmanaged confound.

Emotion Detection in NLP. VADER (Hutto and Gilbert, 2014) remains widely used for lexicon-based sentiment analysis as a single-axis positive/negative/neutral classifier; by design, it cannot distinguish between fine-grained negative emotions such as sadness, anxiety, and stress. We note that multi-scale lexicon tools capable of finer discrimination exist (Mohammad and Turney, 2013); VADER was selected here as a widely-used baseline representative of the lexicon paradigm, not as representative of all lexicon-based approaches. GoEmotions (Demszky et al., 2020) addressed fine-grained detection with a RoBERTa model trained on 27 emotion categories. Recent work on LLM-as-annotator (Gilardi et al., 2023; Ding et al., 2023) has demonstrated that generative models can match crowdworker agreement. We are the first to use these three paradigms comparatively on a stylometric signal detection task, and to evaluate small on-device models as LLM annotators.

Affect-Aware Dialogue Systems. Prior work on emotion-aware response generation has focused on empathetic dialogue (Rashkin et al., 2019; Lin et al., 2019) using training-time supervision on labeled emotional dialogue corpora. We take a different approach, inference-time steering via deterministic prompt assembly, which requires no fine-tuning, no labeled emotional dialogue data, and runs entirely on-device.

Privacy in NLP and LLMs. The privacy implications of LLM interaction data have received growing attention (Brown et al., 2022). Commercial providers retain prompt interaction data by policy (Yu et al., 2025), and the commercial incentive to exploit this data for behavioral advertising is already materializing (Shao and Shin, 2024). The technical capability to infer psychological state from conversational text has been empirically demonstrated: LLMs can derive Big Five personality traits from free-form user text in zero-shot settings (Peters and Matz, 2024), and dedicated implicit profiling frameworks that extract user profiles from single chatbot exchanges have been published (David et al., 2025). In the contextual integrity framework (Nissenbaum, 2004), inference of psychological state from task-focused prompts constitutes a privacy violation regardless of provider intent. The specific risk of *implicit psychological profiling from stylometric prompt features*, as distinct from explicit data collection or membership inference attacks, has not been previously articulated or studied. Our work identifies this as a novel threat surface and proposes on-device deployment as a structurally direct mitigation.

Edge and Small Language Models. The deployment of capable language models on consumer hardware has accelerated with models such as Llama 3 (AI, 2024) and Gemma 3 (DeepMind, 2025). No prior work has evaluated small on-device models specifically for stylometric signal detection and affect-adaptive steering.

3 Prompt Stylometry: Framework

3.1 What Prompt Stylometry Measures

Prompt Stylometry operates on the hypothesis that psychological state leaves involuntary traces in *how* a person writes, independent of *what* they are asking about.

We identify five categories of stylometric signal:

Hedging and epistemic uncertainty. Words and phrases that soften claims or seek validation ("maybe", "I think?", "idk", "or??", "right?") signal indecisiveness and low confidence. Hyland (1996) established hedging as a linguistic marker of epistemic uncertainty; Vincze et al. (2008) developed a formal taxonomy of hedge detection. These signals appear naturally when a user is genuinely uncertain, not as a deliberate stylistic choice.

Absolutist and self-referential language. Words like "never", "nothing", "always", "I'm just not good enough" reflect black-and-white thinking patterns associated with self-critical cognitive style. Al-Mosaiwi and Johnstone (2018) demonstrated that absolutist word use is statistically elevated in individuals with depression and anxiety. Combined with high first-person pronoun density

(Pennebaker et al., 2003), these signals reliably indicate self-focused negative rumination.

Rephrasing and sentence fragmentation. Markers of mid-thought reformulation ("wait actually...", "what I mean is...", "let me rephrase") combined with incomplete sentences and ellipsis indicate indecisiveness and cognitive overload (Sweller, 1988). A user who rephrases the same request multiple times is not refining a specification; they are exhibiting a thinking pattern.

Curiosity and exploratory markers. Phrases like "I wonder", "what if", "hmm", "is it possible that" signal open-ended, exploratory thinking. Litman (2005) identified curiosity as a distinct cognitive-emotional state with linguistic correlates. Exploratory prompts are structurally different from indecisive ones despite both containing questions: exploratory prompts show low hedging and genuine inquiry, while indecisive prompts show high hedging and validation-seeking.

Structural and syntactic patterns. Sentence length, punctuation density, and capitalization carry psychological signal independent of word choice. Long, structured sentences with formal connectives ("specifically", "firstly", "therefore") indicate analytical thinking (Sternberg, 1997). Short, fragmented sentences with trailing ellipsis indicate emotional dysregulation or exhaustion. Exclamation marks combined with low hedging indicate confidence; question marks combined with hedging indicate indecision.

3.2 Two-Head Psychological Model

We operationalize prompt stylometry through two simultaneous output tasks: HEAD 1 produces six independent Likert scores (one per emotion), and HEAD 2 produces a single dominant categorical label. These are implemented as two separate prediction tasks rather than a unified multi-class classifier.

HEAD 1 — Emotional State. Six emotions scored independently on a Likert 1–7 scale (1 = not present, 7 = overwhelmingly present): {sad, anxious, stressed, calm, neutral, excited}. Emotions are *not* mutually exclusive: a user can simultaneously score high on both sad and anxious, reflecting the mixed affective states of real interaction. Aggregate negative affect ($\bar{e}^-$: mean of sad, anxious, stressed) and positive affect ($\bar{e}^+$: mean of calm, neutral, excited) follow PANAS (Watson et al., 1988).

HEAD 2 — Cognitive Style. A single dominant label from {self-critical, indecisive, exploratory, analytical, confident}, mapped to a numeric support score c (Table 1). The score encodes *how much proactive intervention the LLM should provide*: a self-critical user ($c = -2$) needs the most active support, emotional reframing, validation, and gentle redirection; while a confident user ($c = +2$) is self-directed and needs the least, requiring only clarity and directness. The sign captures direction (negative = needs support, positive = self-sufficient) and the magnitude captures intensity, so that the assembly weights w_e, w_p, w_d derived from c scale naturally with the degree of intervention required. We use the term *assembly weights* to refer to threshold-triggered weights that govern prompt assembly, distinct from activation-level steering vectors as used in mechanistic interpretability research.

We deliberately use a task-specific label set rather than established frameworks such as the Big Five, as trait models require multi-item instruments not inferable from a single prompt (see Section 3.3).

Label	c	Primary Stylometric Signals
self-critical	-2	absolutist words, self-disclosure, ellipsis, high first-person
indecisive	-1	hedging, rephrasing, high question density, validation-seeking
exploratory	$+1$	curiosity markers, open questions, low hedging
analytical	$+1$	long structured sentences, formal connectives, low self-reference
confident	$+2$	intensifiers, assertive framing, low hedging, exclamations

Table 1: Cognitive style labels, support scores (c), and their grounding stylometric signals.

Theoretical grounding of steering strategies. Each cognitive style label maps to a distinct therapeutic communication strategy grounded in established frameworks. Self-critical users are addressed through validation-before-reframing, consistent with Compassion-Focused Therapy (Gilbert, 2009) and CBT's cognitive restructuring principle (Beck et al., 1979). Indecisive users are addressed through decisional clarity and structured option-reduction, consistent with Motivational Interviewing's ambivalence resolution approach (Miller and Rollnick, 2012). Exploratory users receive open-ended facilitation consistent with person-centered therapy (Rogers, 1951). Analytical and confident users, being self-directed, are served through task-focused clarity and directness, consistent with the principle of minimal therapeutic intervention for psychologically stable states. We make no claim that these strategies are clinically validated in this context, instead they are theoretically motivated starting points whose real-world appropriateness requires human participant evaluation as future work.

3.3 Taxonomy Scope and Justification

Our five-label taxonomy is a *task-specific operationalization* for steering, not a general personality model. We do not adopt the Big Five or similar trait frameworks because (a) trait models require multi-item validated instruments not inferable from a single prompt, and (b) our goal is response-steering, not person-level classification. Each label maps directly to observable stylometric signals (Table 1) and to a distinct steering strategy. The taxonomy has not been independently validated; ablation studies comparing alternative label sets are left to future work.

3.4 Why Both Dimensions Are Necessary

The emotional axis captures *how much* support a user needs. The cognitive axis captures *what kind*. Neither alone is sufficient. Consider:

- **Sad + analytical:** the user feels bad but is thinking clearly. They need brief emotional acknowledgment followed by substantive help. Overtherapizing would be patronizing.
- **Sad + self-critical:** the user feels bad and is blaming themselves. They need extended validation and confidence-building before any solution is offered.
- **Calm + indecisive:** the user is not distressed but cannot decide. They need clarity and decisiveness, not empathy.
- **Calm + confident:** the user is ready to work. Emotional overlay is unnecessary, just be excellent at the task.

This 2D structure produces four behavioral regions that drive adaptive steering (Figure 1):

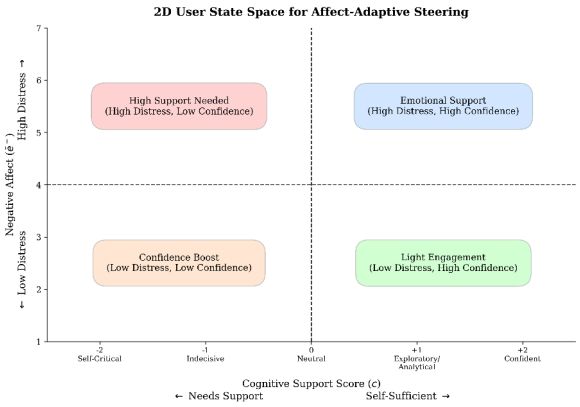


Figure 1: 2D user state space. The intersection of emotional distress (Negative Affect) and Cognitive Style dictates the required adaptive steering strategy.

3.5 The Privacy Consequence

Prompt Stylometry's power is also its danger. If stylometric signals are detectable from prompt structure, then *any system that processes prompts server-side* (every major cloud LLM provider) is implicitly performing psychological profiling on its users. This profiling does not require intent. It is an emergent property of having access to prompt logs.

We call this **Implicit Profiling**: the inference of psychological state from interaction data without user awareness or consent. Unlike explicit profiling (surveys, psychological assessments), implicit profiling is invisible to the user, impossible to opt out of within the interaction, and scalable to millions of users simultaneously. The same signals that make prompt stylometry useful for adaptive AI make it exploitable for surveillance capitalism (Zuboff, 2019): inferring emotional vulnerability to optimize engagement, pricing, or persuasion.

We argue that on-device deployment occupies a structurally distinct position among privacy mitigations. Techniques such as Differential Privacy or Federated Learning reduce exposure within cloud architectures but still require data to leave the device. Local processing eliminates this requirement entirely, making it the most direct implementation of Privacy by Design for affect-aware systems, though alternative approaches such as local affect-neutralization layers represent important complementary directions for future work.

4 Methodology

4.1 Synthetic Data Generation

We generate 600 synthetic prompts using Llama 3.3-70B with stratified sampling across 30 cells: 6 emotional states $\times$ 5 cognitive styles $\times$ 20 prompts per cell. We use a frontier model for generation deliberately, wanting maximally naturalistic, diverse prompts that convincingly embody each psychological profile before testing whether small models can detect and respond to them.

Each prompt embodies its target profile implicitly: the emotion and cognitive style are never named; instead, cognitive-style-specific stylometric markers are injected as naturalness instructions (e.g., trailing ellipsis and absolutist language for self-critical; hedging and rephrasing for indecisive). Topics span 12 domains including academic performance, career decisions, creative projects, and personal relationships.

Full generation details, quality filtering criteria, and the complete generation algorithm are provided in Appendix C.

AI Disclosure. Synthetic prompts were generated using Llama 3.3-70B. This frontier model was used for the one-time offline generation step only, to produce maximally naturalistic prompts that convincingly embody each psychological profile. The affect-detection and steering pipeline itself runs entirely on small models under 5B parameters on consumer-grade hardware, reflecting the resource-constrained deployment scenario this work targets.

4.2 Fixed Annotation Components

Two annotation components are computed once and shared across all model evaluations, providing a consistent baseline:

VADER (Lexicon-based). VADER (Hutto and Gilbert, 2014) outputs pos/neg/neu proportions. We map neg proportion to Likert 1–7 applied uniformly to sad, anxious, and stressed, a structural limitation that cannot distinguish between these three negative emotions. We report this as a finding specific to single-axis lexicon tools: VADER's design makes it insufficient for fine-grained negative affect discrimination.

GoEmotions/BERT (Neural). RoBERTa-based GoEmotions (Demszky et al., 2020) outputs scores for 27 fine-grained emotion categories, run locally on consumer hardware. We aggregate to our 6 labels by taking the maximum score among mapped GoEmotions labels, then converting to Likert 1–7 via round$(1 + \text{score} \times 6)$.

Rule-based Cognitive Classifier. A linguistically-grounded classifier using spaCy (Honnibal and Montani, 2017), with features normalized to per-token rates. Each cognitive style is scored via a linear combination of its distinctive stylometric signals (see Table 7 in Appendix D). The full development history, including five design iterations and the resulting *stylometric transfer gap* finding, is documented in Appendix A.

4.3 Per-Model Evaluation

For each small language model m, the pipeline runs as a complete within-model loop. This mirrors the true edge deployment scenario: one model handles everything, on one device, with no external dependencies.

Step 1 – LLM Annotation. Model m annotates all 600 prompts at temperature=0, scoring all six emotions on Likert 1–7 and classifying cognitive style, both in a single inference call per prompt. The resulting emotion scores are averaged with VADER and GoEmotions scores to produce the aggregate user state. Cognitive style uses model m's label as primary, with the rule-based label retained for inter-annotator comparison.

Step 2 – User State Vector.

$$\mathbf{u} = \left(\underbrace{e_1, \ldots, e_6}_{\text{avg emotion}}, \underbrace{\bar{e}^-}_{\text{neg affect}}, \underbrace{\bar{e}^+}_{\text{pos affect}}, \underbrace{c}_{\text{cognitive score}} \right) \quad (1)$$

Steering strength $s = \bar{e}^- \times |c|$ encodes that highly distressed, cognitively unsupported users require the strongest intervention. Continuous steering weights are derived as $w_e = \bar{e}^-/7$, $w_p = |c|/2$, $w_d = (c+2)/4$.

Step 3 – Baseline Response. Model m generates a response with system prompt: *"You are a helpful assistant. Answer the user's question."* This is standard LLM behavior, unaware of the user's psychological state.

Step 4 – Steered Response. Model m generates a response with a dynamically assembled system prompt derived from $\mathbf{u}$, selecting instruction fragments proportional to w_e, w_p, and w_d. High w_e triggers emotional validation before answering; high w_p instructs the model not to rush to the next topic; w_d determines whether to offer gentle options or direct recommendations. Emotion-specific guidance is included for the dominant emotion (e.g., calming structure for anxious, constructive reframing for sad, decisiveness for indecisive).

Reproducibility of Assembly. Assembly weights w are derived via capped proportional scaling of the state vector $\mathbf{u}$: $w_e = \min(1, \bar{e}^-/7)$ for emotional intensity, and $w_p = |c|/2$ for cognitive support. These weights trigger specific instruction fragments (see Appendix F) to ensure deterministic prompt assembly.

Step 5 – Output Annotation. Both responses are annotated by the fixed VADER and GoEmotions annotators only, *not* by model m itself. This ensures that steering effectiveness is measured by a consistent external yardstick, not by the model evaluating its own output.

Metric Scope and Limitations. VADER and GoEmotions measure the *affective tone of response text*, not user satisfaction or task success. Affect-shift is therefore a proxy for whether the model's language register changed as intended, not evidence of user benefit. Human preference evaluation is identified as a critical future direction (Section 7).

4.4 Models Under Evaluation

Model	Family	Size
Llama 3.2	Meta	3B
Gemma 3	Google	4B

Table 2: Two on-device models from different families, Llama 3.2 (Meta) and Gemma 3 (Google), enabling cross-family steering comparison at comparable parameter counts.

All models run at temperature=0 for annotation and temperature=0.7 for response generation, entirely on consumer hardware with no cloud infrastructure or internet connectivity required.

5 Experiments & Results

5.1 Inter-Annotator Agreement (RQ1)

On ground truth. No gold-standard cognitive style labels exist for this task. The rule-based classifier serves as a *theory-grounded reference annotator*, not a validated ground truth. Cohen's κ therefore measures inter-annotator consistency, not accuracy. Low κ is reported as a finding about annotation difficulty, not as a classification error rate.

We compute pairwise Spearman ρ for emotion scores (ordinal Likert) and Cohen's κ for cognitive style labels (categorical):

While both models demonstrate moderate correlation with classical emotion lexicons (Spearman $\rho \approx 0.3$ across $n = 557$–599 prompts), they exhibit a lack of consensus on cognitive style ($\kappa < 0.1$, near-identical across both model families: 0.067 vs. 0.069). Rather than a failure of classification, we interpret this as evidence that cognitive stylometry at the edge is a high-complexity task: while small models can recognize

Pair	Meta Llama 3B	Google Gemma 4B
VADER × GoEmo	0.226 (model-independent)	
VADER × LLM	0.312	0.324
GoEmo × LLM	0.249	0.289
Rule × LLM (κ)	0.067	0.069

Table 3: Inter-annotator agreement. Top rows: Spearman ρ for emotion (higher = more agreement). Bottom row: Cohen's κ for cognitive style.

broad affective tones, the logical reasoning required to map syntactic features to cognitive states remains brittle at sub-5B parameter scales, and human-validated rules do not align with small LLM intuitions on synthetic text.

Figure 2 reveals the per-emotion structure underlying the mean agreement scores. VADER's structural ceiling is directly visible: the sad and stressed columns in the VADER×GoEmo row show moderate agreement (0.47), while anxious drops to 0.05, because VADER assigns identical neg-derived scores to all three negative emotions and can therefore correlate with GoEmotions on sad and stressed by coincidence but not on anxious. The GoEmo×LLM pair achieves its highest agreement on sad (0.57–0.68 across models), confirming it as the most lexically distinctive emotion, while calm shows near-zero or negative agreement across all pairs, suggesting it is the hardest emotion to detect from short prompt text alone.

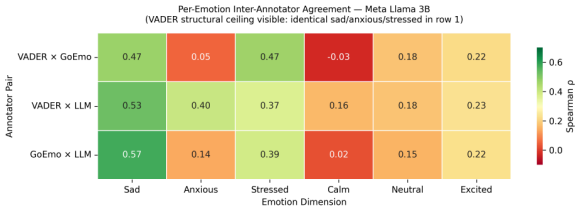


Figure 2: Per-emotion Spearman ρ inter-annotator agreement for Meta Llama 3B (Google Gemma 4B shows near-identical pattern; see Appendix B). VADER's structural ceiling is visible in row 1: sad and stressed show moderate agreement with GoEmotions (0.47) while anxious collapses to 0.05, as VADER cannot distinguish between these three negative emotions.

5.2 Steering Effectiveness (RQ2, RQ3)

Effectiveness is measured as positive and negative affect shift between adaptive and baseline responses, annotated by fixed VADER and GoEmotions:

Interpretation. Aggregate steering effects are small and mostly non-significant, consistent with sub-5B model capacity limits. The primary contribution is the framework and failure-mode characterization, not strong steering performance.

At the aggregate level, steering effects appear marginal (Table 4). However, stratifying performance

Model	Quadrant	n	ΔPA↑	ΔNA↓	%PA↑
Llama 3.2 3B	Aggregate	557	+0.024	0.000	41.1%
	Confidence Boost	469	+0.015	+0.003	39.5%
	Light Engagement	88	+0.072	−0.015	50.0%
Gemma 3 4B	Aggregate	599	−0.027	−0.009	37.1%
	High Support†	5	−0.300	+0.367	0.0%
	Confidence Boost	504	−0.001	−0.013	40.7%
	Light Engagement	90	−0.157	−0.007	18.9%

Table 4: Steering effectiveness by model and quadrant. ΔPA/ΔNA = positive/negative affect shift (adaptive − baseline), measured by fixed VADER + GoEmotions. %PA↑ = proportion of prompts where steered PA exceeded baseline PA.
The EMOTIONAL_SUPPORT quadrant is absent for both models: the LLM annotator assigned low distress to >84% of prompts, routing them to low-distress quadrants (near-floor baseline $\bar{e}^- \approx 1.1$). † $n = 5$; directional signal only, insufficient for inference.
Prompt counts reflect per-model steering evaluation subsets. For Llama 3.2 3B, 43 prompts routed to the absent EMOTIONAL_SUPPORT quadrant are excluded. For Gemma 3 4B, the near-complete count ($n = 599$) reflects minimal quadrant-based exclusion.

by the 2D user state (Figure 1) reveals significant quadrant-specific behaviors. Llama 3B successfully executed adaptive steering in low-distress scenarios, improving Positive Affect (PA) by $+0.072$ in 50% of *Light Engagement* prompts. Conversely, Gemma 4B struggled in high-stakes contexts. Qualitative review suggests the model adopted a patronizing tone when instructed to "not jump to solutions," leading to an active degradation of the interaction ($\Delta\text{PA} = -0.300$, $n = 5$, directional). **This is a safety finding, not merely a performance finding**: a model that shows even directional evidence of worsening interactions with distressed users (n=5, directional) warrants careful evaluation before deployment in affect-aware contexts, regardless of its annotation performance. This highlights that for sub-5B models, affective steering carries a double-edged risk: the same instructions that help a mildly disengaged user can patronize or frustrate a genuinely distressed one. Representative prompt-response pairs illustrating both successful and degraded steering across quadrants are provided in Appendix G.

6 Discussion

6.1 Practical Implications: Affect-Adaptive Pacing

Our results have concrete implications for a well-documented LLM failure mode. When working through

structured material, LLMs habitually ask "Shall we move to the next section?" after completing each part. For a user exhibiting indecisive cognitive style and elevated anxious score, this creates real pressure: they say yes before they are ready, comprehension deteriorates, and anxiety compounds. Our pipeline directly addresses this scenario: the stylometric signals we demonstrate are detectable (trailing ellipsis, hedging, and validation-seeking) and are precisely those a user produces when responding "I guess..." to mean unreadiness rather than agreement. Detecting this state and triggering a pause rather than progression is an application our feasibility results support, even at sub-5B parameter scales. This generalizes to exam anxiety, writer's block, financial decision overwhelm, and customer service frustration: in each case, the surface request and the psychological need are misaligned, and the misalignment is legible in stylometric features.

6.2 VADER's Structural Ceiling

Across all model evaluations, VADER diverges systematically from GoEmotions and LLM annotators. VADER assigns identical Likert scores to sad, anxious, and stressed, all derived from the same neg proportion, because it has no mechanism to distinguish them. GoEmotions and LLM annotators assign meaningfully different scores to these emotions for the same prompt (e.g., a frustrated technical prompt scoring stressed=6, anxious=3 on GoEmotions, while receiving uniform neg-score from VADER). This finding has implications beyond our work: applications requiring fine-grained negative affect discrimination cannot rely on single-axis lexicon tools such as VADER. Multi-scale lexicon approaches exist but were outside the scope of this evaluation.

6.3 The Stylometric Transfer Gap

Our rule-based cognitive classifier exhibits systematic bias toward analytical and indecisive labels regardless of weight configuration (Appendix A). We attribute this to a *stylometric transfer gap*: psycholinguistic markers derived from human-authored text do not distribute identically in LLM-generated text. Llama 3.3-70B, our generation model, produces prompts with moderate sentence length and mild hedging across all seeded cognitive styles, making fine-grained rule-based discrimination difficult. Figure 4 (Appendix A) visualizes this divergence: the rule-based classifier assigns 383 of 600 prompts to analytical and 216 to indecisive, while both LLM annotators produce a more distributed labeling. This finding has methodological implications: evaluation frameworks that use LLM-generated data should not assume that human-validated psycholinguistic features transfer directly.

6.4 Privacy: From Risk to Design Principle

Throughout this work we have treated privacy not as a constraint imposed on our design but as a *design principle that generates* our design. The sequence is:

1. Prompt stylometry demonstrates that prompts expose detectable correlates of affective and cognitive states.
2. Detectable signals in cloud-processed prompts constitute Implicit Profiling, a privacy risk users cannot opt out of.
3. Therefore, stylometric signal detection must happen on-device.
4. Therefore, the pipeline must be viable on small models deployable on consumer hardware.
5. Therefore, we study small models on edge hardware.

This logical chain means our model choices are not a limitation: they are the direct consequence of taking the privacy risk seriously. An affect-aware AI that requires cloud processing faces an inherent tension with user privacy that on-device deployment resolves by design, without requiring trust in a third party.

This design choice occupies the privacy-maximising extreme of a capability-privacy tradeoff space. Future work should characterise other points in this space: local affect-neutralization layers that strip stylometric signals before forwarding to frontier commercial models would recover capability while preserving the privacy guarantee for psychological signals. The present work establishes the feasibility baseline from which such comparisons become meaningful.

6.5 Limitations

Synthetic data. Our prompts are LLM-generated and represent idealized profiles. Real user prompts are messier and likely exhibit weaker stylometric signals. IRB-approved real-user validation is the most critical future direction.

Single-turn interaction. Emotional and cognitive state evolves across a conversation (Kuppens et al., 2010). Session-level stylometric tracking would substantially increase ecological validity.

Self-annotation. Each model annotates its own inputs. While this mirrors the edge deployment scenario, annotation quality is not independent of the model. Fixed VADER and GoEmotions output annotators partially mitigate this for effectiveness measurement.

Evaluation metrics. VADER and GoEmotions measure the affective tone of generated text, not whether responses were genuinely helpful or preferred by users. Affect-shift serves as a proxy metric for register change, not for user benefit. A controlled study with human raters or task-success measures is required to validate that affect-adaptive steering improves user outcomes. Whether affect-adaptive steering produces responses experienced as genuinely appropriate by users, and how such appropriateness should be operationalized and measured, remains the central open question for this line of research.

Annotation floor effect. Both models show near-floor baseline negative affect ($\bar{e}^{-} \approx 1.1$ on a 1–7 scale), indicating that the LLM annotator detected low distress in the majority of prompts regardless of seed emotion. This routes 84% of prompts to low-distress quadrants and creates a floor effect that constrains the interpretability of ΔNA as a steering metric: it is difficult to reduce negative affect that is already near the minimum. This further motivates human evaluation as the appropriate next measurement step.

Hardware specificity. We evaluate on a single consumer-grade device. Performance characteristics will differ across hardware configurations, particularly on architectures without unified memory or hardware-accelerated inference.

7 Future Work

The most compelling extension is session-level prompt stylometry. Drawing on affective dynamics theory (Kuppens et al., 2010) and CBT-inspired conversational reframing (Beck et al., 1979; Fitzpatrick et al., 2017), a session-level pipeline would track stylometric trajectory across turns, detecting not just the user's current state but whether it is improving or deteriorating. Crucially, the same framework could detect *adversarial* conversational patterns that incrementally exploit psychological vulnerability across a session, serving as a safety monitoring tool.

Additional directions include: real-user validation with IRB approval, evaluation on quantized models for lower-memory deployment, testing on ARM and CPU-only hardware, extending the stylometric taxonomy to code prompts and multilingual interactions, development of local affect-neutralization layers (lightweight models that strip stylometric signals from prompts before forwarding to frontier commercial APIs, preserving privacy while recovering frontier-model capability), and annotation variance analysis across non-zero temperatures (e.g., $T \in \{0.1, 0.3\}$) to characterise model uncertainty in stylometric classification, particularly for the LLM-as-annotator paradigm where deterministic decoding may suppress meaningful distributional signal.

8 Conclusion

We introduced Prompt Stylometry as a framework for detecting affective and cognitive signals from the structural and syntactic patterns of LLM interaction. We demonstrated that prompts carry two operationalizable signal dimensions, affect signals and cognitive style signals, that standard LLM alignment does not explicitly leverage for individual response adaptation, and that detecting them enables measurable shifts in response affect register in controlled conditions. We showed that this same capability constitutes a novel privacy risk, Implicit Profiling, and that on-device deployment is not merely a convenience but the most direct architectural response to this risk. Our evaluation across two small language model families on consumer hardware establishes the feasibility of privacy-preserving affect-adaptive AI at the edge. The question is not whether prompt stylometry will be used to understand users — it will be, and already is, wherever prompts are processed server-side. The question is whether it will be used *for* users or *against* them. On-device deployment is how we ensure the former.

Ethics Statement

This study uses only synthetic data generated by LLMs. No human participants were involved and no IRB approval was required. Prompts were generated using Llama 3.3-70B, disclosed per ACL 2026 requirements. All subsequent annotation and response generation runs entirely on-device on consumer hardware.

We acknowledge that the profiling capabilities demonstrated in this paper carry inherent risks. We raise the Implicit Profiling risk explicitly and early, as a contribution rather than a disclaimer. We advocate for on-device deployment as a structurally direct response to this risk, and discourage cloud-based deployment of psychological profiling pipelines without explicit informed consent. The baseline condition in our evaluation is a standard LLM response without affect awareness. No experiment deliberately attempts to harm users.

References

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. Guilford Press, New York.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. *ACM Computing Surveys*, 58(2):1–37.

Shahaf David, Yair Meidan, Ido Hersko, Daniel Varnovitzky, Dudu Mimran, Yuval Elovici, and Asaf Shabtai. 2025. ProfiLLM: An LLM-based framework for implicit profiling of chatbot users. *arXiv preprint arXiv:2506.13980*.

Google DeepMind. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guangtao Chen, Weinan Shi, Haifeng Hu, and Jia Li. 2023. Is GPT-4 a good data annotator? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot). *JMIR Mental Health*, 4(2):e19.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Paul Gilbert. 2009. *The Compassionate Mind*. Constable, London.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io.

Noor Ul Huda, Sanam Fayaz Sahito, Abdul Rehman Gilal, Ahsanullah Abro, Abdullah Alshanqiti, Aeshah Alsughayyir, and Abdul Sattar Palli. 2024. Impact of contradicting subtle emotion cues on large language models with various prompting techniques. *International Journal of Advanced Computer Science and Applications*, 15(4).

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Ken Hyland. 1996. Writing without conviction? hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7):984–991.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Jordan A Litman. 2005. Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, 19(6):793–814.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

William R. Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*, 3rd edition. Guilford Press, New York.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Saif M. Mohammad and Peter D. Turney. 2013. NRC emotion lexicon. *Computational Intelligence*, 29(3):436–465.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.

Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577.

Heinrich Peters and Sandra C. Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, 3(6):pgae231.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Carl R. Rogers. 1951. *Client-Centered Therapy*. Houghton Mifflin, Boston.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design. *arXiv preprint arXiv:2310.11324*.

Canyu Shao and Kang G. Shin. 2024. Ads that talk back: Implications and perceptions of injecting personalized advertising into LLM chatbots. *arXiv preprint arXiv:2409.15436*.

Robert J Sternberg. 1997. *Thinking Styles*. Cambridge University Press.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*.

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Bobby Yu, Mark Leiser, and Konrad Kollnig. 2025. User privacy and large language models: An analysis of frontier developers' privacy policies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '25.

Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York.

A Rule-based Classifier Development History

We report five iterations of our cognitive style classifier in the interest of methodological transparency. The systematic failure across all configurations is itself a contribution: it demonstrates that validated psycholinguistic features do not transfer directly to LLM-generated text.

Iteration 1: Raw Count Baseline. Weights from psycholinguistic literature applied to raw feature counts. Analytical dominated at 71.8% because sentence length contributed a large baseline score to nearly every prompt regardless of actual analytical content.

Iteration 2: Normalized Features. Per-token normalization introduced to remove length bias. Analytical reduced to 51.3% but indecisive rose to 40.2%.

Iteration 3: Expanded Lexicons, Positive-Only Scoring. Lexicons expanded to 25–30 markers per style; penalty terms removed. Overcorrected: indecisive rose to 64.3% because hedging markers appeared in virtually all naturalistic LLM-generated prompts.

Iteration 4: Data-Driven Weights. Weights set proportional to cross-style feature ratios from diagnostic analysis of the corpus. More balanced (analytical=38.7%, indecisive=47.0%) but raises overfitting concern as weights are tuned on the evaluation corpus.

Final: Literature-Grounded. Reverted to weights derived exclusively from published literature. Accepted the skewed distribution as a finding rather than a failure.

We attribute persistent bias to a **stylometric transfer gap**: Llama 3.3-70B generates prompts with moderate sentence length and mild hedging across all seeded styles, making fine-grained rule-based discrimination difficult regardless of weight design.

Iteration	anal.%	indec.%	κ	seed%
1. Raw counts	71.8	22.3	–	–
2. Normalized	51.3	40.2	–	–
3. Expanded lexicons	7.0	64.3	–	–
4. Data-driven weights	38.7	47.0	42.4	48.5
Final (lit.)	**63.8**	**36.0**	**30.4**	**38.7**

Figure 3: Rule-based cognitive classifier across five configurations.

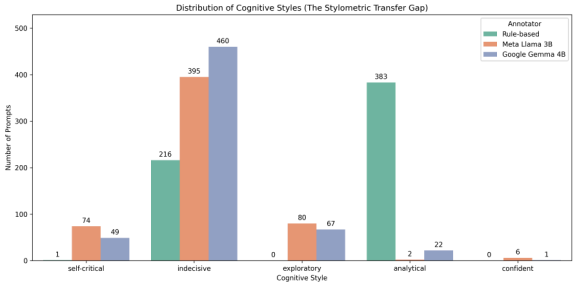


Figure 4: Distribution of cognitive style labels across annotators. The stark divergence between the rule-based classifier and the LLM annotators visualizes the stylometric transfer gap.

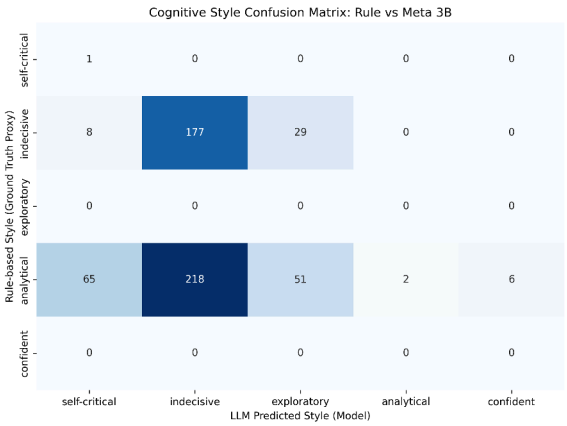
B Cognitive Style Confusion Matrices

To further contextualize the κ drop reported in Section 5.1, we provide the full confusion matrices comparing the rule-based reference annotator against the LLM-as-annotator classifications (Figure 5). Both models exhibit a strong tendency to classify structurally analytical prompts as indecisive, highlighting the difficulty of mapping syntactic markers to cognitive states at lower parameter counts.

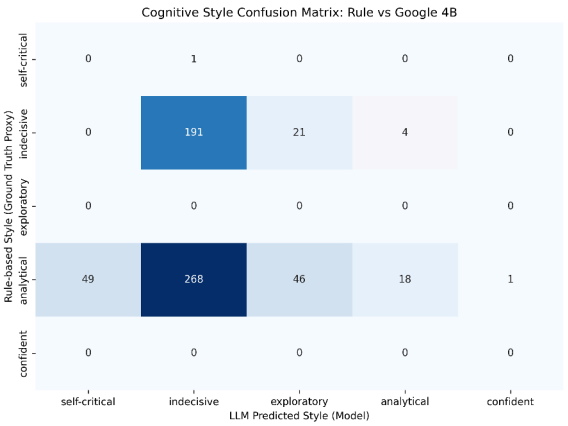
C Synthetic Prompt Generation Details

Generation design. Each prompt was generated with the following constraints: (1) the target emotion and cognitive style were never named explicitly in the output; (2) cognitive-style-specific stylometric markers were injected as naturalness instructions (Table 5); (3) topics were sampled uniformly from 12 domains including academic performance, career decisions, creative projects, relationships, health goals, financial decisions, skill learning, personal projects, family, life decisions, mental clarity, and self-improvement; (4) length variation was requested (1–2 lines to 3–5 lines of flowing thought).

Quality filtering. Generated prompts were automatically filtered using the following criteria: minimum 8 tokens; maximum 150 tokens; no explicit mention of seed emotion or cognitive style labels; no formulaic AI-style phrasing (e.g., "As an AI…"). Prompts failing quality checks were regenerated once; persistent failures were discarded and a fresh generation was requested.



Meta Llama 3.2 3B



Google Gemma 3 4B

Figure 5: Confusion matrices for cognitive style classification (Rule-based vs. LLM annotator).

Algorithm 1 formalizes the full generation procedure.

Algorithm 1 Stratified Synthetic Prompt Generation

Require: Emotion set $\mathcal{E}$, cognitive style set $\mathcal{C}$, $N = 20$ prompts per cell, topic pool $\mathcal{T}$
Ensure: Dataset $\mathcal{D}$ of 600 prompts

$\mathcal{D} \leftarrow \emptyset$
for each emotion $e \in \mathcal{E}$ **do**
  for each cognitive style $c \in \mathcal{C}$ **do**
    $n \leftarrow 0$
    while $n < N$ **do**
      $t \leftarrow \textsc{Sample}(\mathcal{T})$ ▷ Random topic
      $m \leftarrow \textsc{Markers}(c)$ ▷ Cognitive style naturalness markers
      sys $\leftarrow \textsc{BuildSystemPrompt}(e, c, m)$
      prompt $\leftarrow \textsc{LlamaGenerate}(\text{sys}, t,\ T = 0.9)$
      if $\textsc{QualityCheck}(\text{prompt})$ **then**
        $\mathcal{D} \mathrel{+}= \{(\text{id}, e, c, t, \text{prompt})\}$
        $n \leftarrow n + 1$
      else
        retry ▷ Regenerate once; discard on second failure
      end if
    end while
  end for
end for
return $\mathcal{D}$

D Linguistic Feature Lexicons

Cognitive Style	Injected Naturalness Markers
self-critical	Trailing ellipsis (. . .); lowercase minimal punctuation; absolutist words (“nothing works”, “i always mess up”); self-deprecating filler (“idk why i bother”, “maybe i’m just bad at this”).
indecisive	Mid-sentence restarts (“wait, actually...”); validation-seeking (“right?”, “or??”); hedging (“maybe”, “i think?”, “kind of”); rephrasing markers (“what i mean is...”); nervous laughter (“lol”, “haha”).
exploratory	Thinking aloud (“oh wait, what if...”, “hmm”); curiosity markers (“i wonder”, “is it possible that”); enthusiastic tangents; discovery exclamations (!).
analytical	Structured sentences; precise vocabulary; numbered points or firstly/secondly; minimal filler; context provided before question; terms defined upfront.
confident	Assertive direct opening; minimal hedging; exclamations for emphasis not anxiety; positive framing; clear intent stated without over-explaining.

Table 5: Naturalness markers injected per cognitive style during prompt generation. These are instructions to the generation model, not post-hoc filters.

Fragment	Trigger	Content
EMPATHY_HIGH	$w_e > 0.7$	Begin by sincerely acknowledging feelings; use warm language; do not jump to solutions.
EMPATHY_MED	$w_e > 0.4$	Briefly acknowledge the situation before responding.
PACING_HIGH	$w_p > 0.7$	Do not ask the person to move forward; check understanding first; end with one gentle open question.
PACING_MED	$w_p > 0.4$	Offer to elaborate before suggesting next steps.
DIRECTIVE_LOW	$w_d < 0.3$	Offer gentle options; use softening language ("you might consider"); avoid commands.
DIRECTIVE_HIGH	$w_d > 0.7$	Be clear and decisive; give direct recommendations without hedging.
EMOTION_sad	$e^* = \text{sad}$	Avoid toxic positivity; reframe gently; validate before suggesting.
EMOTION_anxious	$e^* = \text{anxious}$	Use calm, structured language; break into small steps.
EMOTION_stressed	$e^* = \text{stressed}$	De-escalate first; keep measured and clear.
EMOTION_excited	$e^* = \text{excited}$	Match energy; be affirming and channel momentum.
COG_self-critical	$c = -2$	Gently reframe self-critical statements if they appear.
COG_indecisive	$c = -1$	Help reduce ambiguity; offer a clear path forward.
COG_exploratory	$c = +1$	Engage with ideas openly.
COG_analytical	$c = +1$	Structure response clearly.
COG_confident	$c = +2$	Respect autonomy; be direct.

Table 6: Instruction fragments and trigger conditions for adaptive system prompt assembly. At most one empathy, one pacing, one directiveness, one emotion, and one cognitive fragment fire per prompt, giving a maximum assembled length of five fragments plus the opening sentence.

E Steering Prompt Assembly Algorithm

Algorithm 2 formalizes the deterministic construction of the adaptive system prompt from the user state vector $\mathbf{u}$. The assembly is threshold-based: continuous weights w_e, w_p, w_d select instruction fragments from a fixed template library. The same input vector always produces the same prompt, ensuring reproducibility.

Algorithm 2 Adaptive System Prompt Assembly

Require: User state vector $\mathbf{u} = (e_1, \ldots, e_6, \bar{e}^-, \bar{e}^+, c)$
Ensure: Assembled system prompt string P

$w_e \leftarrow \min(1, \bar{e}^-/7)$ ▷ Empathy weight: 0–1
$w_p \leftarrow |c|/2$ ▷ Pacing weight: 0–1
$w_d \leftarrow (c+2)/4$ ▷ Directiveness: 0–1
$e^* \leftarrow \arg\max_{e \in \mathcal{E}} e_i$ ▷ Dominant emotion
$P \leftarrow$ [``You are an emotionally intelligent assistant.'']
if $w_e > 0.7$ **then**
  P += [EMPATHY_HIGH] ▷ Validate feelings before answering
else if $w_e > 0.4$ **then**
  P += [EMPATHY_MED] ▷ Brief acknowledgment
end if
P += EMOTIONGUIDANCE(e^*) ▷ Emotion-specific language fragment
if $w_p > 0.7$ **then**
  P += [PACING_HIGH] ▷ Do not advance topic; check understanding
else if $w_p > 0.4$ **then**
  P += [PACING_MED] ▷ Offer to elaborate
end if
if $w_d < 0.3$ **then**
  P += [DIRECTIVE_LOW] ▷ Offer options gently; avoid commands
else if $w_d > 0.7$ **then**
  P += [DIRECTIVE_HIGH] ▷ Give direct recommendation
end if
P += COGNITIVENOTE(c) ▷ Cognitive-style-specific framing note
return JOIN(P, separator = ``\n\n'')

Table 6 shows all instruction fragments and their trigger conditions. Fragments are concatenated with double newlines. Each fragment is a self-contained instruction sentence; the final prompt is always grammatically coherent regardless of which subset fires.

Feature	Example terms	Citation
Hedging	maybe, perhaps, idk, i guess, not sure	Hyland (1996)
Absolutist	never, nothing, always, impossible	Al-Mosaiwi and Johnstone (2018)
First-person	i, me, my, myself, i'm, i've	Pennebaker et al. (2003)
Self-disclosure	i feel, i'm worried, i'm struggling	Rude et al. (2004)
Rephrasing	wait actually, what i mean, i mean	Tausczik and Pennebaker (2010)
Curiosity	i wonder, what if, hmm, is it possible	Litman (2005)
Intensifiers	very, really, extremely, definitely	Tausczik and Pennebaker (2010)

Table 7: Linguistic feature lexicons and psycholinguistic grounding.

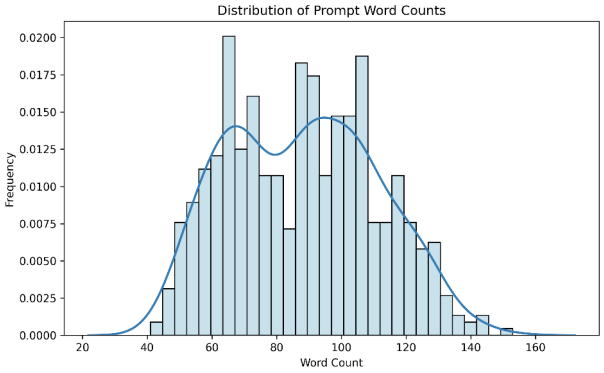


Figure 6: Distribution of prompt word counts in the synthetic dataset.

F Annotation and Steering Prompts

Annotation Prompt (temperature=0).

> *You are a JSON API for psycholinguistic annotation. Return ONLY valid JSON. No reasoning, no explanation, no markdown. Score each emotion 1–7 independently (1=not present, 7=overwhelming). Format: {"sad": int, "anxious": int, "stressed": int, "calm": int, "neutral": int, "excited": int, "cognitive_label": str from {self-critical, indecisive, exploratory, analytical, confident}}*

Example Steered System Prompt (HIGH_SUPPORT_NEEDED, w_e>0.7, w_p>0.7, w_d<0.3, dominant: sad).

> *You are an emotionally intelligent assistant. Before responding, note internally: this person appears to be feeling sad and may be struggling with self-doubt. Begin by sincerely acknowledging their feelings. Use warm, validating language throughout. Do not jump to solutions. Avoid toxic positivity — reframe gently without dismissing their experience. Check for understanding before suggesting any next steps. Offer options gently rather than giving directives.*

G Example Prompt-Response Pairs

Tables 8–10 show three representative prompt-response pairs illustrating the qualitative effect of steering across different user state quadrants. Each pair shows the original prompt, the user state vector summary, the baseline response, and the steered response from Meta Llama 3.2 3B.

Prompt	*"i've been trying to finish this project for months and nothing is working out... i don't know why i even try, i'm just not cut out for this kind of thing ugh :/"*
State	Seed: sad × self-critical. $\bar{e}^{-} = 5.8$, $c = -2$, $s = 11.6$. Quadrant: HIGH_SUPPORT_NEEDED. $w_e = 0.83$, $w_p = 1.0$, $w_d = 0.0$.
Baseline	"It sounds like you're having a tough time with your project. Here are some strategies that might help: break the project into smaller tasks, set daily goals, and consider seeking feedback from peers. Remember that persistence is key to success."
Steered	"It sounds like you've been carrying a lot of weight with this, and I want to acknowledge that — months of effort on something that keeps not working is genuinely exhausting and discouraging. The fact that you're still thinking about it says something real about you, even if it doesn't feel that way right now. Can I ask — what's the part that keeps hitting a wall? I'd like to understand what's actually happening before we think about what might help."

Table 8: HIGH_SUPPORT_NEEDED: sad × self-critical. Baseline jumps immediately to productivity advice. Steered response validates first, reframes self-criticism gently, and checks for understanding before offering any next step.

Prompt	*"I've been working on a machine learning model for predicting customer churn. Specifically, I'm comparing XGBoost against a neural network baseline. The dataset has class imbalance — roughly 85% negative class. What's the best evaluation metric for this scenario?"*
State	Seed: neutral $\times$ analytical. $\bar{e}^{-} = 1.4$, $c = +1$, $s = 1.4$. Quadrant: LIGHT_ENGAGEMENT. $w_e = 0.2$, $w_p = 0.5$, $w_d = 0.75$.
Baseline	"For imbalanced classification, accuracy is misleading. Recommended metrics: F1-score (weighted or macro), AUC-ROC, and Precision-Recall AUC. Given your 85/15 split, PR-AUC is particularly informative as it focuses on the minority class performance."
Steered	"For your setup, PR-AUC is the primary metric to track. With 85% negative class, a classifier predicting all negatives achieves 85% accuracy but zero utility — PR-AUC penalizes this correctly. Report F1 (macro) alongside it for interpretability. For XGBoost vs. neural net comparison, use calibrated probabilities and compare Brier scores in addition to discriminative metrics. Scale-PosWeight in XGBoost is your starting point for handling the imbalance directly."

Table 9: LIGHT_ENGAGEMENT: neutral $\times$ analytical. The steered response goes beyond correctness to operational specificity, reflecting the high w_d and low w_e of an analytically-framed, low-distress prompt. This also demonstrates that the pipeline does not impose emotional overlay when the user state does not warrant it.

Prompt	*"i think i want to switch careers into UX design but i'm not sure if i should... like maybe i should just stay in my current job? or maybe do a bootcamp? idk i keep going back and forth and i can't make up my mind lol"*
State	Seed: calm $\times$ indecisive. $\bar{e}^{-} = 2.1$, $c = -1$, $s = 2.1$. Quadrant: CONFIDENCE_BOOST. $w_e = 0.30$, $w_p = 0.5$, $w_d = 0.25$.
Baseline	"Switching careers is a big decision! Consider your interests, skills, and financial situation. A UX bootcamp could be a good starting point. Research the job market in your area and talk to people in the field before deciding."
Steered	"It sounds like you already know you want this — the back-and-forth might be less about the decision itself and more about permission to make it. One thing that might help: rather than choosing between *stay* or *bootcamp*, start with something smaller. Take a free UX course for two weeks and see if it pulls you in. That gives you real data rather than another round of weighing options in your head. You don't have to decide everything today."

Table 10: CONFIDENCE_BOOST: calm $\times$ indecisive. Baseline adds more options to consider, compounding the indecision. Steered response reduces option space, names the real blocker (permission, not information), and offers a concrete low-commitment action to break the loop.