CONVODRIFT: A Multi-Turn Conversational Dataset for Modeling Stylistic Tone Evolution

Vihindi Kotalawala[♠], Pamoda Dilranga[♢], Gayani Thoradeniya[♣], Prasan Yapa[♡]
[♠]Informatics Institute of Technology, Sri Lanka
[♢]Independant Researcher, Sri Lanka
[♣]National Health and Medical Research Council, Australia
[♡]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg
vihindi.20220672@iit.ac.lk, pamoda9@gmail.com
gayani.thoradeniya@nhmrc.gov.au, prasan.yapa@uni.lu

Abstract

The evolution of linguistic style in conversations is an underexplored issue in NLP. Most style-control datasets focus on sentences or assume a static style throughout, missing the dynamic shifts that occur as user preferences change during interactions. We introduce CONVODRIFT, a dataset designed to model progressive stylistic conversational tone drift under fixed semantic intent. It is built on 15,727 shared multi-turn conversational structures for adaptation and persona-conditioned alignment methods. It consists of six prompt–response pairs per conversation, each with the annotation of style drift and style direction labels. These pairs cover a range of communication genres. We further derive a complementary pairwise dataset by pairing semantically equivalent but stylistically distinct responses and annotating persona-conditioned preferences using five distinct style communication personas, enabling the controlled study of personalisation and pluralistic alignment in language tone. In addition to dataset construction, we conduct a comprehensive evaluation involving human validation, LLM-as-judge assessment, and automatic lexical and semantic evaluations. Across seven Likert criteria annotated by three human annotators, the average *Krippendorff's* α is 0.88, and our lexical and semantic analyses show that drift events induce lexical changes while preserving semantic similarity.

Keywords: Pluralistic alignment, Preference learning, Conversational tone drift

1 Introduction

Conversational interactions involving language models are inherently dynamic: users often adjust linguistic style and tone across multiple turns as goals, preferences, and stylistic expectations evolve. However, existing Natural Language Processing (NLP) datasets and benchmarks provide limited support for modelling this kind of progressive, turn-by-turn stylistic change. In particular, most style-control and style-transfer benchmarks operate at the sentence level (Rao and Tetreault, 2018) and explicitly ignore conversational context, making them ill-suited for studying stylistic evolution over multi-turn dialogue. (Roy et al., 2023). This highlights the lack of datasets that capture how tone progressively drifts during a conversation while preserving the context and maintaining the original semantic intent.

We introduce CONVODRIFT[1], a flexible dataset built around a shared conversational benchmark. This corpus facilitates the study of progressive stylistic tone drift in multi-turn dialogue and the derivation of style and tone-relevant preference data suitable for personalised alignment via Reinforcement Learning From Human Feedback (RLHF), addressing existing limitations in the field. By explicitly separating semantic content from stylistic modification, CONVODRIFT enables controlled analysis of tone variation across turns without confounding content changes, a property that is difficult to achieve with existing dialogue corpora (Cheng et al., 2020). The same conversational trajectories can be reused for multiple purposes, including modelling stylistic dynamics or serving as preference supervision, and can be evaluated under different preference profiles through persona-conditioned tone drift. This design naturally supports pluralistic alignment, in which multiple stylistic preferences are treated as valid rather than collapsed into a single reward signal, addressing concerns highlighted in recent alignment research (Poddar et al., 2024). To further demonstrate the practical utility of the dataset, we conducted a downstream preference modelling task using Vari-

[1]Dataset is available at Github and HuggingFace.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 12–24
November 6-7, 2026 ©2026 Association for Computational Linguistics

ational Preference Learning (VPL) (Poddar et al., 2024). The results show that CONVODRIFT can support persona-conditioned preference learning and is suitable for training preference models that capture diverse stylistic preferences.

In summary, CONVODRIFT contributes: (1) Explicit supervision for progressive tone drift in multi-turn dialogue; (2) A clear separation between semantic content and stylistic modification; (3) Support for studying persona-conditioned and pluralistic stylistic preferences within a unified conversational setting.

2 Related Work

2.1 Stylistic Tone Variation and Drift in Text

Research on stylistic control in text has historically focused on style transfer, in which a single input sentence is rewritten to exhibit a target stylistic attribute while preserving its semantic content. Foundational datasets such as the Grammarly Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) enable supervised informal–formal rewriting and have become standard benchmarks for formality transfer. The existing research has explored stylistic attributes, including politeness (Danescu-Niculescu-Mizil et al., 2013; Niu and Bansal, 2018), sentiment (Fu et al., 2018) , professionalism (Ficler and Goldberg, 2017), and controlled generation (Hu et al., 2017) without parallel data (Shen et al., 2017). Despite covering a range of stylistic attributes, these approaches predominantly operate at the sentence level, framing stylistic control as a single-turn rewriting task applied to isolated utterances. In these formulations, stylistic transformation is treated as a context-independent operation, where each input is processed without access to conversational history or discourse structure (Ficler and Goldberg, 2017; Hu et al., 2017). As a result, such methods do not capture stylistic dynamics across a conversation, where tone may evolve gradually in response to prior turns or communicative goals.

More recent work has begun to incorporate dialogue context into stylistic modelling. Roy et al. (2023) propose a few-shot conversational style transfer framework that rewrites entire dialogues conditioned on exemplar conversations. Nevertheless, this line of work focuses on global style transformation rather than planned, turn-by-turn stylistic drift, leaving fine-grained conversational tone evolution underexplored.

Although conversational style transfer incorporates dialogue context, it still provides limited supervision for *when* stylistic changes occur and *how* tone evolves progressively across turns. Modelling such tone evolution also introduces a personalisation dimension, as different users may exhibit and prefer different stylistic behaviours throughout an interaction. This observation naturally connects stylistic drift modelling to research on personalised and persona-conditioned dialogue systems, which explicitly account for user-specific conversational behaviours.

This connection between stylistic variation and user-specific behaviour motivates consideration of persona-based dialogue research. While such research captures user identity and preferences at the level of content and conversational consistency it offers limited mechanisms for modelling how stylistic preferences (e.g., tone and writing style) are learned, compared, or optimised during training, despite their central role in personalisation (Zhang et al., 2024). Addressing this limitation requires preference-based learning frameworks that explicitly model human judgements, motivating the growing role of preference data and RLHF.

2.2 Personalised Dialogues and Alignment

Personalisation in dialogue systems has traditionally been approached through persona conditioning, where responses are generated based on user-specific attributes or profiles (Zhang et al., 2018). Li et al. (2016) demonstrated that conditioning on textual persona descriptions improves speaker consistency and engagement in open-domain dialogue. Subsequent research extended this paradigm by incorporating demographic and contextual attributes such as age, gender, and emotion into dialogue systems (Zhao et al., 2025b), influencing response content and style. While prior work shows that personas shape stylistic behaviour (Joshi et al., 2017; Zhao et al., 2025a; Lee et al., 2025), these approaches typically treat style as a static property of a user or conversation, rather than as a signal that evolves across interaction turns.

At the same time, surveys on personalised language models highlight tone and writing style as key personalisation factors, alongside relevance and accuracy (Zhang et al., 2024). Effective personalisation, therefore, requires models to adapt to users' stylistic preferences across different contexts, rather than producing an average response.

However, most existing preference datasets do not explicitly label style or capture how stylistic alignment changes over multi-turn conversations. This limitation motivates the development of conversational datasets that jointly model style, persona, and preference dynamics over time.

RLHF has become the dominant paradigm for aligning language models with human preferences, typically using pairwise comparisons aggregated into a single reward signal (Ouyang et al., 2022; Bai et al., 2022; Ziegler et al., 2019). While recent work explores personalisation and pluralistic alignment by modelling heterogeneous or user-conditioned preferences (Poddar et al., 2024; Park et al., 2024; Cheng et al., 2023), existing RLHF datasets primarily focus on global qualities such as helpfulness, safety, and correctness. Stylistic factors, particularly tone and writing style, are rarely annotated explicitly, and preference signals are generally treated as static rather than context-dependent.

Taken together, prior work on personalised dialogue and preference-based alignment provides mechanisms for modelling *who* the speaker is and *what* they prefer, but offers limited support for capturing *how* stylistic expression evolves over the course of a conversation. Most datasets either abstract away conversational context or collapse preferences into a single reward signal, limiting their ability to capture fine-grained, progressive stylistic changes. These limitations highlight shortcomings in existing resources for studying dynamic tone adaptation and stylistic alignment in multi-turn conversations.

3 Methodology

ConvoDrift is a multi-stage data generation and validation dataset designed to support both the analysis of multi-turn stylistic tone drift and persona-conditioned preference learning for personalised RLHF. Accordingly, the dataset consists of two core stages, each producing a distinct but tightly coupled dataset. (1) ConvoDrift–Conversation: A multi-turn conversational dataset that captures progressive stylistic tone drift under fixed semantic intent. (2) ConvoDrift–RLHF: A persona-conditioned pairwise preference dataset constructed from conversational trajectories to enable personalised and pluralistic alignment.

As shown in Figure 1, both datasets reuse the same multi-turn prompts and responses. The RLHF subset restructures responses from these conversations into style-contrasting preference pairs. This unified design ensures consistency between conversational structure and preference supervision, enabling ConvoDrift to serve as both a diagnostic benchmark for tone drift and a resource for studying personalised and pluralistic alignment (Poddar et al., 2024). To provide empirical evidence of the dataset's effectiveness, we conducted a downstream preference modelling task using VPL with representations of GPT-2 and Llama 3.1 8B, achieving accuracies of 0.579 and 0.65, respectively, with further details provided in Appendix A.

3.1 Conversational Dataset Creation

This section describes how ConvoDrift was constructed to capture progressive conversational tone changes while preserving semantic intent across multiple turns.

3.1.1 Model Selection and Generation Setup

To select appropriate models for data generation, we conducted a preliminary comparison across 12 Large Language Models (LLMs) using an identical set of conversational prompts. Models were evaluated based on output quality, adherence to stylistic instructions, latency, and estimated token-level cost under a fixed input–output budget. Based on this comparison, we selected `GPT-4.1-mini` as the primary model for conversational response generation due to its consistent output quality and reliable stylistic control, and `GPT-5-nano` for generating explicit drift, direction labels due to its higher quality and lower latency outputs. Using this setup, we generated 15,727 multi-turn conversations, each containing six prompt–response pairs, resulting in a total of 94,362 response turns. This has been discussed in detail in Appendix B.

3.1.2 Prompt Design and Quality Control

ConvoDrift generates conversations across multiple genres as shown in Table 1 with distinct stylistic norms, spanning both professional and informal contexts. Each genre is paired with diverse communicative scenarios, and prompts enforce semantic preservation while allowing controlled stylistic variation. Generated conversations are then filtered using a rule-based quality pipeline to ensure structural consistency, including a fixed conversation length, valid drift annotations, and the presence of multiple stylistic drift events.

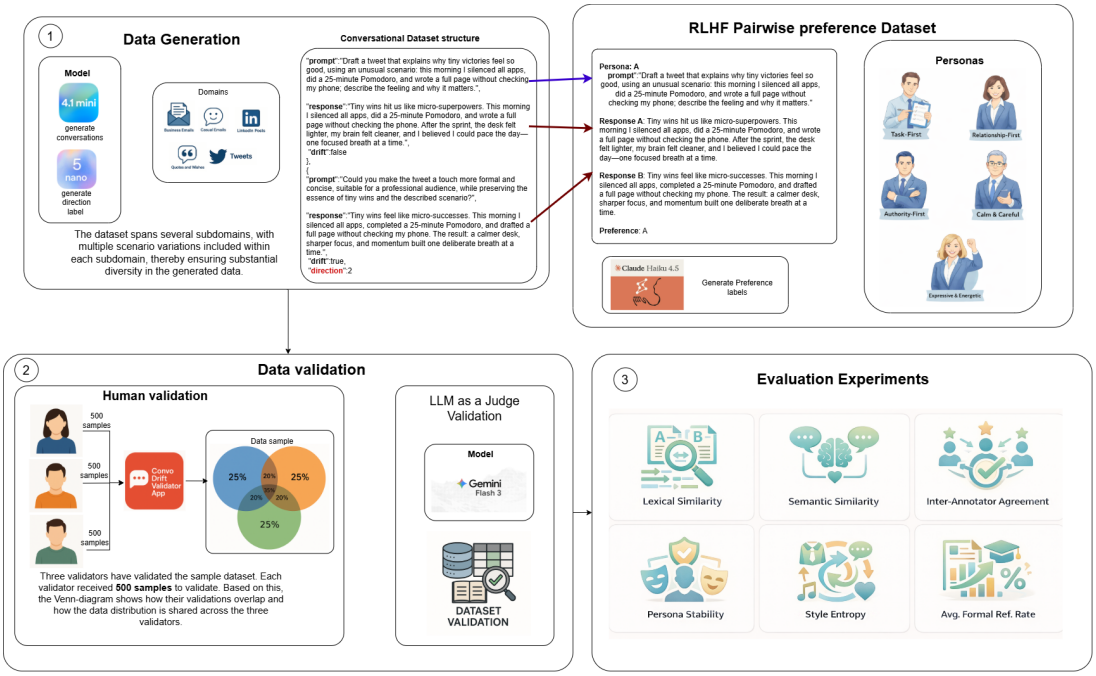

Figure 1: Overview of the ConvoDrift data construction and evaluation pipeline. [1] **Data Generation** illustrates the creation of structured multi-turn conversations across diverse domains and scenarios using selected language models, followed by the derivation of stylistically contrasting responses for RLHF pairwise preference data with persona-conditioned profiling. [2] **Data Validation** shows the quality assurance process, combining human validation via a custom annotation interface and LLM-as-a-judge validation. [3] **Evaluation Experiments** summarises the downstream evaluation protocols applied to the resulting datasets.

Communication genre	Count
Business email	2,997
Casual email	3,246
Quotes/Wishes	3,340
LinkedIn post	3,150
Tweet	2,994

Table 1: Conversation counts by communication genre in the CONVODRIFT dataset.

3.1.3 Multi-turn Conversational Structure

Each conversation is structured as a sequence of six prompt–response pairs:

$$C = \{(p_t, r_t)\}_{t=1}^{6}. \quad (1)$$

where p_t denotes the user prompt at turn t and r_t the corresponding response generated conditionally on the preceding conversational context. Figure 2 provides an illustrative example of this six-turn structure. The first prompt reflects the user's original request, denoted as T1 is always annotated as `Drift=False`. Subsequent turns refine the previous response based on the user's prompt; we set `Drift=True` only when the prompt explicitly requests a style/tone change.

Importantly, semantic content is preserved across turns, enabling the dataset to capture drift trajectories rather than isolated stylistic rewrites (Roy et al., 2023).

3.1.4 Drift, Direction Labels, and Metadata

Each prompt–response pair is annotated with two complementary labels, which are shown in the example conversation in Figure 2.

- A *binary drift label* (`Drift`), indicating whether the user prompt requests a stylistic/tone change; and
- A *direction label* (`Direction`), indicating the direction of the stylistic shift relative to the previous response.

The direction label is defined as:

$$\texttt{direction} = \begin{cases} 0, & \textit{no stylistic change} \\ 1, & \textit{casual / expressive tone} \\ 2, & \textit{formal / precise tone} \end{cases}$$

The drift label is primarily driven by the user prompt, whereas direction is inferred from the stylistic properties of the generated response. For example, in Figure 2, T4 is marked `Drift=True` because the user asks for a "more casual" rewrite, and the corresponding response is labelled with `Direction=1` to reflect a shift toward a more casual/expressive style. These two labels are evaluated independently to avoid conflating intended stylistic instructions with their realised linguistic

Example conversation (Tweet domain; scenario: share a thought about learning through challenges).

T1 (Drift=False)
User: I want to share a tweet about how recovery should be viewed as an essential, ongoing part of resilience, not just an end goal.
Assistant: Recovery isn't the finish line—it's a crucial part of the journey. Every step of healing, no matter how small, builds resilience. Embrace the process and honor your progress. #Resilience #Healing

T2 (Drift=True; Direction=2)
User: Rewrite it in a more formal and precise tone, emphasizing that recovery is a gradual process.
Assistant: Recovery is a gradual process rather than a fixed endpoint. Each stage contributes to resilience and supports sustained personal growth.

T3 (Drift=False)
User: Make it shorter and more direct for a busy feed.
Assistant: Recovery isn't the finish line—it's every brave step forward. Healing builds resilience, one moment at a time. #Resilience #Healing

T4 (Drift=True; Direction=1)
User: Rewrite with a more casual, approachable tone.
Assistant: Hey, recovery's not just about getting to the end—it's every little step you take to heal. Even when it's slow, you're building strength. Keep going. #Resilience

T5 (Drift=True; Direction=0)
User: Include a nature metaphor to symbolise recovery and resilience.
Assistant: Recovery is like a tree growing through the seasons—sometimes slow, sometimes tough, but always reaching for the light. Keep reaching. #Resilience #NatureHeals

T6 (Drift=False)
User: Keep the tree-growth metaphor and condense the message into two clear sentences.
Assistant: Recovery parallels a tree's growth through seasons—steady and resilient. Each phase of healing strengthens the foundation for future growth.

Figure 2: Example CONVODRIFT conversation illustrating progressive tone edits under fixed semantic intent.

effects. Main statistics about the dataset has been discussed in Appendix D.

3.2 Conversion to RLHF Pairwise Data

This section explains how conversational trajectories are transformed into semantically aligned but stylistically contrasting response pairs for persona-conditioned preference learning.

3.2.1 Pairwise Preference Construction

Each multi-turn conversation in CONVODRIFT consists of a sequence of prompts that progressively request stylistic refinements to a response, producing a trajectory of responses with shared semantic intent but differing stylistic realisations.

To derive preference supervision suitable for RLHF, we construct pairwise preference instances by selecting a prompt p from the conversation and pairing two responses associated with it. Specifically, the initial response generated for the task is treated as one candidate (y_a), while a later response from the same conversation corresponding to an explicit stylistic drift with a known direction (e.g., casual or formal) is selected as the contrasting candidate (y_b). The process of constructing the RLHF pairwise dataset section is shown in Figure 1.

Response pairs may be drawn from adjacent or non-adjacent turns, depending on the desired degree of stylistic contrast. By construction, both responses are semantically equivalent, and preference distinctions arise solely from stylistic variation rather than content differences. Each preference instance is represented as a tuple (p, y_a, y_b), following the standard pairwise preference-learning formulation used in RLHF (Christiano et al., 2017; Ouyang et al., 2022).

Persona	Description
Task-First	Focuses on efficiency, clarity, and structured reasoning, prioritising actionable outcomes with minimal unnecessary elaboration.
Relationship-First	Emphasises warmth, empathy, and cooperative language to support positive and sustainable interpersonal interactions.
Authority-First	Communicates with confidence and decisiveness, favouring clear direction, assertive tone, and leadership-oriented language.
Calm & Careful	Uses measured, precise language to minimise risk, ensure safety, and promote stable and cautious decision-making.
Expressive & Energetic	Engages with enthusiasm and vivid language, favouring expressive phrasing that creates excitement and momentum.

Table 2: Persona profiles used for persona-conditioned preference labelling. Each persona encodes a distinct communication focus that guides stylistic preference judgements.

3.2.2 Persona-based Preference Learning

CONVODRIFT supports persona-conditioned preference learning by leveraging the explicit stylistic annotations available in the underlying conversational dataset. Since each response is labelled with drift and direction tags (e.g., toward casual/expressive or formal/precise), stylistic preferences can be systematically grouped and evaluated without relying on implicit or noisy preference inference.

As an experimental instantiation, we define five communication personas, each encoding a consistent stylistic bias shown in Table 2. These personas are not intended to model user identity, but rather to represent distinct and interpretable evaluation perspectives over stylistic variation.

For each example, which includes a prompt and two responses, each persona independently chooses the response that best matches its communication style. This produces separate preference labels for each persona. By keeping these persona-specific labels instead of combining them into a single decision, the dataset preserves differences in stylistic preferences. This design allows us to study personalised and pluralistic alignment, in which multiple communication styles are treated as valid at the same time (Cheng et al., 2023; Poddar et al., 2024).

4 Evaluation and Experiments

4.1 Conversational Dataset

This section evaluates the reliability and quality of the conversational dataset through human validation, LLM-as-a-judge assessment, and automatic semantic and lexical analyses.

4.1.1 Human Evaluation

We conducted a human evaluation to assess the correctness of tone drift annotations and the overall quality of multi-turn conversations in CONVO-DRIFT. Given the cost and time of manual validation, three independent annotators evaluated a stratified subset of the dataset using a sampling strategy designed to balance coverage with reliable agreement estimation.

Since human validation is particularly valuable but limited, we used a controlled-overlap design to maximise the number of human-validated conversations while still enabling agreement analysis: 175 conversations were annotated by all annotators (gold set), 300 conversations were annotated by two annotators (100 conversations per annotator pair), and 375 conversations were annotated by a single annotator (125 unique conversations per annotator). This results in 1,500 conversation-level ratings in total (500 per annotator), covering 850 unique conversations. The communication genre balance was maintained across all subsets, ensuring each genre was represented in similar proportions. Inter-annotator agreement metrics were computed exclusively on the gold set, while the remaining annotated data was used for qualitative analysis and downstream evaluation.

Annotators evaluated each conversation on seven ordinal criteria using a five-point Likert scale. We measure agreement on these ordinal ratings using Krippendorff's α for ordinal data, which supports multiple annotators (Krippendorff, 2011), and report the resulting scores in Table 3.

In addition to the scalar ratings, annotators were instructed to correct drift flags and tone direction labels whenever inconsistencies were identified. Drift is annotated as a binary value indicating whether a tone change is requested in the prompt, while direction labels specify the direction of tone change (no change, formal to casual, or casual to formal). For validating these drift and direction labels on the common subset, inter-annotator agreement was Drift α: *0.7628* and Direction α: *0.7772*. Overall, the agreement results indicate substantial reliability for studying conversational stylistic drift and personalised alignment.

Criterion	α	95% CI
Tone consistency and drift annotation accuracy	0.763	[0.732, 0.794]
Tone direction label accuracy	0.777	[0.751, 0.803]
Drift distribution realism	0.891	[0.853, 0.926]
Fluency and clarity	0.946	[0.916, 0.973]
Cross-turn coherence	0.964	[0.939, 0.985]
Prompt-following accuracy	0.900	[0.863, 0.936]
Domain-specific naturalness	0.916	[0.883, 0.947]

Table 3: Inter-annotator agreement on the common subset measured using Krippendorff's alpha for ordinal data. Confidence intervals correspond to 95% bootstrap intervals.

4.1.2 LLM-as-a-Judge Evaluation

In addition to human validation, we employed an LLM-as-a-judge framework (Zheng et al., 2023). After human evaluation, the remaining portion of the dataset was evaluated using `Gemini 3 Flash` as an automatic judge. The LLM was prompted to assign drift labels, direction labels, and conversation-level quality ratings using the same annotation guidelines followed by human annotators. To assess the reliability of the LLM judge, its outputs were compared against a gold-standard human reference constructed via majority voting over human annotations. Agreement for the drift and direction labels was measured using Cohen's κ. The LLM demonstrates strong agreement with human validation for tone drift detection, achieving Drift of *0.888*, and substantial agreement for tone direction labelling with Direction of *0.930* (Cohen, 1960). These results indicate that the LLM reliably identifies when tone changes occur and correctly captures their directional shift, while aligning with human annotator decisions.

4.1.3 Automatic Evaluation

Semantic Similarity: We evaluate semantic preservation across refinement turns using sentence-level semantic similarity computed with `SBERT` embeddings (Reimers and Gurevych, 2019). We report three complementary measures: (i) *adjacent-turn similarity*, which compares consecutive turns and captures local semantic continuity; (ii) *anchor similarity*, which compares each turn against the initial response and captures global meaning preservation; and (iii) *end-to-start similarity*, which compares the final turn of a conversation with its initial response to measure start-to-end semantic retention.

Similarity Type	n	Mean	Median	Min
Adjacent (SBERT)	2500	0.8993	0.9207	0.3287
Anchor (SBERT)	3000	0.8903	0.9034	0.2382
End-to-Start (SBERT)	492	0.8519	0.8715	0.4423

Table 4: Overall turn-level semantic similarity across refinement turns.

As shown in Table 4, both adjacent-turn and anchor similarity remain high. Adjacent-turn similarity has a mean of 0.899, and anchor similarity has a mean of 0.890, indicating that responses preserve semantic intent both locally and globally as stylistic refinements progress. Additionally, end-to-start similarity is also high, with a mean of 0.852, indicating that meaning is largely retained from the first to the final turn.

Unlike lexical similarity, we do not stratify semantic similarity by drift labels: drift in CONVO-DRIFT is intended to change style, but regardless of whether a prompt requests drift or not, the underlying meaning of responses should be preserved throughout the conversation. Therefore, semantic similarity is evaluated using these three turn-level views of meaning preservation.

We further analyse semantic preservation across communication genres, as depicted in Table 5, which presents domain-wise end-to-start similarity statistics.

Lexical Similarity: While semantic similarity remains high across turns, lexical overlap can change substantially when tone edits are introduced. We evaluate lexical variation across consecutive turns to quantify these surface-level changes. Lexical similarity is computed using ROUGE-L (longest common subsequence overlap), Jaccard-1 (unigram overlap), and Jaccard-2 (bigram overlap).

Genre-wise lexical similarity within non-drift regions (Within; Drift=False) and at drift boundaries (Boundary; Drift=True) in turns within the conversations is summarised in Table 6. Across all metrics and genres, within transitions consistently exhibit higher overlap than boundary transitions. Overall, the drops in ROUGE-L, Jaccard-1, and Jaccard-2 at boundary transitions indicate that tone drift is accompanied by meaningful lexical rephrasing, although the underlying semantic intent is preserved.

4.2 Persona-conditioned RLHF Dataset

This section evaluates the quality of the pairwise preference labels and their ability to capture consistent yet distinct persona-conditioned communication preferences.

Genre	#Convs	Mean	Median	Min
Business email	93	0.9123	0.9255	0.7594
Casual email	100	0.8551	0.8679	0.6223
LinkedIn post	99	0.8842	0.8919	0.7380
Quotes/Wishes	100	0.7739	0.7971	0.4423
Tweet	100	0.8385	0.8580	0.5517

Table 5: Domain-wise end-to-start semantic similarity measured using SBERT across different communication genres.

4.2.1 Experimental study

We evaluate candidate LLM judges using stability, persona sensitivity, style entropy, and formal preference rate to select a reliable model for persona-conditioned labelling. During the construction of this pairwise dataset, we removed data points with semantic similarity below 0.6 to keep the dataset semantically similar but stylistically distinct.

Stability Consistency: Stability measures how consistently an LLM returns the same preference label when asked multiple times under identical conditions. For each sample and persona, we query the model five times and report stability as the fraction of runs that match the majority label (e.g., 5/5 indicates perfect stability). High stability indicates that the model's outputs are not dominated by randomness.

Persona Sensitivity: Persona sensitivity measures how strongly an LLM differentiates between personas by quantifying how often personas disagree on the preferred response (formal vs. casual) for the same sample. For each sample annotated by K personas, we compute sensitivity using the *average pairwise disagreement rate* across persona pairs:

$$\text{Sensitivity} = \frac{1}{\binom{K}{2}} \sum_{i<j} \mathbf{I}(s_i \neq s_j), \tag{2}$$

where s_i and s_j denote the preference labels inferred for personas i and j. We report sensitivity by averaging this disagreement score across samples. Low values indicate that most personas agree on the same response, while high values indicate substantial disagreement across personas.

Style Entropy: Style entropy measures how consistent or flexible a persona's preferences are across samples, based on how balanced its selections are between the two styles (formal vs. casual). For each persona, we compute *Shannon entropy* (Shannon, 1948) (in bits) based on how often it chooses

Genre	Type	n	ROUGE-L	Jaccard-1	Jaccard-2
Business email	Within (Drift=False)	6,242	0.606	0.543	0.374
	Boundary (Drift=True)	8,742	0.538	0.469	0.281
Casual email	Within (Drift=False)	7,031	0.684	0.639	0.501
	Boundary (Drift=True)	9,199	0.492	0.457	0.267
LinkedIn post	Within (Drift=False)	6,823	0.686	0.624	0.508
	Boundary (Drift=True)	8,927	0.393	0.342	0.162
Quotes/Wishes	Within (Drift=False)	5,757	0.492	0.422	0.255
	Boundary (Drift=True)	10,943	0.340	0.282	0.109
Tweet	Within (Drift=False)	5,886	0.572	0.487	0.301
	Boundary (Drift=True)	9,084	0.386	0.314	0.128
Overall	Within (Drift=False)	31,739	0.614	0.550	0.396
	Boundary (Drift=True)	46,895	0.426	0.369	0.186

Table 6: Genre-wise lexical similarity between consecutive turns in conversations within non-drift regions (Within; Drift=False) and at drift boundaries (Boundary; Drift=True), reported with `ROUGE-L` and unigram/bigram `Jaccard` (n transitions).

formal vs. casual responses:

$$\text{style entropy} = -\sum_{s\in\{\text{formal,casual}\}} p(s)\log_2 p(s), \quad (3)$$

where $p(s)$ denotes the fraction of times the persona selects style s across the evaluation set. Low entropy indicates that a persona almost always prefers the same style, reflecting highly consistent preferences, while high entropy indicates that a persona selects both styles at more similar rates, reflecting more context-dependent preferences.

Average Formal Preference Rate: The average formal preference rate summarises the model's overall tendency to prefer formal responses across all comparisons, independent of persona. Low values indicate a global bias toward casual selections, while high values indicate a global bias toward formal selections. A value close to 0.5 indicates that, overall, the model selects formal and casual at similar rates.

Considering the results in Table 7, `Claude Haiku 4.5` provides the most suitable balance among the evaluated criteria. It combines very high stability with the strongest persona sensitivity and relatively low global stylistic bias, while maintaining sufficient contextual flexibility.

Based on this analysis, `Claude Haiku 4.5` is selected as the LLM for generating persona-conditioned preference labels in the final dataset.

4.2.2 Relation to Agreement Analysis

Inter-persona agreement was assessed using Cohen's κ to examine whether the personas make similar or contrasting stylistic preference judgments. The plot shows two broad preference patterns. Personas A, C, and D each have two scores above the $\kappa = 0$ line, indicating positive agreement with one another, while Personas B and E have one positive score representing their agreement with each other. The remaining scores fall below zero, showing disagreement between these two groups. Although some personas share similar preferences, differences in the distribution and magnitude of their agreement scores indicate that each persona retains distinct evaluative characteristics. Further details, including the exact pairwise agreement values, are provided in Appendix C.

Model	Stability	Sensitivity	Average	
			Entropy	Formal Rate
Claude Haiku 4.5	0.990	0.533	0.853	0.542
Gemini 2.5 flash	0.942	0.504	0.868	0.581
GPT-5 Mini	0.932	0.483	0.886	0.585
GPT-5	0.956	0.442	0.884	0.604

Table 7: Comparison of candidate LLMs for persona-conditioned preference labelling.

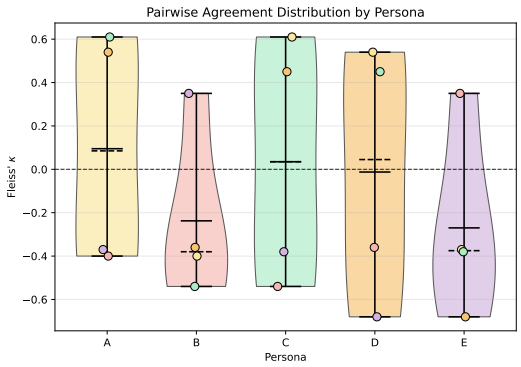

Figure 3: Distribution of pairwise agreement scores for each persona using Cohen's κ.

5 Conclusion

In this paper, we introduced CONVODRIFT, a benchmark for modelling progressive stylistic tone drift in multi-turn conversations. By utilising a shared conversational structure, we explicitly sep-

arate semantic intent from stylistic variation, enabling the derivation of both multi-turn drift trajectories and persona-conditioned preference datasets for pluralistic alignment. Comprehensive validation via human assessment, LLM-as-a-judge, and lexical metrics confirms that this preserves semantic meaning while capturing structured stylistic dynamics.

6 Limitations and Future Directions

There are several avenues remain for future exploration to address the constraints of our current approach. First, as a synthetically generated dataset designed to emulate human–LLM conversations, it lacks the variability and spontaneity inherent in real-world organic data; thus, future work may include real human conversations to capture this realism. Second, we currently measure tone variations only between "formal" and "casual" styles, so future research can explore more complex emotions and subtle style shifts. Third, all conversations are constrained to a fixed length of six turns, limiting the analysis of longer-range stylistic dependencies and naturally varying dialogue lengths. Fourth, as the dataset is predominantly synthetically generated and LLM-annotated, it may inherit model-specific stylistic patterns and annotation biases that do not fully reflect naturally occurring human–LLM interactions. Finally, persona conditioning is restricted to five predefined personas; expanding this space to include a broader range of user preferences would further strengthen the dataset's relevance for personalisation and pluralistic alignment research.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Jihyun Lee, Yejin Jeon, Seungyeon Seo, and Gary Lee. 2025. PicPersona-TOD : A dataset for personalizing utterance style in task-oriented dialogue with image persona. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 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*, 35:27730–27744.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. 2024. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shamik Roy, Raphael Shu, Nikolaos Pappas, Elman Mansimov, Yi Zhang, Saab Mansour, and Dan Roth. 2023. Conversation style transfer using few-shot learning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025a. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. *arXiv preprint arXiv:2505.15456*.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025b. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Downstream Validation with VPL

To examine the effectiveness of the proposed dataset beyond direct label annotation, we conducted a downstream validation experiment using Variational Preference Learning (VPL) (Poddar et al., 2024). The purpose of this task was to show the effectiveness of the proposed drift and direction labels in providing useful learning signals for preference modelling, and to assess whether the dataset captures meaningful persona-specific preference patterns. VPL is well-suited for this analysis because it represents preference learning as a latent-variable problem, where the latent variable z captures hidden user or annotator context and enables the reward model to predict preferences conditioned on that context, rather than averaging all preferences into a single reward function.

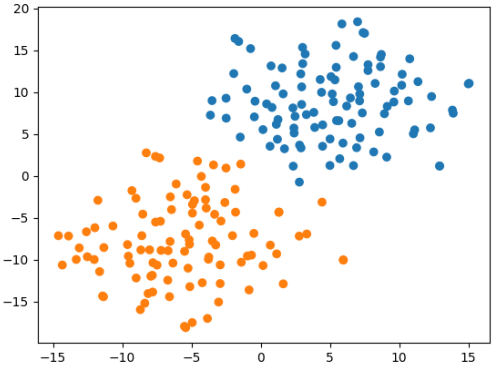

Figure 4: t-SNE visualisation of the latent representations learned by the Llama 3.1 8B-based VPL model.

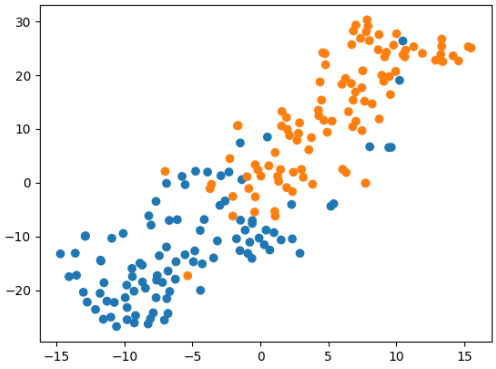

Figure 5: t-SNE visualisation of the latent representations learned by the GPT-2-based VPL model.

Following this setup, we used the prompt together with response A and response B from our pairwise dataset, treating Persona A and Persona B as two distinct preference contexts. The model was trained to predict the preferred response under the corresponding persona-conditioned setting. We evaluated the VPL-based model using two representation backbones: GPT-2 and Llama 3.1 8B. The GPT-2-based model achieved an accuracy of 0.579, while the model using Llama 3.1 8B representations achieved a higher accuracy of 0.65. This indicates that the proposed labels provide predictive signal for distinguishing persona-dependent preferences. The stronger performance of Llama 3.1 8B is likely due to its larger parameter capacity and richer representation quality, which allow it to capture more fine-grained stylistic and persona-conditioned differences.

We further visualised the learned latent representations using t-SNE plots. The resulting visualisations show clearer clustering for the Llama 3.1 8B-based model, with two distinct clusters corresponding to Persona A and Persona B. This separation suggests that the model learned structured differences between the two persona-conditioned preference distributions. Therefore, the downstream experiment provides additional evidence that the proposed drift and direction labels are meaningful, and that the dataset supports persona-sensitive preference learning. Since VPL is designed to learn separable latent representations for divergent user preferences, these results strengthen the validity of our dataset for pluralistic and adaptive preference modelling.

B Prompts used for dataset generation

The dataset was generated using a structured prompting strategy designed to produce multi-turn conversations exhibiting controlled stylistic tone drift while preserving semantic intent. For each communication domain, a dedicated prompt template was used to guide the generation process. All prompts enforce a fixed conversation length of six prompt–response pairs and require the assistant to rewrite the full response at every turn based on incremental user feedback.

Across all domains, the first prompt in each conversation requests an initial draft and is explicitly constrained to have no tone drift. Subsequent prompts introduce natural, user-like revision requests. Tone drift is annotated at the transition level and is allowed only when a noticeable change in tone occurs, such as formal to casual or vice versa. To ensure realism, each conversation contains a minimum of two and a maximum of three tone-drift transitions distributed organically across turns two to six.

All prompts enforce strict output constraints, including a fixed JSON schema, full response rewrites at every turn, and prohibition of placeholders. This design ensures consistency across domains while allowing stylistically diverse refinements. The following sections provide the exact prompt templates used for each communication genre.

Business Email Prompt

```
You are an Email Writing Assistant. Generate only 1 conversation about a BUSINESS EMAIL in the
sub-domain: "{domain}".

The FIRST prompt describes {scenario}. Think out of the box and think about a scenario related
to this and write prompt.

- When asking the first prompt, phrase it naturally as if a human user is requesting help from
an AI assistant to draft the email by explaining the context and purpose of the email and any
key details to include.
- Avoid using placeholders like [company], [lastname], [date] — always substitute with natural
details.
- The conversation always should be as a conversation between a human user and an AI assistant.

STRICT REQUIREMENTS:
- Output exactly 6 prompt-response pairs.
- Pair 1 must have drift=false.
- Pairs 2-6 must contain 2-3 drift=true labels, distributed organically.
- Responses must rewrite the full email each time.
- Output a single JSON object following the specified schema.
- Output a single JSON object following the specified schema.
```

Figure 6: Business Email prompt template used for dataset generation.

Casual Email Prompt

```
You are a Casual Email Writing Assistant. Generate only 1 conversation about a CASUAL EMAIL in the
sub-domain: "{domain}".

The FIRST prompt describes {scenario} and should sound like a natural human request.

- Keep the style personal, warm, and informal.
- Avoid placeholders and use realistic details.

STRICT REQUIREMENTS:
- Output exactly 6 prompt-response pairs.
- Pair 1 must have drift=false.
- Pairs 2-6 must contain 2-3 drift=true labels.
- Rewrite the full email at each turn.
- Output a single JSON object in the required format.
```

Figure 7: Casual Email prompt template used for dataset generation.

LinkedIn Post Prompt

```
You are a LinkedIn Post Writing Assistant. Generate only 1 conversation about a LINKEDIN POST in
the sub-domain: "{domain}".

The FIRST prompt explains the scenario as a natural user request.
- The first response must be a complete LinkedIn post.
- Avoid placeholders and use realistic details.

STRICT REQUIREMENTS:
- Output exactly 6 prompt-response pairs.
- Pair 1 must have drift=false.
- Pairs 2-6 must contain 2-3 drift=true labels.
- Rewrite the full post at each turn.
- Output a single JSON object in the required format.
```

Figure 8: LinkedIn post prompt template used for dataset generation.

Quotes & Wishes Prompt

```
You are a Quotes & Wishes Writing Assistant. Generate only 1 conversation about a QUOTE
or WISH in the sub-domain: "{domain}".

The FIRST prompt describes the scenario naturally.
- The first response must be a complete quote or wish.
- Avoid placeholders.

STRICT REQUIREMENTS:
- Output exactly 6 prompt-response pairs.
- Pair 1 must have drift=false.
- Pairs 2-6 must contain 2-3 drift=true labels.
- Rewrite the full quote or wish at each turn.
- Output a single JSON object in the required format.
```

Figure 9: Quotes and wishes prompt template used for dataset generation.

Tweet Prompt

```
You are a Tweet Writing Assistant. Generate only 1 conversation about a TWEET in
the sub-domain: "{domain}".

The FIRST prompt explains the scenario naturally.
- The first response must be a complete tweet (max 280 characters).
- Avoid placeholders.

STRICT REQUIREMENTS:
- Output exactly 6 prompt-response pairs.
- Pair 1 must have drift=false.
- Pairs 2-6 must contain 2-3 drift=true labels.
- Rewrite the full tweet at each turn (max 280 characters).
- Output a single JSON object in the required format.
```

Figure 10: Tweet post prompt template used for dataset generation.

LLM as a Judge Prompt

```
SECTION 1: INSTRUCTIONS & SCORING GUIDELINES

You are an expert conversation evaluator. Your task is to evaluate
the provided conversation based on specific criteria (Q1-Q8).

Scoring guideline for Q1: Tone Consistency & Drift Annotation Accuracy
- Drift Definition: Detects if the PROMPT explicitly or implicitly
(roundabout way) requests a shift in tone/style (specifically casual
to formal OR formal to casual).
  - IF prompt requests tone shift -> drift=true.
  - IF prompt requests ONLY content edits, length changes, or
    other non-tone updates -> drift=false.
- Assume there are n pairs; pair 1 must be drift=false.
- Start from 5.
- For each pair with an incorrect drift label subtract 1 point,
  down to a minimum of 1.

Q2: Direction Label Correctness
- Detect direction by comparing CURRENT response vs PREVIOUS response.
  - Casual shift -> 1
  - Formal shift -> 2
  - No change -> 0

Q3: Realistic Drift Distribution
Q4: Overall Conversation Quality
Q5: Fluency & Clarity
Q6: Coherence
Q7: Task Relevance
Q8: Domain Naturalness

KEY RULES:
- Leave feedback empty if correct
- First pair always drift=false
- Consider implicit tone shifts as drift
"""

SYSTEM_PROMPT = f"""
{SCORING_CRITERIA_TEXT}

OUTPUT FORMAT:
Return ONLY JSON:
{
  "ratings": {...},
  "refined_drift_label": [...],
  "refined_direction_label": [...],
  "feedbacks": {...}
}
```

Figure 11: Prompt template used for LLM as a Judge.

C Pairwise Inter-Persona Agreement

Table 8 presents the complete pairwise κ agreement scores underlying the persona-level distributions shown in Figure 3. Positive values indicate similar stylistic preferences between persona pairs, whereas negative values indicate consistent preferences for opposing stylistic choices.

	A	B	C	D	E
A	1.00	-0.26	0.49	0.51	-0.29
B	-0.26	1.00	-0.36	-0.29	0.38
C	0.49	-0.36	1.00	0.34	-0.24
D	0.51	-0.29	0.34	1.00	-0.66
E	-0.29	0.38	-0.24	-0.66	1.00

Table 8: Pairwise Cohen's κ agreement between personas.

D Dataset Distribution Statistics

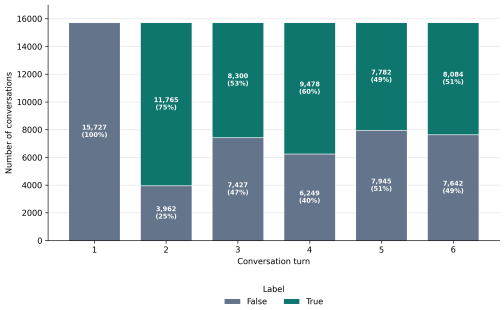

Figure 12: Distribution of drift labels across conversation turns.

Figure 12 shows how drift changes throughout the conversations. All first turns are labelled as non-drift because they establish the initial conversational context. Drift is most common in the second turn, occurring in 74.8% of conversations. The proportion decreases to 52.8% in turn three, increases to 60.3% in turn four, and remains approximately balanced between drift and non-drift in turns five and six.

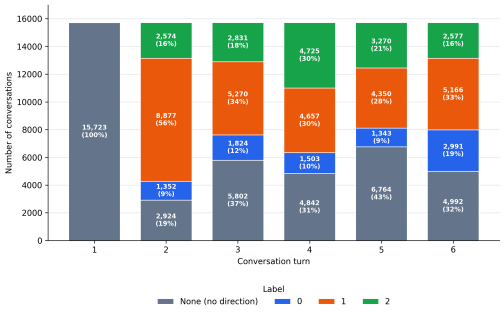

Figure 13: Distribution of direction labels across conversation turns.

Figure 13 presents the distribution of direction labels. The first turn predominantly has no direction because it does not contain drift. Direction 1 is most frequent in turn two, accounting for 56.4% of conversations. The direction distribution becomes more balanced in later turns, particularly in turn four, where Directions 1 and 2 each represent approximately 30% of conversations.

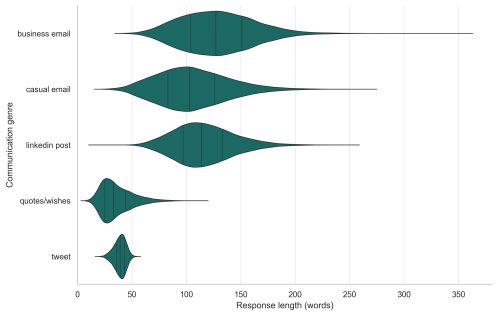

Figure 14: Distribution of response length by communication genre.

Figure 14 shows response length, measured in words, for each communication genre. Business emails have the longest responses, with a median of 127 words, followed by LinkedIn posts with 114 words and casual emails with 103 words. Quotes and wishes and tweets are substantially shorter, with median lengths of 33 and 39 words, respectively. These differences are consistent with the expected structural and stylistic requirements of each communication genre.