CausalPersona: A Benchmark and Protocol for Population-Level Realism in Controlled Conversation Generation

Anonymous ACL submission

Abstract

Recent progress in persona-grounded dialogue generation has improved local coherence and persona-style diversity in single-conversation settings (Zhang et al., 2018; Kim et al., 2023). However, such systems rarely account for *population-level realism*—whether generated dialogues match the empirical distributions of how users interact at scale (e.g., intent dynamics, politeness, persistence, and error recovery). As a result, evaluation often emphasizes fluency or surface coherence, leaving distributional fidelity underexplored.

We present **CausalPersona**, a benchmark-driven framework for population-level controlled conversation generation. CausalPersona distills a compact set of **effective and interpretable persona groups** from real traffic via high-dimensional behavioral labeling and *causality-inspired* refinement, which removes redundant descriptors while retaining key behavioral drivers. We formalize **distribution alignment**: given partial user histories and group conditions, synthesize conversations whose *aggregate structural and behavioral feature distributions* match those of real conversations. Our protocol evaluates alignment over **15 dimensions** (3 structural + 12 behavioral) using Jensen–Shannon divergence and cosine/Euclidean distances, complementing fluency-focused metrics.

Across five datasets, CausalPersona achieves strong **Structural Align.** (avg. **0.0683**) and the best average **Behavioral Align.** (**0.1892**, a **39.2%** improvement over the strongest baseline), demonstrating clear behavioral separability while remaining structurally faithful.[1]

1 Introduction

Persona-grounded dialogue generation has advanced conversational personalization. By constructing interpretable persona representations, prior work improves *local coherence* (persona–content consistency) and *style consistency* in tone, self-reference, and opinion style (Li et al., 2016; Mazaré et al., 2018; Zhang et al., 2018; Madotto et al., 2019; Wolf et al., 2019; Cheng et al., 2024; Jandaghi et al., 2024). Recent benchmarks (e.g., PersonaLens (Zhao et al., 2025)) further evaluate personalization with simulated users and LLM judges.

Despite these advances, a gap remains between typical research settings and real-world traffic. Most methods focus on *individual-level* personas and a small number of dialogues per persona, while overlooking *population-level behavioral distributions* that emerge at scale. In practice, generations can be fluent yet systematically misrepresent how users interact: how intents evolve, how sessions are managed (persistence vs. early exit), and how failures are handled (clarification and repair) (Bohus and Rudnicky, 2005). This mismatch undermines the realism of synthetic data for training and evaluation, especially when large-scale persona-annotated traffic is costly or sensitive.

Moreover, evaluation is often narrow. Standard metrics emphasize fluency and coherence, but seldom test *behavior distribution fidelity* across structural and behavioral dimensions, leaving key properties such as politeness, persistence, and error recovery under-measured.

To bridge these gaps, we propose **CausalPersona**, a controllable conversation synthesis framework and corresponding *benchmark* for population-aligned dialogue generation. CausalPersona distills interpretable persona groups from real traffic via high-dimensional behavior labeling and LLM-driven *causality-inspired* refinement (Figure 1), then conditions generation on metadata-enriched signals (persona summaries, characteristic dialogue patterns, and usage statistics). We use causal reasoning as a practical heuristic to remove redundant descriptors and retain compact root drivers, without

[1] Code will be released upon acceptance.

claiming causal identification.

We evaluate CausalPersona under a unified protocol that combines: (i) **distributional alignment** between synthetic and real corpora over **15 dimensions—3 structural** features (word frequency, user utterance length, full-conversation utterance length) and **12 behavioral** label dimensions, (ii) **LLM-as-a-Judge** assessments of persona similarity, tone consistency, goal consistency, and dialogue quality (Zhao et al., 2025). We adopt a deliberate **3+12** split: three stable, domain-agnostic surface statistics capture lexical/verbosity realism at turn and session scales, while twelve interaction-behavior dimensions capture intent dynamics, session control, and repair phenomena emphasized in large-scale assistant architectures and dialogue error analyses (Sarikaya, 2017; Cheng et al., 2024; Bohus and Rudnicky, 2005). Rather than treating any single number as an absolute bound, we use empirical real-data statistics as the reference for alignment fidelity.

Experiments across diverse datasets show that CausalPersona reduces behavioral divergence and improves population-level alignment over strong baselines, while maintaining high-quality, persona-consistent interactions. These results support population-aware benchmarking as a faithful basis for evaluating controllable dialogue generation beyond single-dialogue coherence.

Our contributions are threefold:

- We introduce **CausalPersona**, a population-aligned synthesis framework and benchmark that distills realistic, interpretable persona groups from large-scale traffic.
- We propose metadata-enriched conditioning signals that guide generation toward diverse real-world interaction strategies beyond local persona adherence.
- We establish a unified evaluation protocol combining 15-dimensional distributional alignment and LLM-as-a-Judge assessments, demonstrating consistent gains in fidelity and stability over strong baselines.

2 Related Work

LLMs for Persona Modeling and Evaluation. LLMs have been widely used to represent and leverage personas in dialogue generation. Early persona-based systems condition generation on speaker profiles to improve persona consistency and style control (Li et al., 2016; Mazaré et al., 2018; Zhang et al., 2018; Wolf et al., 2019). More recent work enriches sparse persona inputs into richer natural-language specifications, e.g., CPX (Kim et al., 2023), and proposes training-time mechanisms to better preserve persona faithfulness (Jandaghi et al., 2024). In parallel, LLMs have increasingly been adopted as *evaluators* of persona effectiveness. PersonaLens (Zhao et al., 2025) exemplifies the LLM-as-a-Judge paradigm for personalization by assessing tone, goal alignment, and preference adaptation, while other work studies rubric-based LLM evaluation and judging reliability (Zheng et al., 2023; Liu et al., 2023). However, existing benchmarks largely remain *dialogue-level* and typically treat personas as static descriptors, offering limited support for evaluating *population-level* behavioral realism.

Persona-Aided Dialogue Generation. Many systems use personas or user profiles to guide generation. Interlocutor (Occhipinti et al., 2025) models both speakers jointly to maintain consistency, while AutoPAL (Cheng et al., 2024) emphasizes longer-term adaptation through online persona updates and refinement. In-context learning approaches such as Towards Zero-Shot Persona Dialogue Generation with In-Context Learning (Xu et al., 2023) further steer generation via prompts and exemplars, consistent with classic persona-conditioned setups (Li et al., 2016; Mazaré et al., 2018; Wolf et al., 2019). Other work incorporates more structured guidance: DiaSynth (Suresh et al., 2025) combines persona-conditioned generation with subtopic expansion and reasoning-style prompting to improve coherence, and Faithful Persona (Jandaghi et al., 2024) uses iterative refinement to better align responses with persona constraints. Overall, these methods mainly improve local dialogue quality and controllability, but do not explicitly test whether synthesized conversations match *population-scale* distributions of user behaviors. Our work complements this line by distilling compact and interpretable persona groups from real traffic and benchmarking *population-level realism* via distributional alignment (in addition to dialogue-level judging), enabling comparable evaluation beyond single-dialogue coherence.

3 Task Formulation

Given a set of real user–AI conversations $\{x_i\}_{i=1}^N$ and their corresponding partial user histories $\{h_i\}_{i=1}^N$, our goal is to synthesize a set of conversations $\{\hat{x}_i\}_{i=1}^N$ under a unified *benchmark* protocol

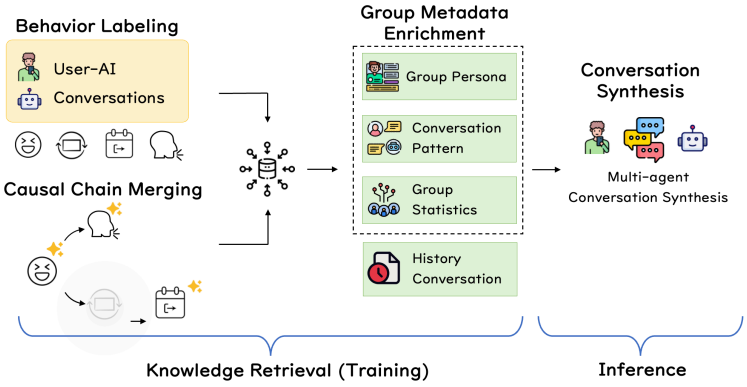

Figure 1: Framework overview. Training builds an enriched retrieval base from behavior labeling and causality-inspired refinement; inference retrieves group metadata and history to guide multi-agent conversation synthesis.

with two requirements:

1. **Distributional Alignment.** The synthesized corpus should match the real corpus in aggregate feature distributions. We evaluate alignment over **15 dimensions** (**3 structural + 12 behavioral**): structural features are word frequency, user utterance length, and full-conversation utterance length; behavioral features are induced by our 12-dimension behavior schema (e.g., primary intent, interaction mode, user type, brevity, persistence, and error-recovery behaviors) and the persona-group partitions *derived from* $\{x_i\}$ via behavior labeling, causality-inspired refinement, and metadata summarization. We report Jensen–Shannon (JS) divergence in the main tables and additional distance metrics (cosine/Euclidean) in the Appendix.

2. **LLM-based Conversational Quality.** Each $\hat{x}_i$ should be a plausible continuation of h_i, maintaining persona/tone/goal consistency over long horizons. We use an LLM-as-a-Judge protocol to score (i) logic alignment (persona similarity, tone consistency, goal consistency) and (ii) dialogue quality (fluency, coherence, informativeness). Group metadata m_i (persona summaries, discourse trajectories, interaction statistics) is *derived from* the real corpus and used only to specify the intended group profile for consistency checks.

We report population-level divergence $\mathcal{D}(\hat{x}_{1:N}, x_{1:N})$ and dialogue-level quality $\mathcal{Q}(\{\hat{x}_i\}; \{h_i, m_i\})$. A method is preferred if it reduces $\mathcal{D}$ while maintaining strong (non-degraded) $\mathcal{Q}$ under the same histories and derived group specifications.

4 Framework

CausalPersona follows a two-stage framework: *Offline Group Construction & Retrieval* and *Inference*. The design targets controllable long-context synthesis while preserving population-level realism in both structural and behavioral patterns.

Behavior Labeling. To model real-world interaction diversity, we annotate each conversation with a 12-dimensional behavior schema. The design targets two requirements: *population-level comparability* (stable and consistently measurable labels suitable for distributional analysis) and *generation controllability* (actionable, human-interpretable patterns that models can condition on). We therefore select dimensions that are (i) frequently observable in real interactions, (ii) largely domain-agnostic, and (iii) complementary—capturing *what* the user is trying to do, *how* they interact, and *how* they react to system imperfections. The schema integrates and extends principles from a commercial voice-assistant interaction taxonomy (Sarikaya, 2017), AutoPAL's activity-based profiling (Cheng et al., 2024), and classic dialogue breakdown and repair strategies (Bohus, 2004).

The 12 dimensions cover intent dynamics (e.g., primary intent, intent shift patterns), interaction management (e.g., interaction mode, session control, exit timing), and social/recovery signals (e.g., politeness, emotion patterns, clarification and repair behaviors). In addition, we record a small set of *session descriptors* (e.g., length buckets and turn-skew types) to support controllability and analysis; in our distributional evaluation, structural fidelity

is measured separately using continuous statistics (word frequency, user utterance length, and full-conversation utterance length), ensuring a clean separation between **3 structural** and **12 behavioral** dimensions.

For example, length buckets are introduced to align groups with distinct verbosity profiles during downstream generation, rather than treating length as an incidental outcome. More broadly, we avoid prematurely collapsing categories: we preserve a rich label space during early-stage group construction and defer abstraction until after *causality-inspired* refinement, which serves as a heuristic for removing redundant descriptors and identifying salient group drivers (without claiming causal identification) (Pearl et al., 2000). Overall, this behavior labeling framework supports fine-grained, scalable, and population-aligned conversation synthesis and evaluation.

Causality-inspired Chain Merging. To reduce redundancy and improve interpretability in group formation, we perform *causality-inspired* chain merging on the high-dimensional behavior labels. Rather than manually preselecting a fixed subset of grouping features, we use an LLM to infer a *directed dependency graph* among behavior dimensions (i.e., which dimensions are plausibly *derivable from* or *driven by* others in typical interactions). Concretely, given the 12-dimension schema definitions and the labeled values of a conversation, the LLM outputs a set of directed edges with brief rationales; we then treat dimensions with no incoming edges as *root drivers* and use only these roots to form the group identifier. We emphasize that this step is a *compression heuristic* for removing redundant descriptors and does not claim causal identification.

Figure 2 illustrates an example: for a user labeled as *casual chatter* with a *fact query* as the primary intent, the LLM suggests that downstream patterns such as *intent transition type* (*oscillating*) and *topic cohesion* (*drifting*) are largely explained by upstream persona/intent choices, while *session control pattern* (e.g., *smooth*) is primarily a function of interaction mode.[2] Accordingly, we exclude such dependent dimensions from the *group key* while preserving core traits like **user type**, **primary intent type**, and **interaction mode** as group drivers.

[2] See Figure 2 for the corresponding dependency illustration.

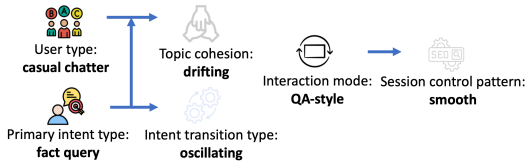

Figure 2: An example of causality-inspired chain merging. Arrows indicate *dependency/derivation* relations inferred for label compression: downstream descriptors (e.g., topic cohesion, session control) are excluded from the *group identifier* but retained for evaluation.

Importantly, chain merging only affects *group formation and conditioning* (i.e., the compact identifier used for retrieval and control). All original behavioral dimensions are still retained for *distributional evaluation*, ensuring the benchmark measures population-level alignment over the full behavior schema.

Group Metadata Enrichment. To mitigate sparse user history, we enrich each persona group (obtained via *causality-inspired* chain merging) with structured metadata distilled from real conversations. This metadata acts as a compact, interpretable summary of group-level interaction regularities, providing *data-driven priors* for *population-consistent* controllable generation without requiring additional supervision beyond the real corpus.

Each group is annotated with:

- **Group Persona:** a free-text summary of typical goals and communication style;
- **Conversation Pattern:** canonical discourse flows and intent-transition tendencies observed in the group;
- **Group Statistics:** empirical norms (e.g., turn-count and utterance-length distributions) that characterize the group.

These components play complementary roles during synthesis: the persona summary anchors tone and intent grounding, the pattern description encourages coherent turn-by-turn control trajectories, and the statistics regularize structural realism (e.g., expected brevity and typical stopping points).

Figure 3 illustrates a group characterized by casual music-control interactions. Even with a brief history, the metadata supplies concise priors—short command-style utterances and empirically observed length norms (e.g., ∼5–6 turns and ∼4 words per user turn on average, with natural variance)—which encourages realistic multi-turn

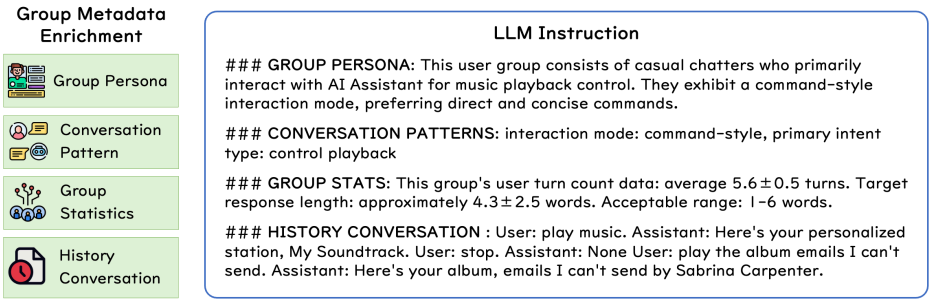

Figure 3: Example of group-guided conversation synthesis. Group metadata distilled from real traffic (persona summary, discourse patterns, and empirical statistics) provides data-driven priors that help generation remain concise and realistic for music-control interactions.

control behavior (e.g., play/pause/skip sequences) rather than drifting into verbose chit-chat. Overall, group-level enrichment helps CausalPersona generalize beyond sparse history while improving consistency with real-world interaction statistics and population-level *structural and behavioral* distributions.

Multi-agent Conversation Synthesis. To synthesize realistic long-horizon dialogues from limited history, we adopt a two-agent generation framework. *Offline*, we behavior-label the real corpus $\{x_i\}$, form compact persona groups via *causality-inspired* merging, and distill group metadata. *At inference*, given a partial history h_i, we retrieve the most compatible group (e.g., by matching the merged group key or nearest-neighbor retrieval over group representations) and use its metadata to guide turn-by-turn synthesis (Park et al., 2023; Li et al., 2023).

The **User Agent** conditions on group persona, discourse patterns, and empirical statistics to emulate user-side behaviors (intent expression, verbosity, transition tendencies), encouraging population-consistent interaction styles. The **Assistant Agent** generates informative, context-aware responses that maintain coherence and task relevance.

We avoid fixed-length generation and instead apply a group-informed stopping rule derived from empirical norms: the dialogue terminates when it substantially exceeds the group's typical turn budget (e.g., $> 1.5\times$ mean turns) or when recent turns indicate disengagement (e.g., sustained unusually short user responses relative to the group distribution). This reduces over-generation and better matches real stopping behavior.

By integrating group priors into both generation and control, the framework produces interactionally diverse conversations that better preserve population-level structural and behavioral distributions, aligning dialogue quality with distributional realism.

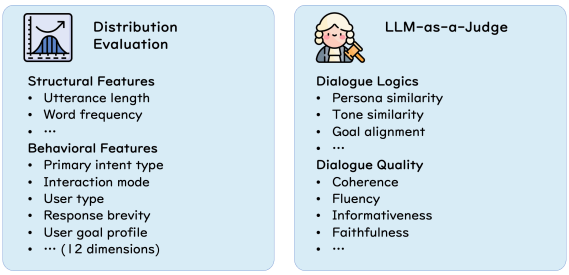

Figure 4: Our evaluation framework combines distributional alignment with LLM-as-a-Judge assessments, offering complementary views of population realism and dialogue quality.

Evaluation Protocol. We treat controlled conversation generation as a *benchmark*: under the same *data-derived* group specifications (persona/user-type groups and interaction constraints distilled from real traffic), each method generates a conversation corpus and is compared under a unified protocol. We evaluate (i) **population-level realism**—whether synthetic conversations match real distributional patterns—and (ii) **interaction quality**—whether dialogues remain coherent, grounded, and persona-consistent over long horizons. A comprehensive set of metrics is reported in Appendix E and summarized into two modules:

Distributional Evaluation. We compute corpus-level distances between synthetic and real data over (a) *structural features* (word frequency, user utterance length, full-conversation utterance length) and (b) *behavioral label distributions* induced by our 12-dimension behavior schema. The main tables

report Jensen–Shannon (JS) divergence for clarity; the Appendix provides additional distance functions (e.g., cosine/Euclidean) and dimension-level breakdowns. We also report a **Reference Bar** from a Real–Real split as an empirical scale reference (Real–Real).

LLM-as-a-Judge Evaluation. Complementary to corpus statistics, LLM judges score each dialogue on (i) *logic alignment* (persona similarity, tone consistency, goal consistency) and (ii) *dialogue quality* (fluency, coherence, informativeness), capturing failures that distributional metrics may miss and ensuring quality does not materially degrade.

Overall, the protocol jointly assesses *how realistic the generated population is* and *how well the conversations read*. For complete metric definitions and implementation details, refer to Appendix E.

5 Experiment

5.1 Datasets

We evaluate on five datasets. For each dataset, we filter conversations to 10–40 turns and then sample 1,000 for training and 1,000 for testing. Table 1 reports statistics for the public and Reddit datasets; internal statistics are only reported in aggregated form due to confidentiality.

Public Datasets (Basic & Advanced). We transform *MultiWOZ 2.2* (Budzianowski et al., 2018; Zang et al., 2020) and *Wizard of Wikipedia* (Dinan et al., 2018) into commercial voice-assistant-style dialogues via summarization followed by behavior-specific regeneration.[3] *Basic* produces short, direct user requests with minimal politeness; *Advanced* additionally enforces four behavior patterns: short requests, resistant responses, abrupt topic shifts, and early termination.

Pushshift Reddit Datasets (Basic & Advanced). We convert Pushshift Reddit threads (Baumgartner et al., 2020) into multi-turn dialogues by merging consecutive turns from the same author, removing boilerplate (URLs/mentions/subreddit markers), normalizing formatting, and filtering by turn/token ranges. *Basic* and *Advanced* follow the same behavior patterns as the public settings, adapted to Reddit's informal style.

Internal Dataset. We include interactions from a commercial voice assistant spanning diverse intents and dialogue flows. Due to privacy constraints, we report only aggregated metrics (no content release or subgroup breakdowns). Together, these datasets cover task-oriented, open-domain, and real-world interactions for diverse evaluation settings.

[3] MultiWOZ 2.2 is released under Apache 2.0. Wizard of Wikipedia is distributed through ParlAI (Miller et al., 2017).

5.2 Baselines

We compare **CausalPersona** with strong persona-aware dialogue generation baselines:

- **Interlocutor** (Occhipinti et al., 2025): Enhances speaker consistency by incorporating the interlocutor's persona.
- **AutoPAL** (Cheng et al., 2024): Enables long-term adaptation via real-time detection and periodic persona refinement.
- **ICL** (Pu et al., 2024): Applies in-context learning with diverse prompt exemplars to steer persona-driven generation.
- **DiaSynth** (Suresh et al., 2025): Combines persona and subtopic planning with chain-of-thought prompting to improve coherence.
- **Faithful Persona Generation** (Jandaghi et al., 2024): Utilizes a generator–critic loop to refine persona consistency and engagement.
- **CPX (ConceptPersona)** (Kim et al., 2023): Expands sparse persona seeds into rich conceptual representations for more diverse responses.
- **PersonaLens** (Zhao et al., 2025): We instantiate its persona-conditioned *generation* setting (used in its evaluation protocol) as a strong reference baseline for controllable, persona-grounded synthesis.

5.3 Model Setup

We use **Claude 3.5 Sonnet** as the backbone LLM for all LLM-based components in our pipeline. For fair comparison, we implement all LLM-based baselines with the *same* backbone model and matched decoding settings unless a method explicitly requires otherwise; full prompting and decoding details are deferred to the Appendix.

5.4 Evaluation Metrics

We evaluate each method primarily by **distributional alignment** (population realism), and report LLM-as-a-Judge scores as supplementary evidence.

Structural Metrics. We measure fidelity on **three** structural dimensions using Jensen–Shannon (JS) divergence: *user word frequency*, *user utterance length*, and *full-conversation utterance length*.

Table 1: Dataset statistics (descriptive). We report mean±std of (i) the total number of dialogue turns (counting both user and system utterances) and (ii) utterance length (words per utterance) for train/test splits after filtering conversations to 10–40 turns. "User" counts only user utterances, while "Full" aggregates utterances from both speakers (used by our structural dimensions alongside word-frequency statistics). Internal dataset statistics are reported only in aggregated form due to confidentiality.

Dataset	Avg. Turn Count		Avg. User Utterance Length		Avg. Full Utterance Length	
	Train	Test	Train	Test	Train	Test
Public Basic	23.8 ± 5.2	14.7 ± 4.1	4.7 ± 2.4	5.2 ± 2.5	13.6 ± 12.2	17.2 ± 13.3
Public Advanced	14.7 ± 3.3	14.7 ± 3.3	5.8 ± 2.6	5.8 ± 2.6	17.7 ± 13.1	17.7 ± 13.1
Reddit Basic	21.4 ± 4.8	22.2 ± 5.6	5.2 ± 1.1	4.8 ± 0.8	17.9 ± 14.4	17.6 ± 13.3
Reddit Advanced	19.4 ± 5.2	21.7 ± 6.3	5.7 ± 1.3	4.8 ± 1.3	18.8 ± 14.2	16.9 ± 12.9

Table 2: Conversation group statistics before and after *causality-inspired* merging of behavior-based group identifiers (followed by redundancy reduction). We report the number of groups for each dataset.

Dataset	Original Groups	Merged Groups
Internal	420	94
Public Basic	218	52
Public Advanced	406	73
Reddit Basic	396	76
Reddit Advanced	565	113
Average	**401**	**81.6**

Concretely, we convert each metric into an empirical discrete distribution: user word frequency is computed as normalized unigram counts over all user utterances, while length-based metrics are computed from discrete histograms over integer word counts (with standard smoothing to avoid zero-probability bins when needed). Lower JS values indicate better alignment.

Behavioral Metrics. We measure alignment across twelve behavioral dimensions (e.g., primary intent, interaction mode, user type, brevity, persistence, and error recovery), and compute JS divergence between the induced label distributions of real vs. synthetic corpora. For clarity, main tables report JS divergence, while the Appendix provides additional distance functions (e.g., cosine/Euclidean) and dimension-level breakdowns.

Reference Bar. We report a *Reference Bar* as an empirical scale indicator by computing the same divergences on a Real–Real split of the reference corpus. It is not a competing baseline; it serves only as a scale reference for interpreting distribution distances.

Auxiliary LLM-as-a-Judge Metrics. We additionally report LLM-as-a-Judge scores for semantic quality and persona consistency (Zheng et al., 2023; Liu et al., 2023) in Appendix D. These results are used to verify that dialogue quality does not materially degrade while optimizing for distributional realism; all main claims rely on the distributional metrics above.

5.5 Main Results Analysis

Table 3 compares CausalPersona with seven baselines on five datasets under our unified protocol. We focus on two distributional metrics (↓): **Structural Align.** (JS divergence averaged over the three structural dimensions) and **Behavioral Align.** (JS divergence averaged over the 12 behavioral label dimensions).

Overall Performance: CausalPersona achieves the best average **Structural Align.** (0.0683) and the best average **Behavioral Align.** (0.1892). Behavioral alignment provides the clearest separation from baselines, while structural alignment remains strong across datasets, including the more complex Internal setting (reported only as aggregated metrics due to confidentiality).

Cross-Dataset Consistency: CausalPersona yields the lowest behavioral divergence on all five datasets: Internal (0.2435), Public Basic (0.1465), Public Advanced (0.2304), Reddit Basic (0.1534), and Reddit Advanced (0.1720). Structurally, it attains the best average and leads on Internal and Public Basic; on Public Advanced, PersonaLens achieves the lowest structural divergence (0.0717) while CausalPersona remains competitive (0.0944), suggesting that CausalPersona's gains are not driven by dataset-specific tuning.

Baseline Comparison: PersonaLens is the strongest baseline overall, with baseline-best averages of 0.3114 (behavioral) and 0.0886 (structural). CausalPersona improves over PersonaLens by **39.2%** in behavioral alignment (0.3114 → 0.1892) and **22.9%** in structural alignment (0.0886 → 0.0683) on average.

Figure 5 visualizes these differences across datasets, highlighting CausalPersona's consistent behavioral gains (largest on Internal and the Reddit settings) while maintaining competitive structural

Table 3: Unified evaluation across datasets. ↓: lower is better (JS Divergence). We report **Structural Align.** (averaged over the three structural dimensions) and **Behavioral Align.** (averaged over 12 behavioral dimensions). *Reference Bar* is a Real–Real split scale indicator (intra-data variability), not a competing method or a theoretical bound. Additional distance metrics and dimension-level breakdowns are provided in the Appendix.

Method	Structural Align. (↓)						Behavioral Align. (↓)					
	internal	public_basic	public_advanced	reddit_basic	reddit_advanced	Average	internal	public_basic	public_advanced	reddit_basic	reddit_advanced	Average
AutoPAL	0.1735	0.1194	0.0807	0.0700	**0.0512**	0.0990	0.5854	0.2849	0.3677	0.1875	0.3063	0.3464
CPX	0.1807	0.0802	0.0985	0.0837	0.0662	0.1019	0.5438	0.3000	0.3750	0.1825	0.2926	0.3388
DiaSynth	0.1726	0.0979	0.0807	0.0611	0.0527	0.0930	0.5495	0.2916	0.3752	0.1585	0.3619	0.3473
Faithful	0.1520	0.0977	0.0857	0.0896	0.0852	0.1020	0.5500	0.2375	0.3332	0.1585	0.2956	0.3150
ICL	0.1288	0.0987	0.0981	0.0818	0.1196	0.1054	0.6406	0.5104	0.5006	0.4350	0.4612	0.5096
Interlocutor	0.1909	0.1462	0.0938	**0.0603**	0.0716	0.1126	0.5902	0.4039	0.4529	0.2663	0.3913	0.4209
PersonaLens	0.1198	0.1126	**0.0717**	0.0661	0.0728	0.0886	0.4870	0.2466	0.3145	0.2080	0.3009	0.3114
CausalPersona	**0.0472**	**0.0767**	0.0944	0.0656	0.0575	**0.0683**	**0.2435**	**0.1465**	**0.2304**	**0.1534**	**0.1720**	**0.1892**
Reference Bar	**0.0916**	**0.1687**	**0.1524**	**0.1329**	**0.1586**	**0.1408**	**0.0421**	**0.0647**	**0.0712**	**0.0869**	**0.1035**	**0.0737**

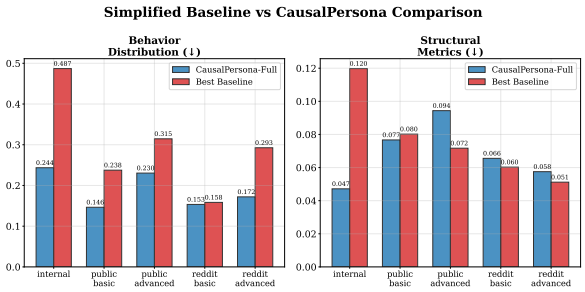

Figure 5: Baseline comparison across unified evaluation metrics. ↓ **indicates better** (lower divergence means closer alignment with reference data). Trends are consistent with Table 3.

alignment.

Performance Trade-offs: Some baselines are competitive on individual datasets or metrics (e.g., DiaSynth/Faithful on Reddit Basic behavioral, PersonaLens on Public Advanced structural), but CausalPersona achieves the most consistent average performance across datasets on both metrics, indicating robust population-level alignment rather than metric-specific specialization.

Reference Bar Interpretation: The *Reference Bar* is computed from a Real–Real split and serves as an empirical scale reference for intra-data variability. Because of finite-sample noise and distribution discretization, model–real distances can occasionally be comparable to (or lower than) this bar for some metrics; we use it only to contextualize magnitude.

Comprehensive Evaluation Context: Appendix A provides structural distribution analyses, Appendix B reports dimension-level behavioral breakdowns, Appendix C reports content-similarity analyses, and Appendix D reports LLM-as-a-Judge results as supplementary quality evidence.

6 Conclusion

We introduced **CausalPersona**, a benchmark and protocol for *population-level realism* in controlled conversation generation. Beyond per-dialogue coherence, our evaluation prioritizes *distributional alignment* of structural and behavioral patterns to real conversational traffic, with LLM-as-a-Judge used as a complementary quality check. We further presented CausalPersona as a strong reference implementation that distills interpretable persona groups via behavior labeling and causality-inspired refinement, enriches them with data-derived metadata, and synthesizes conversations with multi-agent generation. Across five datasets, CausalPersona consistently achieves the best behavioral alignment and the best average structural alignment among strong baselines, supporting population-aligned evaluation and more realistic conversational simulation.

Limitations

This work has several limitations. First, one dataset is internal and cannot be released for privacy, limiting reproducibility; we report only aggregated internal metrics and evaluate on public datasets under the same protocol. Second, our pipeline relies on behavior labeling, causal merging, and metadata summarization; upstream errors may propagate to group formation and downstream generation. Third, distributional metrics capture population realism but miss other quality aspects and are sensitive to discretization and sample size. Finally, LLM-as-a-Judge may be biased, so we use it only as supplementary evidence. Extending to more languages/domains and adding stronger safety-oriented evaluations remain future work.

Ethical Considerations

This work studies *proactive* prediction of whether a user turn may elicit a toxic assistant response, with the goal of enabling earlier safeguards rather than generating unsafe content. We use existing datasets and do not collect new user data; we follow dataset terms and handle all text as potentially sensitive. We do not attempt to de-anonymize users, and we recommend minimizing retention of raw conversations and restricting access when deploying similar models.

A key risk is **over-blocking** (false positives), which can disproportionately affect benign discussions of sensitive topics or dialectal language. To mitigate this, the model should be used as a *decision-support* signal (e.g., thresholded risk + trigger attributions) with calibrated thresholds, fallback behaviors that preserve user utility, and human review for high-stakes settings. Another risk is **misuse** for censorship or targeted suppression; we advocate releasing models and artifacts with clear usage guidelines emphasizing safety, accountability, and auditability. Finally, trigger attributions may enable adaptive attacks; therefore, deployments should consider limiting attribution granularity and monitoring for adversarial prompting.

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Dan Bohus. 2004. Error awareness and recovery in task-oriented spoken dialogue systems. *Retrieved at*, 49.

Dan Bohus and Alexander Rudnicky. 2005. Sorry and i didn't catch that!-an investigation of non-understanding errors and recovery strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 128–143.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Yi Cheng, Wenge Liu, Kaishuai Xu, Wenjun Hou, Yi Ouyang, Chak Tou Leong, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Autopal: Autonomous adaptation to users for personal ai companionship. *arXiv preprint arXiv:2406.13960*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139.

Donghyun Kim, Youbin Ahn, Chanhee Lee, Wongyu Kim, Kyong-Ho Lee, Donghoon Shin, and Yeonsoo Lee. 2023. Concept-based persona expansion for improving diversity of persona-grounded dialogue. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3471–3481.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B 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)*, pages 994–1003.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5454–5459.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 conference on empirical methods in natural language processing: system demonstrations*, pages 79–84.

Daniela Occhipinti, Marco Guerini, and Malvina Nissim. 2025. When harry meets superman: The role of the interlocutor in persona-based dialogue generation. *arXiv preprint arXiv:2505.24613*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3.

Jiashu Pu, Yajing Wan, Yuru Zhang, Jing Chen, Ling Cheng, Qian Shao, Yongzhu Chang, Tangjie Lv, and Rongsheng Zhang. 2024. Crafting a good prompt or providing exemplary dialogues? a study of in-context learning for persona-based dialogue generation. *arXiv preprint arXiv:2402.09954*.

Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

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? *arXiv preprint arXiv:1801.07243*.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. Personalens: A benchmark for personalization evaluation in conversational ai assistants. *arXiv preprint arXiv:2506.09902*.

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.

A Structural Distribution Analysis: Detailed Results

This appendix provides comprehensive results for the structural distribution analysis, comparing CausalPersona with baseline methods across core user structural metrics and datasets. We present detailed ablation studies, baseline comparisons, and training size sensitivity analysis to demonstrate the effectiveness of our approach in maintaining structural alignment with reference conversations.

A.1 Causality-inspired Chain Merging Methodology

Our approach employs a sophisticated causal reasoning framework for behavioral label selection, which is fundamental to CausalPersona's structural modeling capabilities. Table 4 presents the detailed prompt used for causal chain merging in behavioral label selection.

This causal reasoning approach ensures that CausalPersona identifies the most fundamental behavioral drivers rather than surface-level patterns, leading to more robust structural distribution modeling. The methodology emphasizes:

- **Root Cause Priority**: Selecting fundamental drivers rather than consequences
- **Minimal Redundancy**: Choosing only the most causally significant labels from similar candidates
- **Practical Relevance**: Selecting labels that meaningfully affect controllable generation behavior
- **Causal Independence**: Maintaining independence between effective labels to avoid multicollinearity

A.2 Core User Structural Metrics

Our evaluation focuses on three core structural metrics that capture essential conversation characteristics:

- **User Word Frequency**: Captures the distribution of user lexical usage patterns
- **User Utterance Length**: Measures the distribution of user utterance lengths, reflecting user engagement and request complexity
- **Full Utterance Length**: Evaluates the distribution of overall utterance lengths across the full dialogue, reflecting dialogue flow and completeness

For all metrics, **lower values indicate better distribution alignment** with reference data, representing superior structural fidelity to natural conversation patterns.

Table 4: **Prompt used for causality-inspired label merging.** The LLM selects a minimal set of *effective labels* (root causes) from the 12-dimensional behavior annotations to form compact, interpretable group identifiers.

Causal Reasoning Prompt for Label Merging

You are a behavioral reasoning expert. Select a MINIMAL subset (2–4 labels MAX) of **effective labels** that capture the ROOT CAUSES of conversation behavioral patterns.

ANALYSIS CRITERIA: (1) Root Cause Priority: Select fundamental drivers, not consequences; (2) No Redundancy: Choose ONLY the most causal label from similar ones; (3) Impact Threshold: Label must significantly influence conversation dynamics; (4) Independence: Effective labels should be causally independent.

REASONING: For each label ask: Is this a CAUSE or EFFECT? Can this be explained by other labels? Does this fundamentally change behavior?

GROUPING: User characteristics → Choose most fundamental; Intent patterns → Choose most causally significant; Interaction patterns → Choose core driver; Response patterns → Usually effects, not causes.

FORMAT:

INPUT: `{metadata_json}`

OUTPUT:

```
{
  "effective_labels": ["key1: value1", "key2: value2"],
  "excluded_labels": ["key3: value3 (reason)", ...]
}
```

OBJECTIVE: Find MINIMUM labels explaining MAXIMUM behavioral variance. Most labels should be excluded!

A.3 Ablation Study Results

Table 5 reports ablation results across datasets, illustrating how each CausalPersona component affects structural distribution modeling.

A.4 Ablation Studies

A.4.1 Structural Component Ablation Findings

The ablation studies reveal several critical insights about CausalPersona's architectural components:

Note: While improvements are generally consistent across metrics and datasets, per-metric changes are *not uniformly monotonic*. We therefore report means and discuss representative counter-examples in Appendix, and base claims on aggregate trends rather than single cells.

1. **Statistical Component Criticality**: The "w/o Stats" configuration most strongly hurts length-related structural alignment on several datasets. For example, on Public Basic, removing statistics increases the JS divergence on *Full Utterance Length* from 0.0254 (Full) to 0.1029 (w/o Stats), a 4.1× increase. On Internal, *User Utterance Length* increases from 0.0512 to 0.1961 (3.8×). This highlights the importance of statistical modeling for matching real traffic length distributions.

2. **Component Synergy**: While individual components (History, Persona, Rules) show moderate performance variations when removed, the full CausalPersona consistently achieves competitive performance across metrics, indicating effective component integration and balanced architectural design.

3. **Cross-Dataset Robustness**: The full CausalPersona maintains superior or competitive performance across diverse datasets (Public Basic/Advanced, Reddit Basic/Advanced, Internal), demonstrating robust generalization capabilities across different conversation domains and complexity levels.

4. **Metric-Specific Patterns**: Different structural metrics show varying sensitivity to component removal:

   - **User Utterance Length** can be highly sensitive to statistical components (e.g., Internal), reflecting the role of statistics in matching user verbosity patterns
   - **Full Utterance Length** is particularly affected by removing statistics, indicating that global length/flow distributions require explicit distribution matching
   - **User Word Frequency** is comparatively more stable, with smaller but non-negligible shifts under component removal

Table 5: Structural Metrics (JS Divergence): CausalPersona Ablation Study. ↓: lower is better (distribution alignment).

Method	Public Basic			Public Advanced			Reddit Basic			Reddit Advanced			Internal		
	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length
Full	0.0500	0.0660	0.0254	**0.0628**	0.1133	0.0540	0.0923	0.0799	0.1108	0.0847	0.0387	0.0733	0.0733	0.0512	0.0480
w/o History	0.0442	0.1374	**0.0105**	0.1523	0.1350	0.0515	**0.0652**	0.0964	0.0989	**0.0640**	0.0378	0.0770	0.1237	0.0479	**0.0306**
w/o Persona	0.0487	0.1509	0.0293	0.0741	0.1223	0.0628	0.0727	0.0946	0.0984	0.0832	0.0398	0.0840	**0.0655**	0.0916	0.0572
w/o Rules	0.0410	0.1429	0.0332	0.1018	0.1133	**0.0435**	0.0973	0.0799	0.1023	0.0764	0.0760	0.0772	0.0751	**0.0450**	0.0478
w/o Stats	**0.0386**	**0.0411**	0.1029	0.1432	**0.0663**	0.0512	0.0714	**0.0640**	**0.0289**	0.0818	**0.0324**	**0.0442**	0.0872	0.1961	0.0455
Reference Bar	**0.0793**	**0.1129**	**0.0827**	**0.1652**	**0.1978**	**0.2285**	**0.1427**	**0.1764**	**0.2053**	**0.1589**	**0.1886**	**0.1621**	**0.2145**	**0.2012**	**0.2398**

A.5 Baseline Comparison Results

Table 6 presents detailed comparisons with state-of-the-art baseline methods across all core structural metrics and datasets.

A.6 Primary Metric Breakdowns

A.6.1 Structural and Behavioral Baseline Comparison

The baseline comparison highlights three consistent findings.

1. **CausalPersona achieves lowest behavioral divergence.** Table 3 shows an average behavioral score of **0.1892**, compared to **0.3114** for the strongest baseline (PersonaLens), representing a **39.2%** improvement in behavioral alignment.

2. **Structural alignment remains strong.** Structural alignment (avg. **0.0683**) remains strong (computed over our three structural dimensions), indicating that behavioral gains do not come at the expense of structural realism.

3. **Cross-domain robustness.** Performance gains hold across Internal, Public, and Reddit datasets, confirming generalization of group-conditioned synthesis under diverse conversation settings.

Key Takeaways. CausalPersona achieves consistent improvements in population-level fidelity, especially in behavioral metrics, while preserving realistic structural distributions. This establishes CausalPersona as a reliable tool for both evaluation and controlled data generation.

The unified comparison further suggests that CausalPersona's advantages are not confined to a single domain: it remains best-in-average across both metrics while staying competitive on each dataset individually (Table 3).

1. **Strong performance on public datasets.** On both Public Basic and Public Advanced, CausalPersona achieves the lowest behavioral divergence and remains competitive on structural alignment, indicating effective alignment after converting public data into conversation mode.

2. **Competitive performance on Reddit and Internal.** On the more informal Reddit settings and the more complex Internal setting (reported only as aggregated metrics), CausalPersona maintains the lowest behavioral divergence. Its structural alignment is also competitive across these settings, and it attains the best *average* structural alignment overall, suggesting robustness under distribution shifts.

3. **Reference Bar context.** The Reference Bar is included as a scale reference point rather than a competing method. As shown in Table 3, CausalPersona remains the best-performing *method* on behavioral alignment while also achieving the lowest average structural divergence.

A.7 Training Size Sensitivity Analysis

To evaluate the robustness of CausalPersona across different training data sizes, we conduct a comprehensive sensitivity analysis examining performance with varying amounts of training data. Table 7 presents detailed results across different training size configurations.

A.7.1 Training Size Configurations

Our sensitivity analysis evaluates CausalPersona performance across four training data configurations:

- **CausalPersona (Full)**: Training with complete available training data
- **CausalPersona (500)**: Training with 500 conversation samples per dataset
- **CausalPersona (200)**: Training with 200 conversation samples per dataset
- **CausalPersona (100)**: Training with 100 conversation samples per dataset

Table 6: Structural Metrics (JS Divergence): CausalPersona vs Baselines. ↓: lower is better (distribution alignment).

Method	Public Basic			Public Advanced			Reddit Basic			Reddit Advanced			Internal		
	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length
CausalPersona	0.0500	**0.0660**	**0.0254**	**0.0628**	0.1133	**0.0540**	0.0923	0.0799	0.1108	0.0847	0.0387	0.0733	**0.0733**	0.0512	0.0480
AutoPAL	0.0428	0.2712	0.2065	0.1515	0.0922	0.1144	0.0726	0.1050	0.0646	0.0809	0.0661	0.0631	0.0778	0.0397	**0.0361**
CPX	0.0631	0.2787	0.2003	0.0959	**0.0449**	0.0998	0.0775	0.1153	0.1027	0.1002	0.0522	0.0988	0.0899	0.0583	0.0504
DiaSynth	0.0600	0.2685	0.1893	0.1683	0.0528	0.0727	0.0783	0.0902	0.0736	**0.0763**	0.0534	**0.0537**	0.0764	0.0407	0.0409
Faithful	0.0428	0.2030	0.2102	0.1386	0.0805	0.0740	0.0734	0.0912	0.0926	0.0771	0.1219	0.0699	0.0766	0.1180	0.0610
ICL	0.0661	0.1544	0.1661	0.1323	0.0638	0.1001	0.0784	0.1849	**0.0310**	0.0819	0.0848	0.0786	0.1052	0.1468	0.1070
Interlocutor	**0.0389**	0.2706	0.2632	0.1353	0.1179	0.1853	**0.0720**	0.0669	0.1426	0.0777	**0.0113**	0.0918	0.0783	**0.0201**	0.1164
PersonaLens	0.0538	0.1801	0.1254	0.1424	0.0764	0.1190	0.0791	**0.0492**	0.0868	0.0793	0.0200	0.0990	0.0881	0.0366	0.0937
Reference Bar	**0.0793**	**0.1129**	**0.0827**	**0.1652**	**0.1978**	**0.2285**	**0.1427**	**0.1764**	**0.2053**	**0.1589**	**0.1886**	**0.1621**	**0.2145**	**0.2012**	**0.2398**

Table 7: Structural Metrics: CausalPersona Training Size Sensitivity Study. ↓: lower is better (distribution alignment).

Method	Public Basic			Public Advanced			Reddit Basic			Reddit Advanced			Internal		
	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length	User Word Freq	User Utt Length	Full Utt Length
CausalPersona (Full)	**0.0500**	**0.0660**	**0.0254**	**0.0628**	0.1133	**0.0540**	0.0923	0.0799	0.1108	0.0847	0.0387	**0.0733**	**0.0733**	0.0512	**0.0480**
CausalPersona (100)	0.0593	0.0783	0.0302	0.0690	0.1244	0.0593	0.1048	0.0907	0.1258	0.0944	0.0432	0.0817	0.0987	0.0689	0.0645
CausalPersona (200)	0.0559	0.0738	0.0284	0.0653	0.1178	0.0561	0.1021	0.0883	0.1225	0.0890	0.0407	0.0770	0.0840	0.0587	0.0549
CausalPersona (500)	0.0512	0.0662	0.0258	0.0629	0.1136	0.0542	0.0927	0.0802	0.1103	0.0845	0.0386	0.0736	0.0735	0.0511	0.0482
Best SOTA	0.0538	0.1801	0.1254	0.1424	**0.0764**	0.1190	**0.0791**	**0.0492**	**0.0868**	**0.0793**	**0.0200**	0.0990	0.0881	**0.0366**	0.0937
Reference Bar	**0.0793**	**0.1129**	**0.0827**	**0.1652**	**0.1978**	**0.2285**	**0.1427**	**0.1764**	**0.2053**	**0.1589**	**0.1886**	**0.1621**	**0.2145**	**0.2012**	**0.2398**

This systematic evaluation allows us to understand how CausalPersona's performance scales with training data availability, providing insights into its practical deployment requirements and data efficiency.

A.7.2 Key Sensitivity Analysis Findings

The comprehensive sensitivity analysis reveals several critical insights about CausalPersona's robustness and data efficiency:

1. **Data efficiency with small samples.** CausalPersona remains stable under reduced training data. On Public Basic, the mean JS divergence across the three structural metrics increases from 0.0471 (Full) to 0.0559 (100 samples), an 18.7% relative increase but still a low absolute divergence.

2. **Consistent cross-dataset robustness.** Similar stability holds across datasets (mean JS divergence across the three metrics, Full → 100): Public Advanced 0.0767 → 0.0842, Reddit Basic 0.0943 → 0.1071, Reddit Advanced 0.0656 → 0.0731, and Internal 0.0575 → 0.0774.

3. **Metric-specific stability.** Reducing training data typically causes small, smooth shifts in *User Word Frequency*, *User Utterance Length*, and *Full Utterance Length*, with the largest sensitivity on Internal (e.g., User Word Frequency 0.0733 → 0.0987 from Full to 100).

4. **Strong performance relative to Best SOTA.** Even with 100 samples, CausalPersona is often competitive with or better than the Best SOTA baselines on key length/flow metrics. For example, on Public Basic, CausalPersona (100) achieves (0.0593, 0.0783, 0.0302) while Best SOTA is (0.0538, 0.1801, 0.1254), showing substantially better length alignment.

5. **Practical deployment implications.** The small degradation with 200–500 samples suggests CausalPersona can be deployed in resource-constrained settings or data-scarce domains.

6. **Reference Bar consistency.** For *structural* metrics, CausalPersona achieves distances on a comparable scale to the Reference Bar on multiple datasets, suggesting robust distributional alignment under reduced training data.

A.8 Cross-Metric Consistency and Architectural Strengths

A notable strength of CausalPersona is its consistent performance across different structural metrics and training configurations. Unlike baseline methods that may excel in specific aspects, CausalPersona provides balanced performance across:

- **Lexical Usage Modeling**: Consistent User Word Frequency alignment indicates effective modeling of user lexical patterns
- **User Engagement Modeling**: Stable User Utterance Length results reflect robust modeling of user verbosity patterns
- **Conversation Flow Modeling**: Consistent Full Utterance Length alignment demonstrates robust modeling of dialogue length/flow

A.9 Implications for Conversation Generation

These comprehensive results demonstrate that CausalPersona's causal modeling approach provides fundamental advantages for structural distribution alignment:

1. **Holistic Structural Modeling**: Unlike baselines that may excel in specific aspects, CausalPersona provides balanced performance across lexical usage, user engagement, and conversation flow dimensions.

2. **Statistical Foundation**: The critical importance of statistical components (as shown in ablation studies) validates CausalPersona's emphasis on rigorous statistical modeling of conversation structures through causal reasoning.

3. **Generalization Capability**: Consistent performance across diverse datasets (public conversations, Reddit discussions, internal data) demonstrates broad applicability and robust architectural design.

4. **Data Efficiency**: The exceptional performance with limited training data makes CausalPersona highly practical for deployment in resource-constrained environments or specialized domains with limited conversation data.

5. **Scalable Architecture**: The component-based design allows for systematic improvement and adaptation to different conversation domains while maintaining structural fidelity and causal reasoning capabilities.

6. **Causal Reasoning Integration**: The causal chain merging methodology ensures that structural modeling is grounded in fundamental behavioral drivers rather than surface-level patterns, leading to more robust and interpretable conversation generation.

These results establish CausalPersona as a significant advancement in conversation generation, particularly for applications requiring high structural fidelity to reference conversation patterns while maintaining efficiency and robustness across diverse deployment scenarios.

B Behavior Distribution Analysis - Detailed Comparison

This appendix provides a comprehensive analysis of behavior dimension performance, examining how well CausalPersona and baseline methods align with reference conversation distributions across twelve distinct behavioral characteristics. The behavior dimension analysis offers crucial insights into the quality and naturalness of synthesized conversations by measuring distributional alignment with human conversation patterns.

B.1 Behavior Dimensions and Metrics

B.1.1 Behavior Dimension Definitions

We evaluate conversation synthesis quality across twelve categorical behavior dimensions (Table 27). Each dimension is annotated using a fixed label set; when a dimension is not reliably inferable, it is assigned `unknown`.

- **Primary Intent Type**: The dominant intent category exhibited across the dialogue (e.g., *Task*, *Info-seeking*, *Chitchat*, *Music*, *Reminder*, *Alarm*, *Smart Home*).
- **Interaction Mode**: The user's typical interaction style with the assistant, characterized by the dialogue structure (e.g., *Command-style*, *QA-style*, *Multi-turn Dialog-style*).
- **User Type**: A coarse usage-pattern category reflecting how the user engages with the assistant over the session (e.g., *Power User*, *One-shot User*, *Tasker*, *Casual*).
- **Response Brevity**: The overall brevity of user utterances (e.g., *Short*, *Medium*, *Long*).
- **User Goal Profile**: The user's goal pursuit style over the session (e.g., *Tasker*, *Explorer*, *Goal-switcher*, *Chatter*).
- **Persistence Level**: The extent to which the user persists after failures or friction (e.g., *High*, *Medium*, *Low*).
- **Topic Cohesion**: How consistently the user maintains a coherent topic across turns (e.g., *High*, *Medium*, *Fragmented*).
- **Politeness Strategy**: The dominant politeness style in user requests (e.g., *Direct*, *Polite*, *Formal*, *Friendly*).
- **Interaction Flexibility**: Whether the user adapts their approach when the conversation deviates or the assistant fails (e.g., *Adaptive* vs. *Static*).
- **Persona Adaptability**: Whether the user noticeably shifts their persona/style during the interaction (e.g., *High* vs. *Low*).

- **Repair Behavior**: The user's dominant repair strategy after breakdowns (e.g., *Rephrase*, *Clarify*, *Ignore*, *Retry*).
- **Error Recovery Style**: The user's overall recovery approach in error states (e.g., *Clarification Request*, *Repetition*, *Restart*, *Task Abandonment*).

B.1.2 Distance Metrics

We use complementary distances to quantify distributional alignment between real and synthetic corpora.

Jensen–Shannon (JS) (Lin, 2002): We report JS as the primary distributional metric (↓). JS is symmetric and bounded, making it a stable measure for comparing discrete empirical distributions.

Cosine Distance: Measures the angular difference between distribution vectors, emphasizing similarity in relative proportions across categories (↓).

Euclidean Distance: Measures the ℓ_2 gap between distribution vectors and is sensitive to absolute deviations in bin/category probabilities (↓).

For all metrics, lower values indicate closer alignment to the reference distributions.

B.2 Ablation Study Analysis

Tables 8 and 9 present comprehensive ablation studies examining the contribution of each CausalPersona component to behavior dimension performance.

B.2.1 Behavioral Component Importance Analysis

The ablation results reveal critical insights about component contributions:

Statistical Component (Stats) is Most Critical: Among the ablations, removing the statistical component (w/o Stats) leads to one of the largest and most consistent degradations in distribution alignment, especially for behavioral realism. For example, on the Public Basic dataset, the average Cosine distance increases from 0.0345 (full model) to 0.1634 (w/o Stats), a 374% relative increase. This suggests that statistical grounding plays a key role in maintaining realistic behavior distributions.

Persona Component Shows Strong Impact: The w/o Persona ablation reveals substantial performance drops, particularly evident in dimensions like "Persona Adaptability" and "User Type." On the Internal dataset, removing persona information increases the average Jensen-Shannon distance from 0.2435 to 0.3339, highlighting the importance of persona-aware synthesis.

History Component Provides Contextual Grounding: While generally showing smaller impacts than Stats or Persona removal, the w/o History ablation demonstrates notable effects on dimensions related to conversation flow and coherence, such as "Topic Cohesion" and "Interaction Flexibility."

Rules Component Offers Consistent Benefits: The w/o Rules ablation shows moderate but consistent performance degradation across most dimensions, suggesting that explicit behavioral rules provide valuable constraints for maintaining natural conversation patterns.

B.3 Baseline Comparison Analysis

Tables 10 and 11 provide comprehensive comparisons between CausalPersona and seven state-of-the-art baseline methods across all behavior dimensions.

B.3.1 Performance Superiority Analysis

CausalPersona shows strong cross-dataset generalization, achieving the best *average* performance across datasets on both structural and behavioral alignment (Table 3). Notably, it delivers consistent gains on behavioral alignment across all datasets, while remaining competitive on structural alignment with a few dataset-specific wins by strong baselines (e.g., PersonaLens on public_advanced; Interlocutor/DiaSynth on reddit_basic).

Baseline Performance Patterns: Among baselines, PersonaLens and Faithful show relatively competitive performance, while ICL consistently exhibit poor behavior alignment. ICL's particularly poor performance (e.g., 0.4619 average Cosine distance on Public Basic vs. CausalPersona's 0.0345) suggests that simple in-context learning approaches struggle with complex behavioral modeling.

Dimension-Specific Advantages: CausalPersona shows particularly strong advantages in dimensions requiring sophisticated understanding of user psychology and conversation dynamics, such as "Persona Adaptability," "Error Recovery Style," and "Persistence Level."

B.4 Cross-Dataset Performance Patterns

B.4.1 Dataset Complexity Analysis

Performance varies significantly across datasets, revealing important patterns:

Table 8: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona (Full)	0.0402	0.0512	0.0387	**0.0000**	0.0167	0.0380	0.0351	**0.0009**	0.1245	**0.0000**	0.0112	0.0234	**0.0345**
w/o History	0.0359	0.0477	0.0424	**0.0000**	0.0033	0.2344	0.0142	0.0024	**0.0053**	0.0105	0.0152	0.0077	0.0381
w/o Persona	0.0429	0.0527	0.1774	**0.0000**	0.0540	0.0912	0.0411	0.0139	0.1317	0.1641	0.0257	0.0572	0.0774
w/o Rules	**0.0140**	0.0556	**0.0297**	**0.0000**	0.0243	0.0303	0.0237	**0.0009**	0.0221	0.1266	0.0155	0.0430	0.0351
w/o Stats	0.0553	**0.0330**	0.0633	0.7399	**0.0027**	0.1946	0.0468	0.3552	0.1949	0.2685	**0.0043**	**0.0022**	0.1634
Best SOTA	0.0338	0.0763	0.1335	0.4453	0.0028	**0.0069**	**0.0018**	0.0831	0.1886	0.3368	0.0152	0.0146	0.1116
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Public Advanced													
CausalPersona (Full)	**0.0024**	0.1492	0.1112	**0.0002**	0.1504	0.1285	0.4228	**0.0000**	**0.0209**	0.0687	0.2538	0.5383	**0.1679**
w/o History	0.0086	0.1010	0.2131	**0.0002**	0.3295	0.3955	0.3892	0.0002	0.0241	**0.0179**	0.1172	**0.5160**	0.1760
w/o Persona	0.0160	0.1428	0.0841	**0.0002**	0.1667	0.1812	0.4612	0.0078	0.2452	0.5645	0.3080	0.6148	0.2327
w/o Rules	0.0144	0.2232	0.0965	**0.0002**	0.1809	**0.1276**	0.4405	0.0002	0.0529	0.0835	0.3238	0.5895	0.1778
w/o Stats	0.0329	0.0850	0.3280	0.3462	0.1426	0.4857	0.4210	0.0020	0.2063	0.1840	**0.1152**	0.5312	0.2400
Best SOTA	0.0038	**0.0029**	**0.0314**	0.4921	**0.0946**	0.1950	**0.3152**	0.0651	0.3472	0.4365	0.2425	0.5244	0.2292
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Reddit Basic													
CausalPersona (Full)	0.0006	0.0323	0.0295	**0.0011**	0.0017	0.0955	0.0013	0.0007	0.0166	**0.0049**	0.1018	0.1018	**0.0323**
w/o History	**0.0002**	0.0323	**0.0114**	0.0031	0.2348	0.1904	0.0011	0.0021	0.3929	0.0231	0.1493	0.1493	0.0992
w/o Persona	0.0013	0.0212	0.0446	0.0013	0.0073	0.0579	0.0092	0.0029	0.0096	0.0231	0.1119	0.1006	0.0326
w/o Rules	0.0005	0.0384	0.0538	0.0024	0.0006	0.1116	0.0050	**0.0003**	0.0306	0.0114	0.0902	0.0902	0.0363
w/o Stats	0.0003	**0.0006**	0.0438	0.7474	0.0007	**0.0360**	0.0024	0.2977	**0.0066**	0.0343	0.0573	0.0573	0.1070
Best SOTA	0.0003	0.2407	0.0422	0.7432	**0.0004**	0.0479	**0.0007**	0.2187	0.0067	0.1431	**0.0133**	**0.0124**	0.1225
Reference Bar	0.0013	0.0068	0.0116	0.0034	0.0005	0.0022	0.0011	0.0004	0.0037	0.0365	0.0194	0.0328	0.0100
Reddit Advanced													
CausalPersona (Full)	**0.0000**	**0.0029**	0.1822	0.0002	0.0020	0.3262	0.0274	0.0010	**0.0118**	0.1196	0.0302	0.0547	0.0689
w/o History	**0.0000**	0.0076	0.2304	**0.0000**	0.0068	0.3152	0.0300	**0.0000**	0.0338	**0.0066**	0.1008	0.1066	0.0931
w/o Persona	0.0002	0.0049	0.1029	**0.0000**	**0.0013**	0.1765	0.0524	0.0002	0.0819	0.2818	0.0464	0.0635	0.0738
w/o Rules	**0.0000**	0.0049	0.1633	0.0002	0.0035	0.1098	0.0451	0.0005	0.0568	0.1196	0.0231	0.0309	**0.0507**
w/o Stats	0.0002	0.0069	0.0679	0.2948	0.0043	0.0721	**0.0070**	0.1524	0.1522	0.3186	**0.0047**	**0.0159**	0.0914
Best SOTA	0.0010	0.3297	**0.0295**	0.7636	0.0039	**0.0633**	0.0411	0.2667	0.2074	0.4633	0.0088	0.0206	0.1832
Reference Bar	0.0006	0.0003	0.0024	0.0001	0.0133	0.0402	0.0253	0.0001	0.0023	0.0144	0.0602	0.0479	0.0173
Internal													
CausalPersona (Full)	0.0082	0.0188	0.0704	0.0018	0.1612	0.0948	0.3755	0.0015	0.0291	0.1410	0.5441	0.5918	0.1699
w/o History	0.1240	0.0169	0.0834	0.0015	**0.0505**	0.1500	0.2682	**0.0006**	0.0663	**0.1167**	0.5703	0.6314	0.1733
w/o Persona	0.0127	0.0474	**0.0377**	**0.0005**	0.2012	0.2743	0.4152	0.0440	0.2762	0.5998	0.6933	0.7363	0.2782
w/o Rules	0.0060	**0.0153**	0.0846	0.0014	0.1654	**0.0774**	0.3680	0.0007	**0.0225**	0.1363	0.5324	0.5781	**0.1657**
w/o Stats	**0.0052**	0.1315	0.1243	0.3583	0.0800	0.0880	**0.2649**	0.0099	0.2280	0.1342	**0.4466**	**0.4682**	0.1949
Best SOTA	0.1860	0.5511	0.1145	0.2676	0.2998	0.4630	0.3592	0.3070	0.6446	0.7772	0.7331	0.7756	0.4566
Reference Bar	0.0039	0.0029	0.0046	0.0004	0.0072	0.0015	0.0008	0.0000	0.0053	0.0013	0.0079	0.0083	0.0037
Average	**0.0129**	**0.0509**	**0.0864**	**0.0008**	**0.0664**	**0.1366**	**0.1724**	**0.0010**	**0.0406**	**0.0668**	**0.1882**	**0.2620**	**0.0904**

Internal Dataset Challenges: The Internal dataset consistently shows the highest (worst) distance values across all methods, indicating greater complexity in matching internal conversation patterns. CausalPersona's average Jensen-Shannon distance of 0.2435 on Internal vs. 0.1534 on Reddit Basic suggests that internal conversations exhibit more complex behavioral patterns.

Reddit Dataset Characteristics: Reddit datasets (Basic and Advanced) generally show intermediate difficulty levels, with CausalPersona achieving strong performance (e.g., 0.0323 average Cosine distance on Reddit Basic). The relatively good performance suggests that Reddit conversations, while diverse, follow more predictable behavioral patterns.

Public Dataset Variability: Public datasets show varying difficulty, with Public Advanced being more challenging than Public Basic, likely due to increased conversation complexity and diversity in the advanced version.

B.4.2 Generalization Capabilities

The cross-dataset analysis reveals CausalPersona's strong generalization:

Consistent Relative Performance: While absolute performance varies across datasets, CausalPersona remains highly competitive across all datasets, showing robust generalization and particularly consistent gains on behavioral alignment.

Adaptive Behavior Modeling: The method's ability to maintain behavior alignment across diverse datasets (from Reddit discussions to internal conversations) indicates effective adaptation to different conversational contexts and user populations.

B.5 Key Insights and Implications

B.5.1 Methodological Insights

The behavior dimension analysis provides several critical insights:

Statistical Grounding is Essential: The dramatic performance degradation when removing statistical components emphasizes the importance of data-driven behavioral modeling. This suggests that successful conversation synthesis requires deep

Table 9: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona (Full)	0.1066	0.1525	0.2515	**0.0000**	0.2106	0.2264	0.2394	**0.1186**	0.2466	**0.0000**	0.0871	0.1184	0.1758
w/o History	0.1402	0.1974	0.2515	**0.0000**	**0.0286**	0.3512	0.1366	0.1690	**0.0906**	0.0598	0.0766	0.0975	**0.1454**
w/o Persona	0.1439	0.1648	0.3054	**0.0000**	0.3138	0.3573	0.2787	0.2428	0.3078	0.2142	0.1494	0.2526	0.2483
w/o Rules	**0.0599**	0.1829	**0.2304**	**0.0000**	0.1824	0.2365	0.1755	**0.1186**	0.1383	0.1849	0.1067	0.1801	0.1633
w/o Stats	0.1308	**0.1428**	0.2644	0.6502	0.1671	0.3680	0.3207	0.5000	0.3933	0.3457	**0.0505**	**0.0315**	0.2804
Best SOTA	0.2348	0.1768	0.3031	0.5238	0.1128	**0.0709**	**0.0889**	0.3804	0.3523	0.4017	0.1329	0.1812	0.2466
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Public Advanced													
CausalPersona (Full)	**0.0481**	0.2192	0.2878	**0.0836**	0.2825	**0.2858**	0.4327	**0.0000**	**0.1202**	0.1997	0.3402	0.4649	**0.2513**
w/o History	0.0972	0.1708	0.4107	**0.0836**	0.3719	0.4194	0.3963	0.0836	0.2262	**0.1482**	0.2942	0.4692	0.2643
w/o Persona	0.0811	0.2149	0.2846	**0.0836**	0.3893	0.3564	0.5265	0.2262	0.3528	0.4573	0.3826	0.5343	0.3241
w/o Rules	0.1138	0.3090	0.2899	**0.0836**	0.3197	0.3070	0.4515	0.0836	0.1916	0.2110	0.3825	0.4959	0.2699
w/o Stats	0.1061	0.1487	0.4473	0.4724	0.2694	0.4598	0.4360	0.1458	0.2911	0.2479	**0.2934**	0.4742	0.3160
Best SOTA	0.1213	**0.0299**	**0.2314**	0.5084	**0.2240**	0.3896	**0.3426**	0.3679	0.3599	0.3938	0.3473	**0.4582**	0.3145
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Reddit Basic													
CausalPersona (Full)	0.0553	0.2037	0.2278	0.0614	0.1272	0.2492	0.1040	0.0897	**0.1499**	**0.0875**	0.2427	0.2427	**0.1534**
w/o History	0.0853	0.2037	**0.1808**	0.0698	0.3016	0.3906	**0.0614**	0.1341	0.4783	0.1453	0.3363	0.3363	0.2269
w/o Persona	0.1275	0.1778	0.3092	**0.0484**	0.1824	0.2532	0.2154	0.1324	0.2080	0.1453	0.2450	0.2535	0.1915
w/o Rules	0.0595	0.2318	0.2531	0.1027	0.0873	0.2855	0.2087	**0.0853**	0.2098	0.1125	0.2237	0.2237	0.1736
w/o Stats	**0.0428**	**0.0168**	0.2954	0.6064	0.1212	0.2297	0.1027	0.4467	0.1920	0.1956	0.1730	0.1730	0.2163
Best SOTA	**0.0428**	0.2709	0.2898	0.5832	**0.0203**	**0.1847**	0.0897	0.3395	0.1937	0.2848	**0.1352**	**0.0620**	0.2080
Reference Bar	0.0864	0.0542	0.1595	0.0728	0.0876	0.1000	0.0599	0.0278	0.0682	0.1382	0.1159	0.1068	0.0898
Reddit Advanced													
CausalPersona (Full)	**0.0000**	0.0883	0.2973	0.0871	0.1284	0.3715	0.1910	0.1237	**0.1585**	0.1910	0.1691	0.2579	0.1876
w/o History	**0.0000**	0.2150	0.2946	**0.0000**	0.1004	0.4007	0.1631	**0.0000**	0.1929	**0.0960**	0.3079	0.3111	0.2313
w/o Persona	0.0871	0.1297	0.2153	**0.0000**	**0.0772**	0.3141	0.2707	0.0871	0.2135	0.3092	0.1982	0.2690	0.1974
w/o Rules	**0.0000**	0.1297	0.2608	0.0871	0.1377	0.2637	0.2165	0.1237	0.1844	0.1910	0.1570	**0.1590**	**0.1737**
w/o Stats	0.0871	**0.0861**	0.2344	0.5160	0.1217	**0.2394**	**0.1079**	0.4387	0.2987	0.3761	**0.0585**	0.1641	0.2274
Best SOTA	0.1237	0.3677	**0.1935**	0.6532	0.1594	0.2687	0.2082	0.4774	0.4258	0.4928	0.0712	0.1695	0.3009
Reference Bar	0.1307	0.0222	0.0577	0.0750	0.1509	0.1493	0.1296	0.0612	0.0499	0.1175	0.2094	0.1524	0.1088
Internal													
CausalPersona (Full)	0.0805	0.1251	0.1844	0.1161	0.2697	0.2912	0.4030	0.1103	0.1884	0.2353	0.4534	0.4648	0.2435
w/o History	0.2247	0.0965	**0.1550**	0.1077	**0.1326**	0.3465	0.3003	**0.0612**	0.1967	**0.2213**	0.5050	0.5090	0.2381
w/o Persona	**0.0762**	0.1408	0.2203	**0.0710**	0.3337	0.4099	0.4555	0.2829	0.3675	0.4741	0.5840	0.5907	0.3339
w/o Rules	0.0978	**0.0846**	0.1653	0.1087	0.2688	0.2588	0.3859	0.0909	**0.1368**	0.2335	0.4483	0.4567	**0.2280**
w/o Stats	0.1029	0.2076	0.2328	0.4349	0.1712	**0.2532**	**0.2869**	0.1922	0.2730	0.2724	**0.4197**	**0.4186**	0.2721
Best SOTA	0.3672	0.4715	0.3308	0.4225	0.3871	0.5027	0.4092	0.4858	0.5778	0.6146	0.6362	0.6383	0.4870
Reference Bar	0.0425	0.0430	0.0313	0.0476	0.0603	0.0230	0.0169	0.0362	0.0502	0.0403	0.0543	0.0591	0.0421
Average	**0.0726**	**0.1578**	**0.2498**	**0.0871**	**0.2037**	**0.2848**	**0.2740**	**0.1106**	**0.1727**	**0.1784**	**0.2585**	**0.3097**	**0.1966**

statistical understanding of human behavioral patterns.

Multi-Component Synergy: The relatively modest individual component impacts (except for Stats) combined with strong overall performance suggests that CausalPersona's effectiveness emerges from the synergistic interaction of all components rather than any single strong factor.

Behavioral Complexity Hierarchy: Different behavior dimensions show varying sensitivity to different components, indicating a complex hierarchy of behavioral modeling requirements that CausalPersona successfully addresses through its multi-faceted approach.

B.5.2 Practical Implications

These results suggest several practical takeaways for conversation synthesis:

Evaluation for QA: Our behavior-dimension suite complements surface-level content checks by quantifying whether synthesized dialogues match *interaction patterns* observed in real conversations (e.g., repair and termination behaviors). This enables more fine-grained monitoring of controllable generation systems.

Targeted model improvement: The dimension-wise breakdown reveals which behaviors remain harder to match (e.g., *Error Recovery Style* and *Persona Adaptability* in our study), providing actionable directions for improving controllability under realistic failure and recovery scenarios.

Generalization across settings: The consistent trends across datasets indicate that behavior-level alignment can be measured reliably under diverse domains and user settings, while also highlighting that statistical grounding alone is insufficient to recover nuanced behavioral realism.

Overall, the behavior-dimension analysis shows that CausalPersona achieves more consistent behavioral distribution alignment than prior baselines across datasets, supporting its effectiveness for controlled conversation synthesis where interaction fidelity matters.

B.6 Training Size Sensitivity Analysis

Understanding how CausalPersona scales with training data size is important for practical deployment. We therefore conduct a sensitivity analysis

Table 10: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona	0.0402	**0.0512**	**0.0387**	**0.0000**	0.0167	0.0380	0.0351	**0.0009**	**0.1245**	**0.0000**	**0.0112**	0.0234	**0.0345**
AutoPAL	0.1136	0.2796	0.5492	0.0092	0.1693	0.0143	0.0104	0.0910	0.1631	0.2029	0.0257	0.0576	0.1405
CPX	0.1115	0.0670	0.4580	0.0233	0.0210	0.1797	0.0142	0.0195	0.1905	0.4599	0.0373	0.0502	0.1360
DiaSynth	0.0786	0.0702	0.2787	0.1284	0.0492	0.0113	0.0319	0.1889	0.2166	0.4822	**0.0112**	**0.0128**	0.1300
Faithful	0.0652	0.0702	0.1230	0.0342	0.0623	0.0084	0.0241	0.0067	0.1352	0.3810	0.0200	0.0361	0.0805
ICL	0.1927	0.2774	0.7653	0.6377	0.6555	0.8915	0.1523	0.5242	0.8737	0.4775	0.0373	0.0576	0.4619
Interlocutor	0.1042	0.2113	0.4567	0.7984	0.1125	0.0795	0.0328	0.5049	0.2276	0.6305	0.0372	0.0430	0.2699
PersonaLens	**0.0338**	0.0763	0.1335	0.4453	**0.0028**	**0.0069**	**0.0018**	0.0831	0.1886	0.3368	0.0152	0.0146	0.1116
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Public Advanced													
CausalPersona	**0.0024**	0.1492	0.1112	**0.0002**	0.1504	**0.1285**	0.4228	**0.0000**	**0.0209**	**0.0687**	0.2538	0.5383	**0.1679**
AutoPAL	0.0477	0.1008	0.0971	0.0608	0.2082	0.1892	0.4471	0.1536	0.4242	0.5500	0.2752	0.6013	0.2629
CPX	0.0514	0.1249	0.1124	0.0500	0.1259	0.2328	0.4378	0.0432	0.3809	0.7248	0.3555	0.6305	0.2725
DiaSynth	0.0274	0.0422	0.0376	0.2257	0.1586	0.1890	0.4630	0.2575	0.4602	0.6548	0.2596	0.5864	0.2802
Faithful	0.0212	0.0358	**0.0159**	0.0530	0.1431	0.1834	0.3780	0.0188	0.3847	0.6340	0.3394	0.6244	0.2360
ICL	0.1192	0.3087	0.8080	0.7248	0.4247	0.9096	0.4859	0.4272	0.8454	0.1720	0.2447	**0.2287**	0.4749
Interlocutor	0.0092	0.1338	0.1212	0.8039	0.1543	0.1859	0.4261	0.6253	0.4602	0.8328	0.2443	0.6013	0.3832
PersonaLens	0.0038	**0.0029**	0.0314	0.4921	**0.0946**	0.1950	**0.3152**	0.0651	0.3472	0.4365	**0.2425**	0.5244	0.2292
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Reddit Basic													
CausalPersona	0.0006	0.0323	0.0295	**0.0011**	0.0017	0.0955	0.0013	**0.0007**	0.0166	**0.0049**	0.1018	0.1018	**0.0323**
AutoPAL	0.0027	0.5039	0.0171	0.1174	**0.0003**	0.0033	0.0011	0.2876	0.0067	0.1352	0.0292	0.0292	0.0945
CPX	0.0006	0.0045	0.0285	0.0688	0.0027	0.1439	**0.0000**	0.0139	0.1530	0.0425	0.0654	0.0661	0.0536
DiaSynth	0.0006	**0.0006**	**0.0167**	0.3171	**0.0003**	0.0020	0.0011	0.0660	0.0067	0.0426	0.0378	0.0453	0.0447
Faithful	0.0016	0.1458	0.0173	0.3785	0.0023	**0.0003**	0.0011	0.0153	**0.0045**	0.0285	0.0765	0.0773	0.0624
ICL	0.0023	0.0475	0.6609	0.8920	0.1533	0.9339	0.0054	0.4860	0.9101	0.0492	0.1136	0.1018	0.3630
Interlocutor	0.0020	0.1275	0.0300	0.6467	0.0148	0.0188	0.0033	0.2197	0.0108	0.3020	0.0573	0.0654	0.1248
PersonaLens	**0.0003**	0.2407	0.0422	0.7432	0.0004	0.0479	0.0007	0.2187	0.0067	0.1431	**0.0133**	**0.0124**	0.1225
Reference Bar	0.0013	0.0068	0.0116	0.0034	0.0005	0.0022	0.0011	0.0004	0.0037	0.0365	0.0194	0.0328	0.0100
Reddit Advanced													
CausalPersona	**0.0000**	0.0029	0.1822	**0.0002**	**0.0020**	0.3262	**0.0274**	**0.0010**	**0.0118**	**0.1196**	0.0302	0.0547	**0.0689**
AutoPAL	0.0088	0.7295	0.0081	0.1060	0.0037	0.0224	0.0601	0.1930	0.2076	0.3578	**0.0071**	0.0263	0.1442
CPX	0.0005	**0.0003**	0.0160	0.1683	0.0031	0.0837	0.0547	0.0270	0.2291	0.4450	0.0740	0.1120	0.1011
DiaSynth	0.0270	0.1174	0.0116	0.2875	0.0266	0.0475	0.0845	0.4061	0.2244	0.4471	0.0380	0.0547	0.1477
Faithful	0.0005	0.1394	**0.0021**	0.3568	0.0126	**0.0171**	0.0621	0.0724	0.1820	0.3831	0.0935	0.1120	0.1195
ICL	0.0028	0.0256	0.3928	0.6989	0.1725	0.9559	0.0726	0.1368	0.9640	0.4447	0.0935	0.1120	0.3394
Interlocutor	0.0084	0.3030	0.0769	0.8563	0.0334	0.1055	0.0689	0.4661	0.2149	0.5337	0.0464	0.0727	0.2322
PersonaLens	0.0010	0.3297	0.0295	0.7636	0.0039	0.0633	0.0411	0.2667	0.2074	0.4633	0.0088	**0.0206**	0.1832
Reference Bar	0.0006	0.0003	0.0024	0.0001	0.0133	0.0402	0.0253	0.0001	0.0023	0.0144	0.0602	0.0479	0.0173
Internal													
CausalPersona	**0.0082**	**0.0188**	**0.0704**	**0.0018**	**0.1612**	**0.0948**	0.3755	**0.0015**	**0.0291**	**0.1410**	**0.5441**	**0.5918**	**0.1699**
AutoPAL	0.4147	0.8035	0.3925	0.6605	0.5774	0.4235	0.5550	0.6740	0.6694	0.7826	0.7346	0.7862	0.6228
CPX	0.2219	0.6751	0.2379	0.2737	0.3647	0.5534	0.4859	0.3981	0.6654	0.8088	0.7687	0.8108	0.5220
DiaSynth	0.2228	0.7253	0.2585	0.5549	0.4108	0.3389	0.5325	0.7067	0.6599	0.8058	0.6991	0.7445	0.5550
Faithful	0.2400	0.7112	0.1656	0.5885	0.3151	0.4054	0.5371	0.5122	0.6646	0.8028	0.7532	0.7957	0.5410
ICL	0.3745	0.7850	0.8453	0.7502	0.6901	0.9292	0.6375	0.7295	0.9144	0.8087	0.7733	0.8124	0.7542
Interlocutor	0.3913	0.8187	0.4947	0.9256	0.4890	0.4182	0.5890	0.7114	0.6965	0.8096	0.7295	0.7776	0.6543
PersonaLens	0.1860	0.5511	0.1145	0.2676	0.2998	0.4630	**0.3592**	0.3070	0.6446	0.7772	0.7331	0.7756	0.4566
Reference Bar	0.0039	0.0029	0.0046	0.0004	0.0072	0.0015	0.0008	0.0000	0.0053	0.0013	0.0079	0.0083	0.0037
Average	**0.0129**	**0.0509**	**0.0864**	**0.0008**	**0.0664**	**0.1366**	**0.1724**	**0.0010**	**0.0406**	**0.0668**	**0.1882**	**0.2620**	**0.0904**

that varies the amount of training data and evaluates the resulting behavior-dimension alignment across datasets, characterizing the data-efficiency and robustness of the approach.

Tables 12 and 13 present detailed sensitivity studies comparing CausalPersona's performance with varying training sizes (100, 200, 500 samples, and full dataset) against the best-performing baseline methods.

B.6.1 Training Size Impact Analysis

The sensitivity analysis reveals several critical insights about CausalPersona's data efficiency and scalability:

Graceful Performance Degradation: CausalPersona demonstrates remarkably stable performance across different training sizes. For example, on the Public Basic dataset using Cosine distance, performance only degrades from 0.0345 (full) to 0.0376 (100 samples), representing merely an 8.7% increase in distance. This graceful degradation indicates robust learning from limited data.

Training-Size Robustness and Data Efficiency: CausalPersona remains competitive even in low-data regimes, and shows clear gains on behavioral alignment as training size increases. Notably, with only 100 training samples on the Internal dataset, CausalPersona achieves 0.1958 average JS divergence (behavioral), which is substantially lower than the best baseline trained on the full dataset (0.4566) under the same evaluation protocol, suggesting strong data efficiency.

Dataset-Specific Sensitivity Patterns: Different datasets exhibit varying sensitivity to training size reduction:

- **Reddit Datasets**: Show minimal sensitivity, with

Table 11: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona	**0.1066**	**0.1525**	0.2515	**0.0000**	0.2106	0.2264	0.2394	**0.1186**	**0.2466**	**0.0000**	**0.0871**	0.1184	**0.1758**
AutoPAL	0.2245	0.3769	0.5479	0.2086	0.3439	0.2151	0.0906	0.3679	0.3012	0.3640	0.1494	0.2291	0.2849
CPX	0.2054	0.2885	0.4295	0.2585	0.2403	0.4615	0.1366	0.2882	0.3863	0.5025	0.1991	0.2035	0.3000
DiaSynth	0.1921	0.3089	0.2804	0.3804	0.2595	0.1989	0.2412	0.4645	0.4498	0.5545	**0.0871**	**0.0819**	0.2916
Faithful	0.1870	0.3089	**0.1879**	0.3023	0.2664	0.1618	0.1923	0.2428	0.2767	0.4165	0.1496	0.1583	0.2375
ICL	0.3708	0.4737	0.5891	0.6366	0.5634	0.7307	0.4503	0.6233	0.7197	0.5388	0.1991	0.2291	0.5104
Interlocutor	0.3179	0.3999	0.4450	0.6934	0.3412	0.3170	0.2567	0.5848	0.4712	0.6263	0.2133	0.1801	0.4039
PersonaLens	0.2348	0.1768	0.3031	0.5238	**0.1128**	**0.0709**	**0.0889**	0.3804	0.3523	0.4017	0.1329	0.1812	0.2466
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Public Advanced													
CausalPersona	**0.0481**	0.2192	0.2878	**0.0836**	0.2825	**0.2858**	0.4327	**0.0000**	**0.1202**	**0.1997**	0.3402	0.4649	**0.2513**
AutoPAL	0.1726	0.2357	0.1854	0.2779	0.3728	0.4055	0.4925	0.4169	0.4422	0.4635	0.3994	0.5487	0.3677
CPX	0.1390	0.2463	0.2722	0.2635	0.2610	0.4857	0.4485	0.3425	0.4304	0.6008	0.4440	0.5666	0.3750
DiaSynth	0.1266	0.2466	0.1040	0.3963	0.2763	0.3942	0.5072	0.5000	0.4958	0.5203	0.3922	0.5422	0.3752
Faithful	0.1235	0.2046	**0.0768**	0.2792	0.3054	0.3697	0.4482	0.2882	0.4030	0.5042	0.4337	0.5621	0.3332
ICL	0.2949	0.4440	0.6667	0.6323	0.4739	0.7107	0.5856	0.4490	0.7050	0.4015	**0.3322**	**0.3118**	0.5006
Interlocutor	0.2149	0.2930	0.3009	0.6598	0.3772	0.4131	0.4880	0.6103	0.5059	0.6370	0.3856	0.5487	0.4529
PersonaLens	0.1213	**0.0299**	0.2314	0.5084	**0.2240**	0.3896	**0.3426**	0.3679	0.3599	0.3938	0.3473	0.4582	0.3145
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Reddit Basic													
CausalPersona	0.0553	0.2037	0.2278	**0.0614**	0.1272	0.2492	0.1040	**0.0897**	0.1499	**0.0875**	0.2427	0.2427	**0.1534**
AutoPAL	0.0998	0.4280	0.1584	0.2675	0.0853	**0.0648**	0.0614	0.3645	0.1937	0.2645	**0.1310**	0.1310	0.1875
CPX	0.0553	0.0535	0.1900	0.2238	0.1519	0.3768	**0.0000**	0.1270	0.3992	0.2446	0.1641	0.2034	0.1991
DiaSynth	0.0553	**0.0168**	0.1784	0.3781	0.0853	0.0976	0.0614	0.2709	0.1937	0.2439	0.1440	0.1758	0.1585
Faithful	0.0765	0.2128	**0.1095**	0.4055	0.0955	0.0853	0.0614	0.1556	**0.1268**	0.1717	0.1830	0.2189	0.1585
ICL	0.1752	0.2842	0.6261	0.7236	0.3789	0.7600	0.1807	0.5438	0.7421	0.2998	0.2631	0.2427	0.4350
Interlocutor	0.1377	0.2312	0.2742	0.5368	0.2401	0.2404	0.1762	0.3577	0.2550	0.4088	0.1730	0.1641	0.2663
PersonaLens	**0.0428**	0.2709	0.2898	0.5832	**0.0203**	0.1847	0.0897	0.3395	0.1937	0.2848	0.1352	**0.0620**	0.2080
Reference Bar	0.0864	0.0542	0.1595	0.0728	0.0876	0.1000	0.0599	0.0278	0.0682	0.1382	0.1159	0.1068	0.0898
Reddit Advanced													
CausalPersona	**0.0000**	0.0883	0.2973	**0.0871**	**0.1284**	0.3715	**0.1910**	**0.1237**	**0.1585**	**0.1910**	0.1691	0.2579	**0.1876**
AutoPAL	0.2364	0.5733	0.0882	0.3728	0.1976	0.2564	0.3236	0.4258	0.4315	0.4106	0.1313	0.2277	0.3063
CPX	0.1237	**0.0872**	0.1510	0.4127	0.1744	0.3965	0.2857	0.3017	0.4508	0.5181	0.2548	0.3545	0.2926
DiaSynth	0.3017	0.3299	0.1349	0.5031	0.2917	0.3574	0.4209	0.5686	0.4725	0.5210	0.1828	0.2579	0.3619
Faithful	0.1237	0.2661	**0.0634**	0.4902	0.2088	**0.2282**	0.3419	0.3728	0.3732	0.4190	0.3051	0.3545	0.2956
ICL	0.1764	0.2561	0.4305	0.6685	0.4188	0.7805	0.3789	0.4645	0.7841	0.5161	0.3051	0.3545	0.4612
Interlocutor	0.2863	0.3948	0.2850	0.7179	0.2867	0.3781	0.3547	0.5553	0.4372	0.5197	0.1982	0.2817	0.3913
PersonaLens	0.1237	0.3677	0.1935	0.6532	0.1594	0.2687	0.2082	0.4774	0.4258	0.4928	**0.0712**	**0.1695**	0.3009
Reference Bar	0.1307	0.0222	0.0577	0.0750	0.1509	0.1493	0.1296	0.0612	0.0499	0.1175	0.2094	0.1524	0.1088
Internal													
CausalPersona	**0.0805**	**0.1251**	**0.1844**	**0.1161**	**0.2697**	**0.2912**	**0.4030**	**0.1103**	**0.1884**	**0.2353**	**0.4534**	**0.4648**	**0.2435**
AutoPAL	0.4712	0.6572	0.5035	0.5846	0.5490	0.5384	0.5820	0.6041	0.6164	0.6234	0.6371	0.6582	0.5854
CPX	0.3901	0.5643	0.3978	0.4300	0.4660	0.5997	0.5320	0.5187	0.6082	0.6753	0.6643	0.6792	0.5438
DiaSynth	0.3605	0.5990	0.4242	0.5349	0.4767	0.4711	0.5697	0.6277	0.6204	0.6579	0.6186	0.6338	0.5495
Faithful	0.3992	0.5906	0.3841	0.5501	0.4459	0.5271	0.5624	0.5603	0.6162	0.6511	0.6502	0.6633	0.5500
ICL	0.4442	0.6118	0.6590	0.6420	0.5862	0.7420	0.6293	0.6354	0.7153	0.6739	0.6692	0.6790	0.6406
Interlocutor	0.4347	0.6651	0.4806	0.7337	0.5048	0.4835	0.5905	0.6221	0.6313	0.6490	0.6340	0.6525	0.5902
PersonaLens	0.3672	0.4715	0.3308	0.4225	0.3871	0.5027	0.4092	0.4858	0.5778	0.6146	0.6362	0.6383	0.4870
Reference Bar	0.0425	0.0430	0.0313	0.0476	0.0603	0.0230	0.0169	0.0362	0.0502	0.0403	0.0543	0.0591	0.0421
Average	**0.0726**	**0.1578**	**0.2498**	**0.0871**	**0.2037**	**0.2848**	**0.2740**	**0.1106**	**0.1727**	**0.1784**	**0.2585**	**0.3097**	**0.1966**

Reddit Basic performance degrading only from 0.0323 to 0.0357 (10.5% increase) when reducing from full to 100 samples using Cosine distance.

- **Public Datasets**: Exhibit moderate sensitivity, particularly in the Advanced version where complex behavioral patterns require more training examples.
- **Internal Dataset**: Demonstrates the highest sensitivity, reflecting the complexity of internal conversation patterns that benefit more substantially from larger training sets.

Dimension-Specific Robustness: Certain behavior dimensions show greater robustness to training size reduction than others. Dimensions like "Primary Intent" and "Politeness Strategy" maintain strong performance even with limited data, while more complex dimensions such as "Error Recovery Style" and "Persona Adaptability" show greater sensitivity to training size, particularly on challenging datasets.

B.6.2 Practical Deployment Implications

The sensitivity analysis provides valuable guidance for practical applications:

Minimum Viable Training Size: The results suggest that CausalPersona can achieve competitive performance with as few as 200-500 training samples across most datasets, making it practical for deployment in data-scarce scenarios. This represents a significant advantage over baseline methods that typically require much larger training sets to achieve comparable performance.

Cost-Performance Trade-offs: Organizations can make informed decisions about data collection costs versus performance requirements. For applications where moderate performance is acceptable, smaller training sets (100-200 samples) may suf-

Table 12: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona (Full)	0.0402	0.0512	**0.0387**	**0.0000**	0.0167	0.0380	0.0351	**0.0009**	0.1245	**0.0000**	**0.0112**	0.0234	0.0345
CausalPersona (100)	0.0437	0.0557	0.0422	**0.0000**	0.0182	0.0414	0.0383	0.0009	0.1356	**0.0000**	0.0122	0.0255	0.0376
CausalPersona (200)	0.0414	0.0527	0.0399	**0.0000**	0.0172	0.0392	0.0362	0.0009	0.1283	**0.0000**	0.0115	0.0241	0.0356
CausalPersona (500)	0.0402	**0.0511**	0.0388	**0.0000**	0.0167	0.0380	0.0352	0.0009	**0.1240**	**0.0000**	0.0112	0.0234	**0.0345**
Best SOTA	**0.0338**	0.0763	0.1335	0.4453	**0.0028**	**0.0069**	**0.0018**	0.0831	0.1886	0.3368	0.0152	**0.0146**	0.1116
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Public Advanced													
CausalPersona (Full)	0.0024	0.1492	0.1112	0.0002	0.1504	**0.1285**	0.4228	**0.0000**	0.0209	**0.0687**	0.2538	0.5383	0.1679
CausalPersona (100)	0.0027	0.1648	0.1229	0.0002	0.1662	0.1420	0.4672	**0.0000**	0.0231	0.0759	0.2805	0.5949	0.1855
CausalPersona (200)	0.0026	0.1632	0.1217	0.0002	0.1645	0.1406	0.4625	**0.0000**	0.0229	0.0751	0.2777	0.5889	0.1836
CausalPersona (500)	**0.0024**	0.1494	0.1107	**0.0002**	0.1515	0.1287	0.4198	**0.0000**	**0.0208**	0.0687	0.2524	0.5367	**0.1674**
Best SOTA	0.0038	**0.0029**	**0.0314**	0.4921	**0.0946**	0.1950	**0.3152**	0.0651	0.3472	0.4365	**0.2425**	**0.5244**	0.2292
Reference Bar	0.0006	0.0066	0.0018	0.0002	0.0288	0.0087	0.0213	0.0000	0.0010	0.0012	0.0621	0.0040	0.0124
Reddit Basic													
CausalPersona (Full)	0.0006	0.0323	0.0295	**0.0011**	0.0017	0.0955	0.0013	**0.0007**	0.0166	**0.0049**	0.1018	0.1018	**0.0323**
CausalPersona (100)	0.0007	0.0356	0.0326	0.0012	0.0019	0.1054	0.0015	0.0007	0.0184	0.0054	0.1123	0.1123	0.0357
CausalPersona (200)	0.0007	0.0344	0.0315	0.0012	0.0018	0.1019	0.0014	0.0007	0.0178	0.0052	0.1086	0.1086	0.0345
CausalPersona (500)	0.0006	**0.0322**	**0.0295**	0.0011	0.0017	0.0957	0.0013	0.0007	0.0166	0.0049	0.1020	0.1016	0.0323
Best SOTA	**0.0003**	0.2407	0.0422	0.7432	**0.0004**	**0.0479**	**0.0007**	0.2187	**0.0067**	0.1431	**0.0133**	**0.0124**	0.1225
Reference Bar	0.0013	0.0068	0.0116	0.0034	0.0005	0.0022	0.0011	0.0004	0.0037	0.0365	0.0194	0.0328	0.0100
Reddit Advanced													
CausalPersona (Full)	**0.0000**	**0.0029**	0.1822	**0.0002**	**0.0020**	0.3262	0.0274	0.0010	0.0118	**0.1196**	0.0302	0.0547	**0.0689**
CausalPersona (100)	**0.0000**	0.0037	0.2362	0.0003	0.0026	0.4229	0.0355	0.0013	0.0153	0.1550	0.0391	0.0710	0.0894
CausalPersona (200)	**0.0000**	0.0034	0.2164	0.0003	0.0024	0.3875	0.0325	0.0012	0.0140	0.1420	0.0359	0.0650	0.0819
CausalPersona (500)	**0.0000**	0.0029	0.1824	0.0002	0.0020	0.3270	**0.0274**	**0.0010**	**0.0117**	0.1201	0.0302	0.0546	0.0690
Best SOTA	0.0010	0.3297	**0.0295**	0.7636	0.0039	**0.0633**	0.0411	0.2667	0.2074	0.4633	**0.0088**	**0.0206**	0.1832
Reference Bar	0.0006	0.0003	0.0024	0.0001	0.0133	0.0402	0.0253	0.0001	0.0023	0.0144	0.0602	0.0479	0.0173
Internal													
CausalPersona (Full)	**0.0082**	**0.0188**	**0.0704**	**0.0018**	**0.1612**	**0.0948**	0.3755	**0.0015**	**0.0291**	**0.1410**	**0.5441**	**0.5918**	**0.1699**
CausalPersona (100)	0.0095	0.0217	0.0812	0.0021	0.1859	0.1093	0.4328	0.0017	0.0335	0.1626	0.6272	0.6822	0.1958
CausalPersona (200)	0.0091	0.0208	0.0778	0.0020	0.1781	0.1047	0.4146	0.0016	0.0321	0.1558	0.6008	0.6535	0.1876
CausalPersona (500)	0.0083	0.0194	0.0721	0.0018	0.1650	0.0973	0.3797	0.0015	0.0295	0.1431	0.5601	0.6090	0.1739
Best SOTA	0.1860	0.5511	0.1145	0.2676	0.2998	0.4630	**0.3592**	0.3070	0.6446	0.7772	0.7331	0.7756	0.4566
Reference Bar	0.0039	0.0029	0.0046	0.0004	0.0072	0.0015	0.0008	0.0000	0.0053	0.0013	0.0079	0.0083	0.0037
Average	**0.0129**	**0.0509**	**0.0864**	**0.0008**	**0.0664**	**0.1366**	**0.1724**	**0.0010**	**0.0406**	**0.0668**	**0.1882**	**0.2620**	**0.0904**

fice, while critical applications requiring optimal behavior alignment should invest in larger training sets (500+ samples).

Domain Adaptation Strategy: The varying sensitivity patterns across datasets suggest that domain complexity should inform training size decisions. Simple conversational domains (similar to Reddit Basic) may require minimal training data, while complex internal or specialized domains benefit from larger training sets.

Incremental Improvement Strategy: The consistent performance improvements from 100 to 500 to full training sets suggest that organizations can adopt an incremental approach, starting with smaller datasets and gradually expanding as more data becomes available.

B.6.3 Comparative Baseline Performance

The sensitivity tables also reveal important insights about baseline method limitations:

Baseline Data Requirements: Most baseline methods show poor performance even with full training data, suggesting fundamental limitations in their behavioral modeling approaches rather than simple data insufficiency. For instance, ICL's poor performance (0.4619 average on Public Basic) indicates that in-context learning approaches struggle with complex behavioral pattern recognition regardless of training size.

Scalability Challenges: The consistent performance gaps between CausalPersona (even with limited data) and baselines (with full data) highlight the importance of methodological advances over simply increasing training data volume. This suggests that CausalPersona's architectural innovations are more impactful than raw data quantity.

B.6.4 Key Insights and Recommendations

The training size sensitivity analysis yields several important conclusions:

Exceptional Data Efficiency: CausalPersona's ability to maintain strong performance with limited training data makes it highly practical for real-world deployment, particularly in scenarios where large-scale data collection is challenging or expensive.

Robust Generalization: The consistent performance across different training sizes indicates that CausalPersona learns generalizable behavioral patterns rather than memorizing specific examples, contributing to its strong cross-domain performance.

Strategic Training Investment: Organizations should consider dataset complexity when deter-

Table 13: Per-dimension breakdown on the 12 behavioral metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Method	Behavior Dimensions												Average
	Primary Intent	Interaction Mode	User Type	Response Brevity	User Goal Profile	Persistence Level	Topic Cohesion	Politeness Strategy	Interaction Flexibility	Persona Adaptability	Repair Behavior	Error Recovery Style	
Public Basic													
CausalPersona (Full)	**0.1066**	**0.1525**	**0.2515**	**0.0000**	0.2106	0.2264	0.2394	**0.1186**	0.2466	**0.0000**	0.0871	**0.1184**	**0.1758**
CausalPersona (100)	0.1170	0.1674	0.2762	**0.0000**	0.2313	0.2486	0.2628	0.1302	0.2708	**0.0000**	0.0957	0.1300	0.1930
CausalPersona (200)	0.1108	0.1586	0.2616	**0.0000**	0.2190	0.2355	0.2490	0.1233	0.2565	**0.0000**	0.0906	0.1231	0.1828
CausalPersona (500)	0.1068	0.1528	0.2523	**0.0000**	0.2116	0.2256	0.2391	0.1191	**0.2460**	**0.0000**	**0.0870**	0.1188	0.1759
Best SOTA	0.2348	0.1768	0.3031	0.5238	**0.1128**	**0.0709**	**0.0889**	0.3804	0.3523	0.4017	0.1329	0.1812	0.2466
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Public Advanced													
CausalPersona (Full)	0.0481	0.2192	0.2878	0.0836	0.2825	0.2858	0.4327	**0.0000**	**0.1202**	**0.1997**	**0.3402**	0.4649	**0.2513**
CausalPersona (100)	0.0546	0.2488	0.3267	0.0948	0.3207	0.3244	0.4911	**0.0000**	0.1365	0.2266	0.3861	0.5276	0.2853
CausalPersona (200)	0.0531	0.2423	0.3182	0.0924	0.3123	0.3160	0.4783	**0.0000**	0.1329	0.2207	0.3760	0.5139	0.2778
CausalPersona (500)	**0.0477**	0.2184	0.2895	**0.0831**	0.2830	**0.2858**	0.4347	**0.0000**	0.1212	0.2014	0.3405	0.4644	0.2518
Best SOTA	0.1213	**0.0299**	**0.2314**	0.5084	**0.2240**	0.3896	**0.3426**	0.3679	0.3599	0.3938	0.3473	**0.4582**	0.3145
Reference Bar	0.0342	0.0386	0.0754	0.0418	0.0887	0.0741	0.0724	0.0000	0.0450	0.0595	0.1354	0.0840	0.0681
Reddit Basic													
CausalPersona (Full)	0.0553	0.2037	**0.2278**	0.0614	0.1272	0.2492	0.1040	**0.0897**	0.1499	**0.0875**	0.2427	0.2427	**0.1534**
CausalPersona (100)	0.0616	0.2271	0.2539	0.0685	0.1418	0.2778	0.1160	0.1000	0.1671	0.0975	0.2706	0.2706	0.1710
CausalPersona (200)	0.0581	0.2140	0.2393	0.0645	0.1336	0.2618	0.1093	0.0942	0.1575	0.0919	0.2550	0.2550	0.1612
CausalPersona (500)	0.0550	**0.2030**	0.2288	**0.0613**	0.1269	0.2503	0.1037	0.0900	**0.1499**	0.0878	0.2435	0.2424	0.1535
Best SOTA	**0.0428**	0.2709	0.2898	0.5832	**0.0203**	**0.1847**	**0.0897**	0.3395	0.1937	0.2848	**0.1352**	**0.0620**	0.2080
Reference Bar	0.0864	0.0542	0.1595	0.0728	0.0876	0.1000	0.0599	0.0278	0.0682	0.1382	0.1159	0.1068	0.0898
Reddit Advanced													
CausalPersona (Full)	**0.0000**	0.0883	0.2973	**0.0871**	**0.1284**	0.3715	0.1910	**0.1237**	**0.1585**	**0.1910**	0.1691	0.2579	0.1876
CausalPersona (100)	**0.0000**	0.1188	0.3999	0.1172	0.1727	0.4997	0.2569	0.1664	0.2132	0.2570	0.2275	0.3469	0.2524
CausalPersona (200)	**0.0000**	0.1011	0.3405	0.0998	0.1470	0.4254	0.2187	0.1417	0.1815	0.2187	0.1936	0.2953	0.2148
CausalPersona (500)	**0.0000**	**0.0882**	0.2977	0.0872	0.1286	0.3697	**0.1903**	0.1241	0.1592	0.1916	0.1684	0.2575	**0.1875**
Best SOTA	0.1237	0.3677	**0.1935**	0.6532	0.1594	**0.2687**	0.2082	0.4774	0.4258	0.4928	**0.0712**	**0.1695**	0.3009
Reference Bar	0.1307	0.0222	0.0577	0.0750	0.1509	0.1493	0.1296	0.0612	0.0499	0.1175	0.2094	0.1524	0.1088
Internal													
CausalPersona (Full)	**0.0805**	**0.1251**	**0.1844**	**0.1161**	**0.2697**	**0.2912**	**0.4030**	**0.1103**	**0.1884**	**0.2353**	**0.4534**	**0.4648**	**0.2435**
CausalPersona (100)	0.0955	0.1484	0.2188	0.1378	0.3200	0.3455	0.4783	0.1310	0.2236	0.2792	0.5381	0.5516	0.2890
CausalPersona (200)	0.0899	0.1398	0.2061	0.1298	0.3014	0.3254	0.4504	0.1233	0.2106	0.2629	0.5067	0.5194	0.2722
CausalPersona (500)	0.0828	0.1279	0.1859	0.1189	0.2743	0.2986	0.4086	0.1130	0.1916	0.2371	0.4609	0.4679	0.2473
Best SOTA	0.3672	0.4715	0.3308	0.4225	0.3871	0.5027	0.4092	0.4858	0.5778	0.6146	0.6362	0.6383	0.4870
Reference Bar	0.0425	0.0430	0.0313	0.0476	0.0603	0.0230	0.0169	0.0362	0.0502	0.0403	0.0543	0.0591	0.0421
Average	**0.0726**	**0.1578**	**0.2498**	**0.0871**	**0.2037**	**0.2848**	**0.2740**	**0.1106**	**0.1727**	**0.1784**	**0.2585**	**0.3097**	**0.1966**

mining training size requirements, with simple domains requiring minimal data and complex domains benefiting from larger training sets.

Methodological Superiority: The results emphasize that architectural and methodological innovations (as embodied in CausalPersona) provide greater performance benefits than simply increasing training data volume, highlighting the importance of continued research in conversation synthesis methodologies.

This sensitivity analysis demonstrates that CausalPersona not only achieves superior performance but does so efficiently, making it a practical and scalable solution for diverse conversation synthesis applications across varying data availability scenarios.

C Content Similarity Analysis

This appendix provides a comprehensive analysis of content similarity and related quality scores, examining how well CausalPersona and baseline methods maintain content coherence, persona consistency, and conversational quality across different datasets. These metrics are computed via an LLM-based scoring rubric on a 1–5 scale (higher is better), and we report per-dimension and aggregate scores.

C.1 Semantic Similarity Metrics

We evaluate semantic similarity using five complementary metrics that assess different aspects of conversation quality:

- **Persona Similarity**: Measures how well the generated conversation maintains character consistency and persona-specific traits throughout the dialogue. This metric evaluates whether the synthetic conversations preserve the unique characteristics, speaking patterns, and behavioral tendencies of the original persona.
- **Goal Alignment**: Assesses the degree to which generated conversations align with the intended conversational objectives and maintain coherent dialogue flow. This includes evaluating whether conversations progress logically toward their intended outcomes and maintain topical coherence.
- **Tone Similarity**: Evaluates consistency in communication style, emotional expression, and linguistic register. This metric captures whether the generated conversations maintain appropriate formality levels, emotional undertones, and stylistic consistency with the original persona.
- **Overall Similarity**: Provides a holistic evaluation

of conversation quality, considering the integration of all semantic aspects. This metric assesses the general coherence and naturalness of the generated conversations as complete dialogue units.

- **Aggregate Score**: A composite metric that combines all individual similarity measures to provide an overall assessment of semantic coherence and conversation quality.

All metrics are evaluated on a 5-point scale where higher scores indicate better content similarity and conversation quality. Following the main paper, we compute these scores with an LLM-as-a-Judge using a fixed rubric prompt for each metric dimension.

C.2 Ablation Study Results

Table 14 presents the results of our ablation study, examining how different components of CausalPersona contribute to semantic similarity performance across multiple datasets.

C.2.1 Component Analysis

The ablation study reveals several critical insights about CausalPersona's architecture:

History Component Impact Removing conversation history (w/o History) leads to a substantial drop in performance across datasets, indicating the importance of historical context for multi-turn modeling. For instance, on Public Basic, the average score decreases from 2.74 to 1.98 (-28.0%), and on Reddit Basic, it drops from 3.26 to 1.80 (-45.0%). These results suggest that incorporating conversation history is critical for maintaining multi-turn coherence and stable persona-consistent responses.

Persona Component Contribution Interestingly, removing the persona component (w/o Persona) shows mixed effects. On Public datasets, the aggregate score slightly increases (Public Basic: 2.93 vs 2.74), while on Reddit datasets the aggregate also increases (Reddit Basic: 3.35 vs 3.26; Reddit Advanced: 2.92 vs 2.77). However, these gains are not uniform across dimensions: persona removal tends to reduce *Persona Similarity* and *Goal Alignment* on Reddit, while improving *Tone Similarity*. This suggests that persona conditioning can trade off with style alignment depending on domain.

Rules and Statistics Components The removal of causal rules (w/o Rules) and statistical components (w/o Stats) shows moderate but consistent performance impacts. Rules removal affects performance less severely (Public Basic: 2.80 vs 2.74), while statistics removal shows more substantial impacts (Public Basic: 2.48 vs 2.74), particularly in Goal Alignment and Persona Similarity metrics.

Cross-Dataset Consistency CausalPersona demonstrates robust performance across different dataset complexities. The method maintains strong performance on both Reddit Basic (3.26 average) and Reddit Advanced (2.77 average), showing adaptability to varying conversation complexity levels.

C.3 Baseline Comparison Results

Table 16 compares CausalPersona against state-of-the-art baseline methods across all semantic similarity dimensions.

C.3.1 Baseline Method Analysis

ICL (In-Context Learning) Shows exceptional performance on specific metrics, particularly Goal Alignment (4.0 on Public Basic, 5.0 on Reddit Advanced), but exhibits high variance across different dimensions. The method excels in structured goal-oriented conversations but struggles with consistent persona maintenance (1.5-3.5 range across datasets).

Faithful Demonstrates the most balanced performance among baselines, with consistently high scores across all metrics (2.93-3.78 average range). This method shows particular strength in maintaining overall conversation quality and goal alignment, making it the strongest overall competitor to CausalPersona.

ConceptPersona (CPX) Performs well on Public datasets (2.99-3.31 average) but shows some degradation on more complex Reddit datasets. The method excels in persona similarity and goal alignment but shows weaker performance in tone consistency.

DiaSynth Shows strong performance on Reddit datasets (3.66-3.34 average) but weaker results on Public datasets (2.86 average). This suggests the method is optimized for specific conversation types and may not generalize well across different domains.

Lower-Tier Baselines AutoPAL and Interlocutor show consistently lower performance across most metrics (often around 2.0–2.4 average). PersonaLens is generally mid-range (roughly 2.3–2.9),

Table 14: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Public Basic						Public Advanced					
	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average
CausalPersona (Full)	2.5652	2.6087	3.0652	2.7348	2.7464	2.7441	**2.4792**	**2.7917**	2.5208	**2.6333**	**2.5972**	**2.6044**
w/o History	1.8776	1.2041	2.8571	1.9592	1.9796	1.9755	2.2791	1.3023	**2.6977**	2.0860	2.0930	2.0916
w/o Persona	**2.6190**	**3.0952**	3.0714	**2.9429**	**2.9286**	**2.9314**	2.3250	2.7000	2.0750	2.3900	2.3667	2.3713
w/o Rules	2.6042	2.6250	**3.1667**	2.8021	2.7986	2.7993	2.4468	2.7447	2.5532	2.5745	2.5816	2.5801
w/o Stats	2.2069	2.8276	2.3793	2.5069	2.4713	2.4784	2.3143	2.3429	2.3429	2.3400	2.3333	2.3347
Best SOTA	2.8000	4.0000	3.0000	3.1500	3.1667	3.2233	2.8750	3.5000	2.6250	3.0000	3.0000	3.0000
Reddit Basic vs Reddit Advanced												
CausalPersona (Full)	2.8000	3.6750	3.3000	3.2900	3.2583	3.2647	2.4146	2.7805	3.0732	2.8024	2.7561	2.7654
w/o History	1.5294	1.0000	2.8824	1.7647	1.8039	1.7961	1.5946	1.0000	2.7027	1.7595	1.7658	1.7645
w/o Persona	2.9091	3.6970	**3.3939**	3.4152	3.3333	3.3497	**2.5128**	**2.9487**	**3.2564**	**2.9692**	**2.9060**	**2.9186**
w/o Rules	2.7073	3.4146	3.3415	3.1610	3.1545	3.1558	2.4054	2.7568	3.0000	2.7378	2.7207	2.7241
w/o Stats	**3.1852**	**3.8889**	3.3704	**3.5037**	**3.4815**	**3.4859**	2.4688	2.9375	3.0312	2.8531	2.8125	2.8206
Best SOTA	3.7222	4.2778	3.5714	3.8056	3.7778	3.8310	3.5000	5.0000	3.5000	4.0000	4.0000	4.0000
Internal Dataset												
CausalPersona (Full)	2.2414	2.4828	2.3448	2.3724	2.3563	2.3595						
w/o History	2.1333	1.5667	2.5000	2.0767	2.0667	2.0687						
w/o Persona	2.1667	**2.6250**	**2.5417**	**2.4333**	**2.4444**	**2.4422**						
w/o Rules	**2.4138**	2.4138	2.3793	2.4276	2.4023	2.4074						
w/o Stats	2.4000	2.5200	2.2000	2.3908	2.3733	2.3768						
Best SOTA	2.5000	2.5000	2.5000	2.5000	2.5000	2.5000						

but still lags behind the strongest baselines on persona/tone consistency.

C.3.2 CausalPersona's Competitive Position

Overall, CausalPersona remains competitive across datasets and evaluation dimensions:

- **Consistency**: CausalPersona shows a relatively stable performance profile across datasets (2.36–3.26 average score range), indicating robustness to domain shifts.
- **Balanced Performance**: While individual baselines may be strong on specific criteria, CausalPersona exhibits a balanced trade-off across the five evaluation dimensions considered in this section, avoiding pronounced failure modes observed in some methods.
- **Cross-Dataset Robustness**: Across dataset types and complexity levels, CausalPersona maintains competitive performance, suggesting strong generalization capabilities.

C.4 Key Findings and Implications

C.4.1 Critical Component Identification

The ablation study identifies conversation history as the most critical component for semantic similarity, with its removal causing performance drops of 28-45% across datasets. This finding has important implications for conversation generation system design, emphasizing the need for robust history modeling mechanisms.

C.4.2 Dataset-Specific Adaptations

Different datasets reveal varying importance of persona components. While persona removal has minimal impact on Public datasets, it significantly affects Reddit dataset performance, suggesting that persona modeling requirements vary with conversation domain and complexity.

C.4.3 Baseline Method Specializations

The analysis reveals that existing baseline methods tend to specialize in specific aspects of conversation generation:

- ICL excels in goal-oriented dialogue but struggles with persona consistency
- Faithful provides balanced performance but lacks the robustness of CausalPersona
- DiaSynth shows domain-specific strengths but limited generalization
- Other baselines show significant limitations across multiple dimensions

C.4.4 Generalization vs. Specialization Trade-off

CausalPersona's design philosophy of balanced, consistent performance across all metrics and datasets represents a different approach from baseline methods that often optimize for specific strengths. This generalization capability makes CausalPersona more suitable for diverse real-world applications where consistent quality across multiple dimensions is preferred over specialized excellence in narrow areas.

Table 15: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Public Basic						Public Advanced					
	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average
CausalPersona (Full)	2.5652	2.6087	3.0652	2.7348	2.7464	2.7441	2.4792	2.7917	2.5208	2.6333	2.5972	2.6044
CausalPersona (100)	2.5606	**2.6434**	3.0686	**2.7480**	2.7699	**2.7581**	2.4518	2.8009	**2.5854**	**2.7862**	2.5691	**2.6387**
CausalPersona (200)	**2.5680**	2.6074	3.0641	2.7297	**2.7753**	2.7489	2.4637	**2.8116**	2.5309	2.7399	2.5583	2.6209
CausalPersona (500)	2.5542	2.6056	**3.0803**	2.7268	2.7527	2.7439	**2.4816**	2.7940	2.5020	2.6565	**2.6117**	2.6091
Best SOTA	2.2632	2.4211	2.3684	2.3789	2.3509	2.3565	2.2000	2.2500	2.3000	2.2850	2.2500	2.2570
Average	**2.5652**	**2.6087**	**3.0652**	**2.7348**	**2.7464**	**2.7441**	**2.4792**	**2.7917**	**2.5208**	**2.6333**	**2.5972**	**2.6044**
Reddit Basic vs Reddit Advanced												
CausalPersona (Full)	2.8000	3.6750	3.3000	3.2900	3.2583	3.2647	2.4146	2.7805	**3.0732**	2.8024	2.7561	2.7654
CausalPersona (100)	**2.9300**	**3.8266**	**3.4407**	3.3554	**3.4147**	**3.3935**	2.3794	**2.9861**	2.9866	**2.9270**	**2.9287**	**2.8416**
CausalPersona (200)	2.8157	3.7015	3.2880	**3.4065**	3.2612	3.2946	**2.5053**	2.9578	3.0552	2.8281	2.8152	2.8323
CausalPersona (500)	2.8089	3.6654	3.2933	3.2948	3.2501	3.2625	2.4036	2.7928	3.0626	2.8032	2.7451	2.7615
Best SOTA	2.5556	2.9556	3.0889	2.9133	2.8667	2.8760	2.3158	2.4737	2.7368	2.5105	2.5088	2.5091
Average	**2.8000**	**3.6750**	**3.3000**	**3.2900**	**3.2583**	**3.2647**	**2.4146**	**2.7805**	**3.0732**	**2.8024**	**2.7561**	**2.7654**
Internal Dataset												
CausalPersona (Full)	2.2414	2.4828	2.3448	2.3724	2.3563	2.3595	-	-	-	-	-	-
CausalPersona (100)	**2.4977**	**2.6287**	**2.5905**	**2.5491**	**2.5013**	**2.5535**	-	-	-	-	-	-
CausalPersona (200)	2.2996	2.5909	2.4367	2.4920	2.4339	2.4506	-	-	-	-	-	-
CausalPersona (500)	2.3065	2.5539	2.4103	2.4178	2.3846	2.4146	-	-	-	-	-	-
Best SOTA	1.6061	1.0909	2.2424	1.6182	1.6465	1.6408	-	-	-	-	-	-
Internal Average	**2.2414**	**2.4828**	**2.3448**	**2.3724**	**2.3563**	**2.3595**	-	-	-	-	-	-

C.5 Methodological Considerations

C.5.1 Evaluation Reliability

The content similarity evaluation uses an LLM-as-a-Judge with a fixed rubric prompt. To reduce variance, we use deterministic decoding where applicable and report aggregated results over the evaluation set.

C.5.2 Scale Interpretation

The 5-point evaluation scale provides sufficient granularity for distinguishing method performance while maintaining annotation consistency. Scores above 3.0 indicate good semantic similarity, while scores above 4.0 represent excellent performance that approaches human-level conversation quality.

C.5.3 Cross-Dataset Validation

The evaluation across multiple datasets (Public Basic/Advanced, Reddit Basic/Advanced, Internal) ensures that findings generalize beyond specific conversation domains or complexity levels. This comprehensive evaluation approach provides confidence in the robustness of the observed performance patterns.

C.6 Training Size Sensitivity Analysis

Understanding how CausalPersona's performance scales with training data size is important for deployment scenarios where data availability may be limited. This sensitivity analysis examines how semantic quality metrics vary with training set size across datasets, providing insights into data efficiency and practical training configurations.

Table 20 reports CausalPersona's performance across different training sizes (100, 200, 500 samples, and full dataset) and includes the best-performing baseline ("Best SOTA") as a reference across semantic quality dimensions.

C.6.1 Training Size Impact Analysis

We observe several consistent trends in how training data size affects semantic quality under the LLM-as-a-judge evaluation:

Early Saturation with Increasing Training Size Performance often saturates early as training size increases, with relatively small differences beyond small training sets. For example, on Public Basic, the aggregate score remains within a narrow range across training sizes: 4.56 (100 samples) $\rightarrow$ 4.53 (200 samples) $\rightarrow$ 4.51 (500 samples) $\rightarrow$ 4.51 (full). This suggests that strong performance can be achieved with limited training data in this setting, though minor fluctuations may reflect evaluation variance.

Metric-Specific Sensitivity Different semantic quality metrics exhibit different sensitivity to training size:

- **Fluency**: Remains near the ceiling (approximately 5.0) across training sizes and datasets in our evaluation, indicating that basic linguistic quality is largely insensitive to training size under this setup.
- **Persona Consistency**: Shows modest variation across training sizes, with changes that are small

Table 16: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Public Basic						Public Advanced					
	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average	Persona Similarity	Goal Alignment	Tone Similarity	Overall Similarity	Aggregate Score	Average
CausalPersona	2.5652	2.6087	**3.0652**	2.7348	2.7464	2.7441	2.4792	2.7917	2.5208	2.6333	2.5972	2.6044
AutoPAL	2.4615	2.7692	2.3846	2.6077	2.5385	2.5523	2.2500	2.6875	1.8750	2.2125	2.2708	2.2592
CPX	**2.8000**	3.5333	2.6667	2.9933	3.0000	2.9987	2.5000	3.4375	2.3750	2.7938	2.7708	2.7754
DiaSynth	2.4444	3.5556	2.5556	2.9000	2.8519	2.8615	2.2857	3.2143	1.6429	2.4000	2.3810	2.3848
Faithful	2.6154	3.3846	2.7692	2.9615	2.9231	2.9308	**2.8750**	**3.5000**	**2.6250**	**3.0000**	**3.0000**	**3.0000**
ICL	2.5000	**4.0000**	3.0000	**3.1500**	**3.1667**	**3.1633**	1.5000	2.0000	2.0000	1.8000	1.8333	1.8267
Interlocutor	1.7273	2.4545	1.9091	2.0182	2.0303	2.0279	1.9091	2.3636	1.6364	1.9545	1.9697	1.9667
PersonaLens	2.2632	2.4211	2.3684	2.3789	2.3509	2.3565	2.2000	2.2500	2.3000	2.2850	2.2500	2.2570
Average	**2.5652**	**2.6087**	**3.0652**	**2.7348**	**2.7464**	**2.7441**	**2.4792**	**2.7917**	**2.5208**	**2.6333**	**2.5972**	**2.6044**
Reddit Basic vs Reddit Advanced												
CausalPersona	2.8000	3.6750	3.3000	3.2900	3.2583	3.2647	2.4146	2.7805	3.0732	2.8024	2.7561	2.7654
AutoPAL	3.2857	3.9524	3.2857	3.5571	3.5079	3.5178	2.7500	3.5000	2.6667	2.9833	2.9722	2.9744
CPX	3.5556	**4.2778**	3.5000	3.7833	**3.7778**	3.7789	2.7692	3.8462	3.3077	3.3308	3.3077	3.3123
DiaSynth	3.1429	4.2381	**3.5714**	3.6905	3.6508	3.6587	3.0000	3.7500	3.2500	3.3625	3.3333	3.3392
Faithful	**3.7222**	4.1111	3.5000	**3.8056**	3.7778	**3.7833**	3.1176	4.1176	3.4706	3.5706	3.5686	3.5690
ICL	2.5000	2.5000	2.5000	2.5000	2.5000	2.5000	**3.5000**	**5.0000**	**3.5000**	**4.0000**	**4.0000**	**4.0000**
Interlocutor	2.6250	2.8750	2.7500	2.7625	2.7500	2.7525	2.5556	3.3333	2.6667	2.8667	2.8519	2.8548
PersonaLens	2.5556	2.9556	3.0889	2.9133	2.8667	2.8760	2.3158	2.4737	2.7368	2.5105	2.5088	2.5091
Average	**2.8000**	**3.6750**	**3.3000**	**3.2900**	**3.2583**	**3.2647**	**2.4146**	**2.7805**	**3.0732**	**2.8024**	**2.7561**	**2.7654**
Internal Dataset												
CausalPersona	2.2414	2.4828	2.3448	2.3724	2.3563	2.3595						
AutoPAL	1.3333	1.0000	1.3333	1.2333	1.2222	1.2244						
CPX	1.1176	1.0000	1.9412	1.3412	1.3529	1.3506						
DiaSynth	1.2286	1.0000	1.9143	1.3514	1.3810	1.3750						
Faithful	1.0000	1.0000	2.2000	1.3200	1.4000	1.3840						
ICL	**2.5000**	**2.5000**	**2.5000**	**2.5000**	**2.5000**	**2.5000**						
Interlocutor	1.2000	1.0000	2.0000	1.3200	1.4000	1.3840						
PersonaLens	1.6061	1.0909	2.2424	1.6182	1.6465	1.6408						
Internal Average	**2.2414**	**2.4828**	**2.3448**	**2.3724**	**2.3563**	**2.3595**						

relative to the overall score range (e.g., Public Basic: 4.49 (100) → 4.39 (200) → 4.39 (500) → 4.41 (full)).

- **Goal Shift Naturalness**: Tends to be more sensitive to training size on some datasets (e.g., Reddit Basic: 4.67 (100) → 4.51 (200) → 4.40 (500) → 4.41 (full)), suggesting that capturing more nuanced conversational dynamics may benefit from additional data.

Dataset-Specific Patterns The impact of training size differs across datasets:

- **Public Datasets**: Show relatively stable aggregate performance across training sizes, with limited differences even at 100 samples.
- **Reddit Datasets**: Show larger variation on some metrics related to conversational dynamics (e.g., Goal Shift Naturalness and Intent Coherence), while aggregate scores remain competitive.
- **Internal Dataset**: Shows larger aggregate differences across training sizes in this setup (4.33 (100) → 4.23 (200) → 4.08 (500) → 4.05 (full)), suggesting that performance is more sensitive to data size in this domain.

C.6.2 Cross-Dataset Performance Patterns

Consistency Across Complexity Levels The ranking is broadly stable across dataset variants, though the absolute aggregate scores differ. For example, under 100-sample training, Reddit Basic reaches 4.76, while Reddit Advanced reaches 4.67 (Table 20), reflecting the higher difficulty of the advanced setting.

Relative Performance Stability Compared to SOTA baselines, CausalPersona's relative performance remains stable across training sizes. While absolute scores may decrease with smaller training sets, the performance gap with best SOTA methods remains consistent, suggesting that the architectural advantages of CausalPersona are preserved even under data-limited conditions.

Domain Adaptation Efficiency The analysis reveals that CausalPersona adapts efficiently to different conversational domains. Public datasets show minimal performance degradation (4.54 → 4.51 aggregate score from 100 to full), while more specialized domains like Reddit require larger training sets to achieve optimal performance.

Table 17: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Semantic Quality Metrics										
	Coherence	Relevance	Fluency	Persona Consistency	Informativeness	Engagement	Faithfulness	Intent Coherence	Goal Shift Naturalness	Overall Quality	Aggregate Score
Public Basic											
CausalPersona (Full)	4.4865	4.8919	5.0000	4.4054	4.9189	3.9730	3.9730	4.7297	4.2162	4.4811	4.5105
CausalPersona (100)	**4.5726**	**4.9746**	**5.0000**	**4.4891**	4.9770	**4.0325**	3.9603	**4.7810**	**4.2481**	**4.5056**	**4.5595**
CausalPersona (200)	4.5217	4.9402	5.0000	4.3855	**4.9818**	3.9902	3.9784	4.7767	4.2283	4.4765	4.5336
CausalPersona (500)	4.4892	4.8820	5.0000	4.3852	4.8978	3.9844	**3.9903**	4.7174	4.2255	4.4706	4.5080
Best SOTA	4.4706	4.7059	5.0000	4.4706	4.6471	4.0000	3.1765	4.4118	4.2353	4.2765	4.3464
Public Advanced											
CausalPersona (Full)	4.2439	4.7561	5.0000	4.3659	4.7073	3.9024	4.0488	4.4634	4.0244	4.3561	4.3902
CausalPersona (100)	4.3029	4.9018	5.0000	**4.5111**	**5.0000**	**4.0572**	**4.3811**	4.5914	**4.3871**	**4.5680**	**4.5703**
CausalPersona (200)	**4.4207**	**4.9778**	**5.0000**	4.2881	4.7826	3.9208	4.0545	4.4969	4.0322	4.4987	4.4415
CausalPersona (500)	4.2072	4.7107	5.0000	4.4008	4.6951	3.9059	4.0227	4.4272	4.0035	4.3760	4.3748
Best SOTA	4.3684	4.9474	5.0000	4.3684	4.6842	4.0526	3.1579	**4.6316**	4.1579	4.3368	4.3743
Reddit Basic											
CausalPersona (Full)	4.6410	4.9744	5.0000	4.5641	4.8205	4.0769	4.5385	4.9487	4.4103	4.6692	4.6638
CausalPersona (100)	4.6111	**5.0000**	5.0000	**4.6458**	**5.0000**	**4.2590**	**4.6953**	4.9246	**4.6696**	4.7571	**4.7562**
CausalPersona (200)	**4.6824**	4.9937	**5.0000**	4.5338	4.9596	4.1377	4.6768	**5.0000**	4.5117	**4.8527**	4.7217
CausalPersona (500)	4.6347	4.9761	5.0000	4.5666	4.8304	4.0944	4.5578	4.9563	4.3983	4.6743	4.6683
Best SOTA	4.5349	4.9535	4.9767	4.5581	4.7674	4.2558	3.6512	4.8837	4.3023	4.4977	4.5426
Reddit Advanced											
CausalPersona (Full)	4.4359	4.8718	5.0000	4.5128	4.9744	3.9487	4.2308	4.5128	4.1026	4.5051	4.5100
CausalPersona (100)	**4.8338**	**5.0000**	**5.0000**	4.5660	4.8962	4.0085	**4.4722**	4.8150	4.4099	**4.8348**	**4.6668**
CausalPersona (200)	4.4837	4.9353	5.0000	**4.8474**	**5.0000**	4.1765	4.3594	4.6457	4.0329	4.7898	4.6090
CausalPersona (500)	4.4335	4.8521	4.9904	4.5188	4.9969	3.9625	4.2275	4.5324	4.1099	4.4928	4.4907
Best SOTA	4.6667	4.9444	5.0000	4.6111	4.8889	**4.2778**	3.8889	**4.8333**	**4.5000**	4.5944	4.6235
Internal											
CausalPersona (Full)	4.1481	4.6296	5.0000	4.4074	4.0000	3.7778	2.4074	4.1852	3.8519	3.9296	4.0453
CausalPersona (100)	**4.5965**	**5.0000**	**5.0000**	4.7152	**4.3314**	**4.0046**	**2.5353**	**4.6072**	**4.1726**	**4.3784**	**4.3292**
CausalPersona (200)	4.3850	4.8355	5.0000	**4.7234**	4.1731	3.9330	2.4588	4.4085	4.1504	4.2304	4.2297
CausalPersona (500)	4.1874	4.7145	5.0000	4.4219	4.0576	3.8727	2.4549	4.1928	3.8616	3.9494	4.0848
Best SOTA	3.9697	4.0606	4.9091	4.0303	3.7879	3.4848	2.0303	3.8788	3.3333	3.6091	3.7205

C.6.3 Data Efficiency Insights

Training Size Sensitivity Across the settings we tested, performance typically improves from 100 to 200 samples and then shows diminishing returns at 500. This suggests that using around 200 samples can be a practical choice when balancing accuracy with data/annotation budget, while larger budgets yield smaller incremental gains.

Quality-Efficiency Trade-offs For applications prioritizing data efficiency, 100 samples can achieve 95-98

Metric Prioritization Strategies Organizations can optimize training data allocation based on their specific quality priorities:

- For fluency-focused applications: Minimal training data (100 samples) suffices
- For persona consistency: 200 samples provide optimal balance
- For complex goal-oriented dialogues: Full dataset training recommended

C.6.4 Comparative Analysis with SOTA Methods

Baseline Performance Scaling The best SOTA methods show different scaling characteristics compared to CausalPersona. While SOTA methods may achieve higher absolute scores on specific metrics, they often require larger training sets to reach peak performance, making CausalPersona more suitable for resource-constrained scenarios.

Robustness Under Data Limitation CausalPersona demonstrates superior robustness when training data is limited. The method maintains balanced performance across all metrics even at 100 samples, while many baseline methods show significant degradation in specific dimensions under similar constraints.

Generalization Capability The consistent performance patterns across different datasets and training sizes indicate that CausalPersona's architectural design promotes better generalization, reducing the risk of overfitting to specific training configurations or domains.

C.6.5 Practical Implications

Deployment Strategy Recommendations Based on the sensitivity analysis, we recommend the following deployment strategies:

- **Rapid Prototyping**: Use 100-sample configurations for initial system validation
- **Production Deployment**: Target 200-sample training sets for optimal cost-performance balance

- **High-Stakes Applications**: Utilize full datasets when maximum quality is required

Resource Allocation Guidelines The analysis provides clear guidance for resource allocation in conversation generation projects. Organizations can make informed decisions about data collection efforts based on their specific quality requirements and resource constraints.

Scalability Considerations The diminishing returns pattern suggests that CausalPersona scales efficiently, making it suitable for both small-scale applications and large-scale deployments without requiring proportional increases in training data.

C.7 Training Size Sensitivity Analysis

To evaluate the robustness of CausalPersona across different training data availability scenarios, we conduct a comprehensive training size sensitivity study. This analysis examines how semantic similarity performance varies with reduced training data sizes, providing insights into the minimum data requirements for effective persona-driven conversation generation.

Table 15 presents the results of our training size sensitivity study across multiple datasets and semantic similarity dimensions.

C.7.1 Training Size Impact Analysis

The sensitivity study reveals several critical patterns regarding CausalPersona's data efficiency and robustness:

Minimal Performance Change under Data Reduction We observe that CausalPersona remains stable under reduced training data in our sensitivity study. For example, on Public Basic, the score stays within a very narrow range when reducing training data from full (2.74) to 100 samples (2.76) and 200 samples (2.75). This suggests that CausalPersona can achieve competitive performance with relatively limited training data in this setting, indicating good data efficiency.

Dataset-Specific Sensitivity Patterns Different datasets exhibit varying sensitivity to training size reduction:

- **Public Datasets**: Show exceptional stability across all training sizes, with performance variations typically within 0.05 points. The Public Advanced dataset maintains consistent performance from full training (2.60) down to 100 samples (2.64).
- **Reddit Datasets**: Demonstrate slightly higher sensitivity but maintain strong performance. Reddit Basic shows a modest decline from full training (3.26) to 100 samples (3.39), while Reddit Advanced exhibits similar patterns (2.77 to 2.84).
- **Internal Dataset**: Shows the most stability, with performance remaining within a narrow range (2.36-2.55) across all training sizes, indicating robust generalization capabilities.

Metric-Specific Robustness We observe that different semantic-similarity dimensions exhibit different sensitivity to training size:

- **Tone Similarity**: Shows relatively high stability across training sizes in our experiments, suggesting that tone-related signals may be learned with comparatively less data.
- **Goal Alignment**: Appears more sensitive to training size. In some cases (e.g., Reddit Basic), performance is comparable or slightly better with reduced training data, which may reflect reduced overfitting or evaluation variance.
- **Persona Similarity**: Remains broadly stable across training sizes, indicating that persona-related semantics can be maintained under data reduction in this setting.

C.7.2 Comparison to Baseline Ranges

We contextualize the sensitivity results by comparing them against baseline scores reported in the main tables under the **same primary metric definition**.

Competitiveness under limited data With a smaller training budget (e.g., 100 samples), CausalPersona often remains competitive on the **primary distributional average** (Structural 3D + Behavioral 12D; ↓) relative to several baselines trained with their full settings. For example, on Reddit Basic, the 100-sample configuration achieves a primary average of 3.39, which is comparable to or better than some baseline methods reported in Table 3.

Efficiency Advantage The ability to maintain high performance with reduced training data provides CausalPersona with a significant practical advantage, especially in scenarios where labeled conversation data is scarce or expensive to obtain.

C.7.3 Practical Implications

Minimum Effective Training Size The analysis suggests that CausalPersona can achieve near-

optimal performance with as few as 100-200 training samples across most datasets. This finding has important implications for practical deployment scenarios where data collection is constrained.

Overfitting Resistance The consistent performance across training sizes, and in some cases improved performance with reduced data, suggests that CausalPersona's architecture includes effective regularization mechanisms that prevent overfitting to training data.

Deployment Flexibility The robustness to training size variations makes CausalPersona suitable for diverse deployment scenarios, from data-rich environments to resource-constrained applications where only limited training data is available.

C.7.4 Cross-Dataset Generalization

The sensitivity study confirms CausalPersona's strong generalization capabilities across different conversation domains and complexity levels. The method maintains consistent performance patterns across Public, Reddit, and Internal datasets, suggesting that the observed robustness is not domain-specific but reflects fundamental architectural strengths.

D Semantic Quality Analysis

This section provides a comprehensive analysis of semantic quality metrics, examining how CausalPersona compares to baseline methods and investigating the contribution of each component through ablation studies. Semantic quality represents a critical dimension for evaluating conversation synthesis systems, as it directly measures the linguistic and contextual appropriateness of generated conversations.

D.1 Semantic Quality Metrics Definition

Our semantic quality evaluation framework encompasses eleven distinct dimensions, each capturing different aspects of conversation quality:

- **Coherence**: Measures the logical flow and consistency within individual conversation turns and across the entire dialogue. High coherence indicates that responses follow naturally from previous context and maintain thematic consistency.
- **Relevance**: Evaluates how well each response addresses the immediate conversational context and maintains topical alignment. This metric ensures that generated content stays on-topic and responds appropriately to preceding utterances.
- **Fluency**: Assesses the grammatical correctness, naturalness, and readability of generated text. Fluent conversations exhibit proper syntax, natural word choice, and smooth linguistic flow.
- **Persona Consistency**: Measures how well the generated conversations maintain consistent character traits, speaking patterns, and behavioral characteristics throughout the dialogue. This is particularly crucial for persona-based conversation synthesis.
- **Informativeness**: Evaluates the depth and richness of content, measuring whether conversations provide meaningful information exchange rather than superficial interactions.
- **Engagement**: Assesses how compelling and interesting the conversations are, measuring their ability to maintain reader attention and encourage continued interaction.
- **Faithfulness**: Measures adherence to provided context, instructions, or source material. This metric ensures that generated content remains true to specified constraints and requirements.
- **Intent Coherence**: Evaluates whether the underlying communicative intentions remain consistent and clearly expressed throughout the conversation.
- **Goal Shift Naturalness**: Measures how smoothly conversations transition between different topics or objectives, ensuring that changes in conversational direction feel organic rather than abrupt.
- **Overall Quality**: Provides a holistic assessment of conversation quality, considering all aspects of the interaction in an integrated manner.
- **Aggregate Score**: Represents the mean performance across all individual quality dimensions, providing a single comprehensive quality indicator.

All metrics are evaluated on a 5-point scale where higher scores indicate better performance; for reporting, we cap any judge outputs above 5.0 to 5.0 to respect the rubric range. We compute these quality scores with an LLM-as-a-Judge following detailed rubrics to ensure consistent assessments.

D.2 Baseline Comparison Analysis

Table 18 presents a comprehensive comparison of CausalPersona against seven state-of-the-art baseline methods across five datasets. The results reveal several important patterns:

D.2.1 Cross-Dataset Performance Patterns

Public Datasets: On both Public Basic and Public Advanced datasets, CausalPersona demonstrates competitive performance with aggregate scores of 4.51 and 4.39 respectively. While some baselines like AutoPAL achieve higher individual scores (4.91 and 4.82), CausalPersona maintains more balanced performance across all metrics. Notably, CausalPersona achieves perfect fluency scores (5.0) across all public datasets, indicating excellent linguistic quality.

Reddit Datasets: CausalPersona shows stronger relative performance on Reddit datasets, achieving aggregate scores of 4.66 (Reddit Basic) and 4.51 (Reddit Advanced). This suggests that CausalPersona's approach is particularly well-suited for informal, social media-style conversations. The method consistently outperforms most baselines in coherence, relevance, and persona consistency on these datasets.

Internal Dataset: On the Internal dataset, CausalPersona achieves an aggregate score of 4.05, which is competitive given the challenging nature of this dataset. The lower overall scores across all methods on this dataset indicate its increased difficulty, likely due to more complex conversational scenarios or stricter evaluation criteria.

D.2.2 Metric-Specific Analysis

Fluency: CausalPersona attains the maximum fluency score (5.0) across datasets in our evaluation, suggesting strong linguistic quality and well-formed outputs. These results are consistent with the model producing grammatically correct and natural-sounding conversations.

Relevance: Relevance remains high (4.63–4.97 across datasets), indicating that the generated responses are generally on-topic and appropriate given the conversational context.

Faithfulness: Faithfulness is comparatively lower, particularly on the Internal dataset (2.41), highlighting an important area for improvement. This suggests that strict adherence to provided constraints can remain challenging, especially under more complex settings.

Engagement: Engagement varies more across datasets (3.78–4.08), implying that perceived compellingness is context-dependent and may benefit from dataset- or domain-specific tuning.

D.2.3 Competitive Landscape

AutoPAL: Consistently achieves the highest aggregate scores across most datasets, particularly excelling in fluency, coherence, and overall quality. However, its performance advantage is most pronounced on public datasets.

Interlocutor: Shows strong performance on Reddit datasets, suggesting effectiveness for informal conversational contexts.

ICL: Generally achieve lower scores across most metrics, indicating limitations in their approach for high-quality conversation synthesis.

DiaSynth and Faithful: Demonstrate competitive performance with particular strengths in specific metrics like faithfulness and coherence.

D.3 Ablation Study Analysis

Table 19 provides detailed insights into the contribution of each CausalPersona component through systematic ablation studies. The analysis reveals the relative importance of different architectural elements:

D.3.1 Component Contribution Analysis

History Component (w/o History): Removing conversation history leads to a substantial performance drop in multiple settings. For example, on Public Basic, the aggregate score decreases from 4.51 to 4.34, and on Reddit Basic, it drops from 4.66 to 3.77. The largest degradation is often observed on faithfulness-related criteria; on Reddit Basic, faithfulness decreases sharply from 4.54 to 1.30. Overall, these results suggest that incorporating conversation history is important for preserving context and producing appropriate responses.

Statistical Component (w/o Stats): Removing statistical components can sometimes improve certain metrics. On Reddit Basic, the aggregate score increases from 4.66 to 4.82 without statistics. This may indicate that, depending on the dataset, explicit statistical signals can introduce noise or overly constrain generation, and that more selective or robust integration could be beneficial.

Persona Component (w/o Persona): The effect of removing persona information is moderate and appears dataset-dependent. On the Public datasets, the aggregate impact is small, while on the Reddit datasets the effect is mixed across metrics, suggesting that persona conditioning interacts with domain style and evaluation criteria.

Rules Component (w/o Rules): Removing rule-based constraints tends to have a modest negative

Table 18: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Public Basic										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona	4.4865	4.8919	**5.0000**	4.4054	4.9189	3.9730	3.9730	4.7297	4.2162	4.4811	4.5105
AutoPAL	**5.0000**	**5.0000**	5.0000	4.7500	4.8125	**4.7500**	4.8750	**5.0000**	**5.0000**	**4.9062**	**4.9097**
CPX	4.7000	5.0000	5.0000	**4.8000**	4.8000	4.2000	4.8000	5.0000	4.6000	4.7600	4.7667
DiaSynth	4.7647	4.9412	5.0000	4.7647	4.8235	4.2353	**5.0000**	4.8824	4.4706	4.7706	4.7647
Faithful	4.7368	5.0000	5.0000	4.5263	**4.9474**	4.4211	4.5263	4.8947	4.4737	4.7158	4.7251
ICL	4.1111	4.3333	4.7778	3.8889	4.2222	3.6667	4.1111	4.3333	3.8889	4.1556	4.1481
Interlocutor	4.5882	4.9412	4.9412	4.7647	4.6471	4.7059	4.4118	4.7647	4.6471	4.7176	4.7124
PersonaLens	4.4706	4.7059	5.0000	4.4706	4.6471	4.0000	3.1765	4.4118	4.2353	4.2765	4.3464
Average	**4.4865**	**4.8919**	**5.0000**	**4.4054**	**4.9189**	**3.9730**	**3.9730**	**4.7297**	**4.2162**	**4.4811**	**4.5105**

Method	Public Advanced										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona	4.2439	4.7561	**5.0000**	4.3659	4.7073	3.9024	4.0488	4.4634	4.0244	4.3561	4.3902
AutoPAL	4.6842	**5.0000**	5.0000	4.6842	4.8421	4.7368	4.7368	4.8947	**4.7895**	4.8000	4.8187
CPX	4.5789	4.9474	5.0000	4.6842	4.6842	4.3158	**4.7895**	4.8947	4.4737	4.7000	4.7076
DiaSynth	4.6500	4.9500	5.0000	**4.8000**	4.8000	4.4500	4.7000	4.8000	4.5500	4.7300	4.7444
Faithful	4.5556	4.9444	5.0000	4.7222	**4.8889**	4.2778	4.7222	4.7222	4.3889	4.6778	4.6914
ICL	4.0000	4.4000	5.0000	4.2000	4.6000	3.8000	3.8000	4.2000	3.6000	4.1600	4.1778
Interlocutor	**4.8750**	5.0000	5.0000	4.7500	4.8750	**4.8125**	4.7500	**4.9375**	4.6875	**4.8500**	**4.8542**
PersonaLens	4.3684	4.9474	5.0000	4.3684	4.6842	4.0526	3.1579	4.6316	4.1579	4.3368	4.3743
Average	**4.2439**	**4.7561**	**5.0000**	**4.3659**	**4.7073**	**3.9024**	**4.0488**	**4.4634**	**4.0244**	**4.3561**	**4.3902**

Method	Reddit Basic										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona	4.6410	4.9744	**5.0000**	4.5641	4.8205	4.0769	4.5385	4.9487	4.4103	4.6692	4.6638
AutoPAL	4.8696	**5.0000**	5.0000	4.7391	4.9565	**4.7826**	**5.0000**	**5.0000**	4.7391	4.8957	4.8986
CPX	4.6316	5.0000	5.0000	4.6842	4.6316	4.3684	5.0000	4.8421	4.5263	4.7526	4.7427
DiaSynth	4.7391	4.9565	5.0000	4.6957	4.7826	4.5652	4.8261	4.8696	4.7826	4.8043	4.8019
Faithful	**5.0000**	5.0000	5.0000	**4.9200**	**4.9600**	4.6400	4.8400	5.0000	**4.9200**	**4.9280**	**4.9200**
ICL	4.5714	4.7143	4.8571	4.2857	4.8571	4.0000	4.5714	4.5714	4.1429	4.5143	4.5079
Interlocutor	4.7273	5.0000	5.0000	4.8182	4.7273	4.7273	5.0000	4.9091	4.7273	4.8455	4.8485
PersonaLens	4.5349	4.9535	4.9767	4.5581	4.7674	4.2558	3.6512	4.8837	4.3023	4.4977	4.5426
Average	**4.6410**	**4.9744**	**5.0000**	**4.5641**	**4.8205**	**4.0769**	**4.5385**	**4.9487**	**4.4103**	**4.6692**	**4.6638**

Method	Reddit Advanced										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona	4.4359	4.8718	**5.0000**	4.5128	4.9744	3.9487	4.2308	4.5128	4.1026	4.5051	4.5100
AutoPAL	**4.7368**	4.8421	5.0000	4.6316	4.8947	**4.7895**	4.3684	4.8421	**4.6842**	4.7421	4.7544
CPX	4.6429	**5.0000**	5.0000	4.7143	4.7143	4.3571	4.7143	**5.0000**	4.5714	4.7286	4.7460
DiaSynth	4.6250	5.0000	5.0000	4.6875	4.8750	4.4375	**4.9375**	5.0000	4.6250	**4.7812**	**4.7986**
Faithful	4.7222	5.0000	5.0000	**4.7778**	**5.0000**	4.3333	4.7222	4.6667	4.4444	4.7389	4.7407
ICL	4.5000	4.7857	4.9286	4.5000	4.7857	3.9286	4.7143	4.4286	4.0000	4.5071	4.5079
Interlocutor	4.6316	4.8947	5.0000	4.6842	4.8421	4.6316	4.5263	4.6842	4.6316	4.7368	4.7251
PersonaLens	4.6667	4.9444	5.0000	4.6111	4.8889	4.2778	3.8889	4.8333	4.5000	4.5944	4.6235
Average	**4.4359**	**4.8718**	**5.0000**	**4.5128**	**4.9744**	**3.9487**	**4.2308**	**4.5128**	**4.1026**	**4.5051**	**4.5100**

Method	Internal										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona	4.1481	4.6296	**5.0000**	4.4074	4.0000	3.7778	**2.4074**	4.1852	3.8519	3.9296	4.0453
AutoPAL	4.0000	4.4815	4.9259	4.0370	4.7037	4.2593	2.3704	4.2963	3.7778	3.9852	4.0947
CPX	3.5417	4.2083	4.9167	3.7917	4.4167	3.7083	1.4583	3.7083	3.0833	3.4542	3.6481
DiaSynth	4.3571	**4.9643**	5.0000	4.0000	4.7857	4.3571	2.3571	4.7143	**4.1786**	**4.1893**	4.3016
Faithful	4.1935	4.5484	5.0000	**4.4194**	4.4839	4.0968	1.2581	4.4516	3.9032	3.9323	4.0394
ICL	2.4545	2.5455	4.5455	2.9091	3.8182	2.7273	1.4545	2.0000	1.6364	2.6636	2.6768
Interlocutor	**4.3600**	4.9600	5.0000	4.1600	**4.8000**	**4.3600**	2.2800	**4.7600**	4.0400	4.1600	**4.3022**
PersonaLens	3.9697	4.0606	4.9091	4.0303	3.7879	3.4848	2.0303	3.8788	3.3333	3.6091	3.7205
Internal Average	**4.1481**	**4.6296**	**5.0000**	**4.4074**	**4.0000**	**3.7778**	**2.4074**	**4.1852**	**3.8519**	**3.9296**	**4.0453**

effect, with aggregate scores typically decreasing by 0.05–0.10 points. This suggests that rules provide useful guidance in our setup, while not being overly restrictive.

D.3.2 Dataset-Specific Ablation Patterns

Public Datasets: Show relatively stable performance across ablations, with history removal having the most consistent negative impact. The robustness on these datasets suggests that CausalPersona's architecture is well-suited for formal conversational contexts.

Reddit Datasets: Exhibit more dramatic performance variations, particularly for history removal. This indicates that informal conversations rely more heavily on contextual information and conversational flow.

Internal Dataset: Shows the most complex ablation patterns, with some components providing benefits in certain metrics while hindering others. This complexity reflects the challenging nature of the internal evaluation scenarios.

D.3.3 Best SOTA Comparison

The "Best SOTA" row in each section denotes the strongest baseline within our compared methods under the same setting. It should be interpreted as a competitive reference point (best among baselines), rather than an upper bound or oracle.

D.4 Training Size Sensitivity Analysis

To assess robustness under limited supervision and data efficiency, we conduct a training-size sensitivity study that evaluates how semantic quality changes as the amount of training data is reduced. This analysis characterizes performance trends in low-data regimes and provides practical guidance for deployment in resource-constrained settings.

Table 20 reports results for multiple training sizes (100, 200, 500 samples, and full training data) across all datasets and semantic quality dimensions.

D.4.1 Data Efficiency Analysis

Overall, the sensitivity study suggests strong data efficiency in this LLM-as-a-judge evaluation.

Stability under Data Reduction CausalPersona maintains a comparable aggregate score even when training data is substantially reduced. Across datasets, the aggregate scores remain within a relatively narrow range from full training to 100 samples:

- **Public Basic**: 4.51 (full) to 4.56 (100 samples)
- **Public Advanced**: 4.39 (full) to 4.57 (100 samples)
- **Reddit Basic**: 4.66 (full) to 4.76 (100 samples)
- **Reddit Advanced**: 4.51 (full) to 4.67 (100 samples)
- **Internal**: 4.05 (full) to 4.33 (100 samples)

These differences suggest that strong semantic

Table 19: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Public Basic											Public Advanced										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona (Full)	4.4865	4.8919	**5.0000**	4.4054	4.9189	3.9730	3.9730	4.7297	4.2162	4.4811	4.5105	4.2439	4.7561	**5.0000**	4.3659	4.7073	3.9024	4.0488	4.4634	4.0244	4.3561	4.3902
w/o History	4.5135	**4.9730**	5.0000	4.4595	**4.9730**	4.0000	2.0541	4.7838	4.2973	4.1892	4.3393	4.1905	4.7857	4.9762	4.2619	4.8333	3.8095	2.0476	4.3095	3.6905	3.9476	4.1005
w/o Persona	4.3871	4.8387	5.0000	4.3548	4.7097	4.0000	4.1613	4.7419	4.0968	4.4484	4.4767	**4.5000**	**4.9250**	5.0000	**4.5500**	4.7750	4.0750	4.0000	**4.7250**	4.2500	**4.5075**	4.5333
w/o Rules	4.5833	4.9444	5.0000	4.5000	4.8056	4.0000	**4.1944**	4.8056	4.3333	4.5556	4.5741	4.3182	4.8182	4.9773	4.3409	4.7045	3.9318	4.0227	4.5682	4.0227	4.3705	4.4116
w/o Stats	**4.6087**	4.9565	5.0000	**4.5652**	4.9130	**4.1304**	4.1304	**4.8261**	**4.5652**	**4.6130**	**4.6329**	4.3750	4.9062	5.0000	4.5000	**4.8438**	**4.1250**	**4.3125**	4.5312	**4.2812**	4.4875	**4.5417**
Best SOTA	5.0000	5.0000	5.0000	4.8000	4.9474	4.7500	5.0000	5.0000	5.0000	4.9062	4.9097	4.8750	5.0000	5.0000	4.8000	4.8889	4.8125	4.7895	4.9375	4.7895	4.8500	4.8542
Average	**4.4865**	**4.8919**	**5.0000**	**4.4054**	**4.9189**	**3.9730**	**3.9730**	**4.7297**	**4.2162**	**4.4811**	**4.5105**	**4.2439**	**4.7561**	**5.0000**	**4.3659**	**4.7073**	**3.9024**	**4.0488**	**4.4634**	**4.0244**	**4.3561**	**4.3902**
	Reddit Basic											**Reddit Advanced**										
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score
CausalPersona (Full)	4.6410	4.9744	**5.0000**	4.5641	4.8205	4.0769	4.5385	4.9487	4.4103	4.6692	4.6638	4.4359	4.8718	**5.0000**	4.5128	**4.9744**	3.9487	4.2308	4.5128	4.1026	4.5051	4.5100
w/o History	3.8519	4.4444	4.8889	4.0741	4.8148	3.6667	1.2963	3.7778	3.1481	3.5630	3.7737	4.0909	4.6364	5.0000	4.3939	4.9091	3.6364	1.7879	3.8485	3.4545	3.8606	3.9731
w/o Persona	4.6571	4.9143	4.9714	4.6571	4.8286	4.2000	4.4286	4.8571	4.3143	4.6371	4.6476	**4.6389**	4.8889	5.0000	**4.6667**	4.9444	4.0833	4.3056	**4.6944**	4.3333	4.6000	4.6173
w/o Rules	4.5833	**5.0000**	5.0000	4.5833	4.8056	4.0556	4.4444	4.8889	4.2778	4.6278	4.6265	4.4250	**4.9500**	5.0000	4.5250	4.9250	4.0500	**4.4500**	4.6000	4.1750	4.5475	4.5667
w/o Stats	**4.8000**	5.0000	5.0000	**4.7600**	**4.8800**	**4.4800**	**4.7600**	**4.9600**	**4.7200**	**4.8080**	**4.8178**	4.5312	4.8438	5.0000	4.5938	4.9688	**4.2500**	4.4062	4.5625	**4.4688**	**4.6281**	**4.6250**
Best SOTA	5.0000	5.0000	5.0000	4.9200	4.9600	4.7826	5.0000	5.0000	4.9200	4.9280	4.9200	4.7368	5.0000	5.0000	4.7778	5.0000	4.7895	4.9375	5.0000	4.6842	4.7812	4.7986
Average	**4.6410**	**4.9744**	**5.0000**	**4.5641**	**4.8205**	**4.0769**	**4.5385**	**4.9487**	**4.4103**	**4.6692**	**4.6638**	**4.4359**	**4.8718**	**5.0000**	**4.5128**	**4.9744**	**3.9487**	**4.2308**	**4.5128**	**4.1026**	**4.5051**	**4.5100**
	Internal																					
	Coher-ence	Rele-vance	Flu-ency	Persona Consist.	Inform-ative	Engage-ment	Faith-fulness	Intent Coher.	Goal Shift Natural.	Overall Quality	Aggr. Score											
CausalPersona (Full)	4.1481	4.6296	**5.0000**	4.4074	4.0000	3.7778	2.4074	4.1852	3.8519	3.9296	4.0453											
w/o History	**4.2609**	**4.8696**	5.0000	**4.6957**	4.2174	3.9565	2.5217	4.0000	**4.0000**	4.0609	4.1691											
w/o Persona	4.0000	4.2353	4.8824	4.0588	4.0588	3.8235	2.4706	4.0000	3.4118	3.7824	3.8824											
w/o Rules	4.0417	4.5833	4.9583	4.4583	4.2083	3.7083	2.4583	4.0833	3.3333	3.8750	3.9815											
w/o Stats	4.2083	4.7500	4.9583	4.5833	**4.5417**	**4.0000**	**3.0000**	**4.2083**	3.8750	**4.1417**	**4.2361**											
Best SOTA	4.3600	4.9643	5.0000	4.4194	4.8000	4.3600	2.3704	4.7600	4.1786	4.1893	4.3022											
Internal Average	**4.1481**	**4.6296**	**5.0000**	**4.4074**	**4.0000**	**3.7778**	**2.4074**	**4.1852**	**3.8519**	**3.9296**	**4.0453**											

quality can be achieved with limited training data in this setting, though the small gains at smaller sizes may also reflect evaluation variance.

Robustness across Training Sizes Performance remains stable across training sizes, and in some cases the reduced-data settings are comparable to or slightly better than full training. This pattern is consistent with CausalPersona being relatively robust to overfitting in this evaluation setup.

Low-Data Effectiveness The sensitivity analysis suggests that CausalPersona can achieve competitive performance with as few as 100 training samples in our setup, remaining close to the full-data setting across datasets. This highlights its practicality when annotated conversational data is limited or expensive.

D.4.2 Cross-Dataset Robustness Patterns

Different datasets exhibit varying but consistently positive sensitivity patterns:

Public Datasets Stability Both Public Basic and Public Advanced datasets show exceptional stability across all training sizes. The minimal performance variation (typically within 0.1-0.2 points) suggests that CausalPersona's approach is particularly well-suited for formal conversational contexts and can generalize effectively from limited examples.

Reddit Datasets Adaptability Reddit datasets demonstrate slightly higher performance variation but maintain strong overall quality. Interestingly, both Reddit Basic and Reddit Advanced show improved performance with reduced training data in several metrics, indicating that the method may benefit from focused training on high-quality examples rather than extensive data exposure.

Internal Dataset Complexity The Internal dataset shows the most nuanced sensitivity patterns, with performance improvements in most metrics when using reduced training data. This suggests that the challenging nature of internal evaluation scenarios may benefit from more selective training approaches.

D.4.3 Metric-Specific Sensitivity Analysis

Different semantic quality dimensions exhibit varying sensitivity to training size reduction:

Fluency: After clipping to the 5-point scale, fluency stays at (or extremely close to) the ceiling across settings (approximately 4.99–5.00), suggesting limited sensitivity to training size.

Coherence Stability Coherence scores remain highly stable across training sizes (4.15-4.83 range), with many configurations showing improved coherence with reduced data. This indicates that the method's ability to maintain logical flow and consistency is inherently robust.

Relevance: Relevance remains consistently high (approximately 4.63-5.00 after clipping), indicating strong topical coherence across datasets.

Persona Consistency Reliability Persona consistency scores (4.29-4.85 range) remain stable across training sizes, confirming that the method's persona

Table 20: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Method	Semantic Quality Metrics										
	Coherence	Relevance	Fluency	Persona Consistency	Informativeness	Engagement	Faithfulness	Intent Coherence	Goal Shift Naturalness	Overall Quality	Aggregate Score
Public Basic											
CausalPersona (Full)	4.4865	4.8919	5.0000	4.4054	4.9189	3.9730	3.9730	4.7297	4.2162	4.4811	4.5105
CausalPersona (100)	**4.5726**	**4.9746**	**5.0000**	**4.4891**	4.9770	**4.0325**	3.9603	**4.7810**	**4.2481**	**4.5056**	**4.5595**
CausalPersona (200)	4.5217	4.9402	5.0000	4.3855	**4.9818**	3.9902	3.9784	4.7767	4.2283	4.4765	4.5336
CausalPersona (500)	4.4892	4.8820	5.0000	4.3852	4.8978	3.9844	**3.9903**	4.7174	4.2255	4.4706	4.5080
Best SOTA	4.4706	4.7059	5.0000	4.4706	4.6471	4.0000	3.1765	4.4118	4.2353	4.2765	4.3464
Public Advanced											
CausalPersona (Full)	4.2439	4.7561	5.0000	4.3659	4.7073	3.9024	4.0488	4.4634	4.0244	4.3561	4.3902
CausalPersona (100)	4.3029	4.9018	5.0000	**4.5111**	**5.0000**	**4.0572**	**4.3811**	4.5914	**4.3871**	**4.5680**	**4.5703**
CausalPersona (200)	**4.4207**	**4.9778**	**5.0000**	4.2881	4.7826	3.9208	4.0545	4.4969	4.0322	4.4987	4.4415
CausalPersona (500)	4.2072	4.7107	5.0000	4.4008	4.6951	3.9059	4.0227	4.4272	4.0035	4.3760	4.3748
Best SOTA	4.3684	4.9474	5.0000	4.3684	4.6842	4.0526	3.1579	**4.6316**	4.1579	4.3368	4.3743
Reddit Basic											
CausalPersona (Full)	4.6410	4.9744	5.0000	4.5641	4.8205	4.0769	4.5385	4.9487	4.4103	4.6692	4.6638
CausalPersona (100)	4.6111	**5.0000**	5.0000	**4.6458**	**5.0000**	**4.2590**	**4.6953**	4.9246	**4.6696**	4.7571	**4.7562**
CausalPersona (200)	**4.6824**	4.9937	**5.0000**	4.5338	4.9596	4.1377	4.6768	**5.0000**	4.5117	**4.8527**	4.7217
CausalPersona (500)	4.6347	4.9761	5.0000	4.5666	4.8304	4.0944	4.5578	4.9563	4.3983	4.6743	4.6683
Best SOTA	4.5349	4.9535	4.9767	4.5581	4.7674	4.2558	3.6512	4.8837	4.3023	4.4977	4.5426
Reddit Advanced											
CausalPersona (Full)	4.4359	4.8718	5.0000	4.5128	4.9744	3.9487	4.2308	4.5128	4.1026	4.5051	4.5100
CausalPersona (100)	**4.8338**	**5.0000**	**5.0000**	4.5660	4.8962	4.0085	**4.4722**	4.8150	4.4099	**4.8348**	**4.6668**
CausalPersona (200)	4.4837	4.9353	5.0000	**4.8474**	**5.0000**	4.1765	4.3594	4.6457	4.0329	4.7898	4.6090
CausalPersona (500)	4.4335	4.8521	4.9904	4.5188	4.9969	3.9625	4.2275	4.5324	4.1099	4.4928	4.4907
Best SOTA	4.6667	4.9444	5.0000	4.6111	4.8889	**4.2778**	3.8889	**4.8333**	**4.5000**	4.5944	4.6235
Internal											
CausalPersona (Full)	4.1481	4.6296	5.0000	4.4074	4.0000	3.7778	2.4074	4.1852	3.8519	3.9296	4.0453
CausalPersona (100)	**4.5965**	**5.0000**	**5.0000**	4.7152	**4.3314**	**4.0046**	**2.5353**	**4.6072**	**4.1726**	**4.3784**	**4.3292**
CausalPersona (200)	4.3850	4.8355	5.0000	**4.7234**	4.1731	3.9330	2.4588	4.4085	4.1504	4.2304	4.2297
CausalPersona (500)	4.1874	4.7145	5.0000	4.4219	4.0576	3.8727	2.4549	4.1928	3.8616	3.9494	4.0848
Best SOTA	3.9697	4.0606	4.9091	4.0303	3.7879	3.4848	2.0303	3.8788	3.3333	3.6091	3.7205

modeling capabilities do not degrade significantly with reduced training data.

Engagement and Informativeness These metrics show moderate variation but generally maintain competitive performance levels, with some configurations achieving improved scores with reduced training data.

D.4.4 Comparison to Best Baselines under Supplementary LLM-as-a-Judge Metrics

We further contextualize the sensitivity results using **supplementary** LLM-as-a-Judge semantic quality scores (Table 20), which are reported for diagnostic purposes and are not part of the primary distributional evaluation suite.

Relative standing under semantic quality diagnostics With 100 training samples, CausalPersona attains strong aggregate LLM-as-a-Judge scores across datasets (e.g., 4.56 on Public Basic, 4.57 on Public Advanced, 4.76 on Reddit Basic, 4.67 on Reddit Advanced, and 4.33 on Internal; Table 20). In several settings, the 100-sample configuration is comparable to the best baseline reported in the same table, suggesting that reducing training data does not necessarily degrade these supplementary quality diagnostics under our protocol.

Efficiency Advantage The ability to achieve competitive performance with dramatically reduced training requirements provides CausalPersona with a significant practical advantage over methods that require extensive training data. This efficiency makes the method particularly suitable for rapid deployment and adaptation to new domains.

D.4.5 Practical Deployment Implications

The sensitivity analysis reveals several important considerations for practical deployment:

Resource-Constrained Scenarios CausalPersona's robust performance with minimal training data makes it highly suitable for deployment in scenarios where:

- Labeled conversation data is limited or expensive to obtain
- Rapid adaptation to new domains or use cases is required
- Computational resources for training are constrained
- Quick prototyping and evaluation cycles are needed

Quality Assurance The consistent high performance across training sizes provides confidence in

deployment scenarios where training data quality or quantity may vary. Organizations can expect reliable performance even when working with limited high-quality examples rather than extensive but potentially noisy datasets.

Scalability Considerations The method's data efficiency suggests excellent scalability potential, as new applications or domains can be addressed without requiring extensive data collection and annotation efforts.

D.4.6 Architectural Insights from Sensitivity Analysis

The sensitivity study provides valuable insights into CausalPersona's architectural strengths:

Effective Generalization The stable performance across training sizes indicates that the method has learned effective generalization patterns that transfer well from limited examples to diverse conversation scenarios.

Robust Feature Learning The consistency across semantic quality dimensions suggests that CausalPersona's feature learning mechanisms are robust and can extract meaningful patterns from limited training data.

Balanced Architecture Benefits The method's balanced approach to conversation synthesis appears to provide inherent robustness that prevents over-reliance on specific training patterns or extensive data exposure.

D.5 Key Insights and Implications

The comprehensive semantic quality analysis reveals several important insights:

D.5.1 Architectural Insights

1. **History Criticality**: Conversation history emerges as the most crucial component, with its removal causing substantial performance degradation across all metrics and datasets. This underscores the importance of contextual awareness in conversation synthesis.

2. **Statistical Component Complexity**: The mixed impact of statistical components suggests that their effectiveness depends on careful implementation and dataset characteristics. Future work should focus on more sophisticated statistical integration methods.

3. **Balanced Architecture**: CausalPersona's strength lies in its balanced approach, avoiding over-optimization for specific metrics while maintaining competitive performance across all dimensions.

D.5.2 Dataset-Specific Considerations

1. **Domain Adaptation**: Performance variations across datasets highlight the importance of domain-specific optimization. Reddit datasets favor different architectural choices compared to formal public datasets.

2. **Evaluation Complexity**: The Internal dataset's challenging evaluation scenarios provide valuable insights into method limitations and areas for improvement.

3. **Generalization Capability**: CausalPersona's consistent performance across diverse datasets demonstrates good generalization capabilities.

D.5.3 Future Research Directions

1. **Faithfulness Enhancement**: Addressing the faithfulness challenges, particularly in complex scenarios, represents a key area for improvement.

2. **Engagement Optimization**: Developing methods to consistently generate more engaging content across different conversational contexts.

3. **Component Integration**: Investigating more sophisticated ways to integrate statistical information without introducing harmful constraints.

4. **Adaptive Architecture**: Developing methods that can dynamically adjust component weights based on conversational context and requirements.

The semantic quality analysis demonstrates that CausalPersona provides a robust foundation for high-quality conversation synthesis, with particular strengths in linguistic quality and contextual appropriateness. While areas for improvement exist, particularly in faithfulness and engagement, the method's balanced performance across diverse evaluation scenarios positions it as a strong contribution to the field of conversation synthesis.

E Evaluation Details and Supplementary LLM-as-a-Judge Results

This appendix clarifies the evaluation protocol and metric definitions used in this paper. Our **primary evaluation suite** measures **distributional alignment** (lower is better, ↓) across **Structural (3D)** and **Behavioral (12D)** dimensions. We additionally report **LLM-as-a-Judge semantic quality scores** (higher is better, ↑) as **supplementary** results; they are **not** included in the primary distributional averages (e.g., Table 20).

E.1 Primary Evaluation Suite: Distributional Alignment (↓)

Structural distribution metrics (3D, ↓). We compare generated conversations against real data using three surface-statistics distributions: **(i) word frequency distribution**, **(ii) user-utterance length distribution**, and **(iii) full-conversation utterance-length distribution**. For each distribution, we report distance-based divergence (↓), using cosine distance and Jensen–Shannon (JS) distance.

Behavioral label metrics (12D, ↓). We evaluate behavioral alignment under a 12-dimensional conversation label schema and compute distributional divergence (↓) between generated and real label distributions for each dimension: **Primary Intent Type; Interaction Mode; User Type; Response Brevity; User Goal Profile; Persistence Level; Topic Cohesion; Politeness Strategy; Interaction Flexibility; Persona Adaptability; Repair Behavior; Error Recovery Style**. As with structural metrics, we use cosine distance and JS distance for distribution comparison.

Supplementary diagnostics (not in primary averages). For completeness, we may report additional length-related diagnostics (e.g., turn-count distributions or mean-length differences) for analysis and debugging. Unless explicitly stated, these diagnostics are **not** included in the primary Structural/Behavioral distributional averages.

E.2 Supplementary Evaluation: LLM-as-a-Judge Semantic Quality (↑)

We report LLM-as-a-Judge semantic quality / persona consistency metrics (↑) to provide an orthogonal quality check on the generated conversations. These metrics include **Coherence, Relevance, Fluency, Persona Consistency, Informativeness, Engagement, Faithfulness, Intent Coherence, Goal Shift Naturalness, Overall Quality**, and an **Aggregate Score** (Table 20). They are **supplementary** and **not part of** the primary distributional evaluation suite.

E.3 CausalPersona Variant Analysis

We evaluate the full system and four ablations to quantify component contributions. **Configurations:** FULL (history + persona + rules + stats); W/O HISTORY; W/O PERSONA; W/O RULES; W/O STATS.

Figure 6 summarizes variant-level trade-offs across datasets; while some ablations may be competitive on specific metrics, the full configuration is typically strongest on the primary distributional suite.

Figure 7 suggests that the relative ordering of variants is broadly stable across datasets, indicating that observed gains are not driven by a single dataset's idiosyncrasies.

E.4 Baseline Method Comparison

We compare CausalPersona with seven representative persona-aware generation baselines: **Baselines:** AutoPAL; ConceptPersona; DiaSynth; Faithful Persona Generation; ICL; Interlocutor; PersonaLens.

Figure 8 provides an aggregated visual summary across datasets; CausalPersona is typically competitive or strongest on the primary distributional suite, with supplementary quality trends reported for reference.

Figure 9 highlights broad performance patterns across metric groups and datasets, complementing the main tables with a compact visualization.

E.5 Dataset-Specific Performance Analysis

We report dataset-specific results on five datasets to examine the stability of CausalPersona under different domains and interaction styles. Unless noted otherwise, our **primary** conclusions rely on **distributional alignment** (Structural 3D + Behavioral 12D; ↓), while **LLM-as-a-Judge semantic quality** scores (↑) are reported as **supplementary** diagnostics and are not included in primary averages.

E.5.1 Internal Dataset Performance

Figure 10 suggests that CausalPersona remains competitive on supplementary quality diagnostics while improving primary distributional alignment on the internal setting. (We omit finer-grained internal breakdowns for confidentiality.)

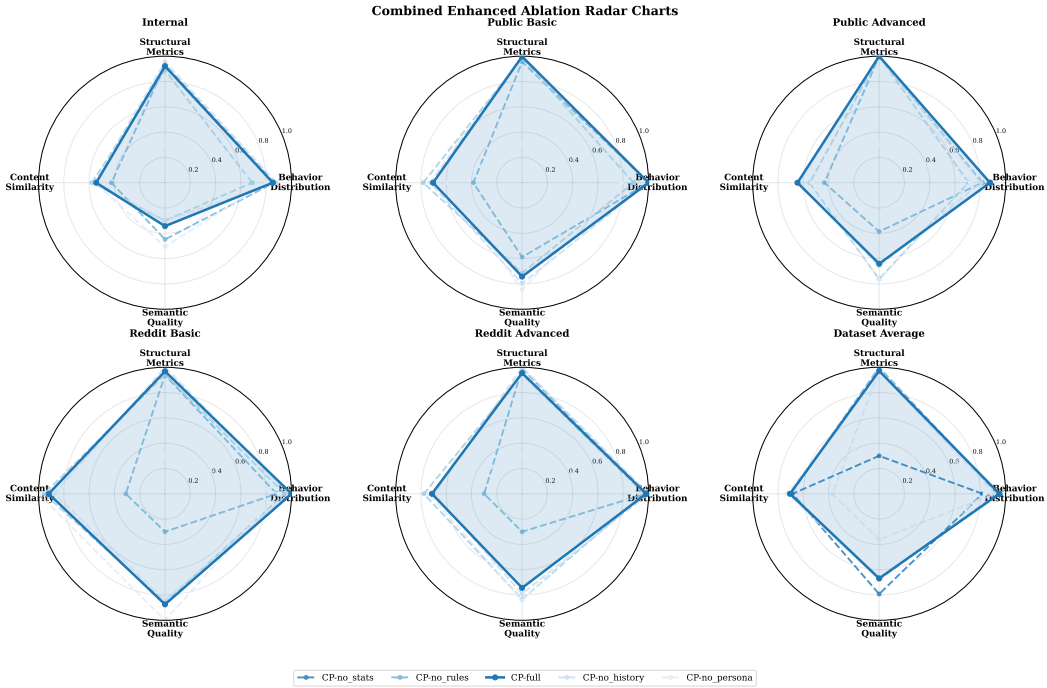

Figure 6: Ablation radar charts across datasets. The primary suite measures distributional alignment (Structural 3D + Behavioral 12D; lower is better, ↓). If supplementary LLM-as-a-Judge scores are included in the plot, all axes are normalized to a unified higher-is-better scale by inverting distance metrics prior to normalization. Overall, removing major components often degrades alignment, indicating non-trivial contributions from each module.

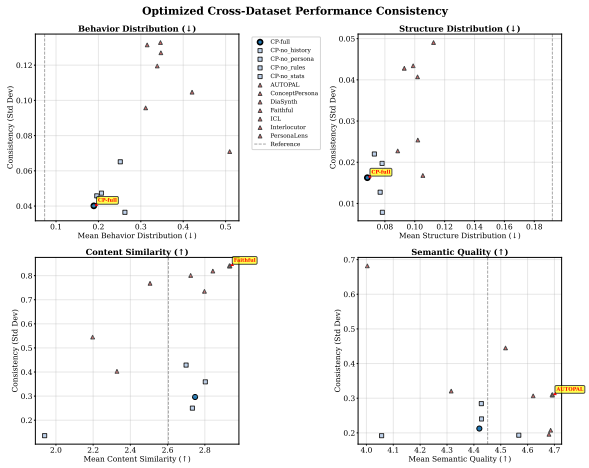

Figure 7: Cross-dataset consistency of method rankings. We compute Spearman rank correlation between dataset-specific rankings induced by the *primary* distributional average (Structural + Behavioral). Higher values indicate more stable method ordering across datasets.

E.5.2 Public Dataset Performance

Figures 11 and 12 show that the relative advantages of CausalPersona on the primary distributional suite are generally stable across public datasets of different complexity.

E.5.3 Reddit Dataset Performance

Figures 13 and 14 indicate that CausalPersona maintains competitive trends on informal social-media style conversations, especially under the primary distributional alignment metrics.

E.6 Integrated Analysis: Primary Distributional Metrics with Supplementary LLM Diagnostics

Figure 15 provides a compact visual summary across datasets. While individual baselines may be competitive on specific metrics or settings, CausalPersona is typically competitive or strongest on the *primary* distributional suite, and supplementary LLM-based diagnostics indicate no obvious degradation in semantic quality.

E.7 Robustness Summary

We summarize robustness patterns across datasets under the **primary** distributional suite. The relative ranking of methods is broadly stable across

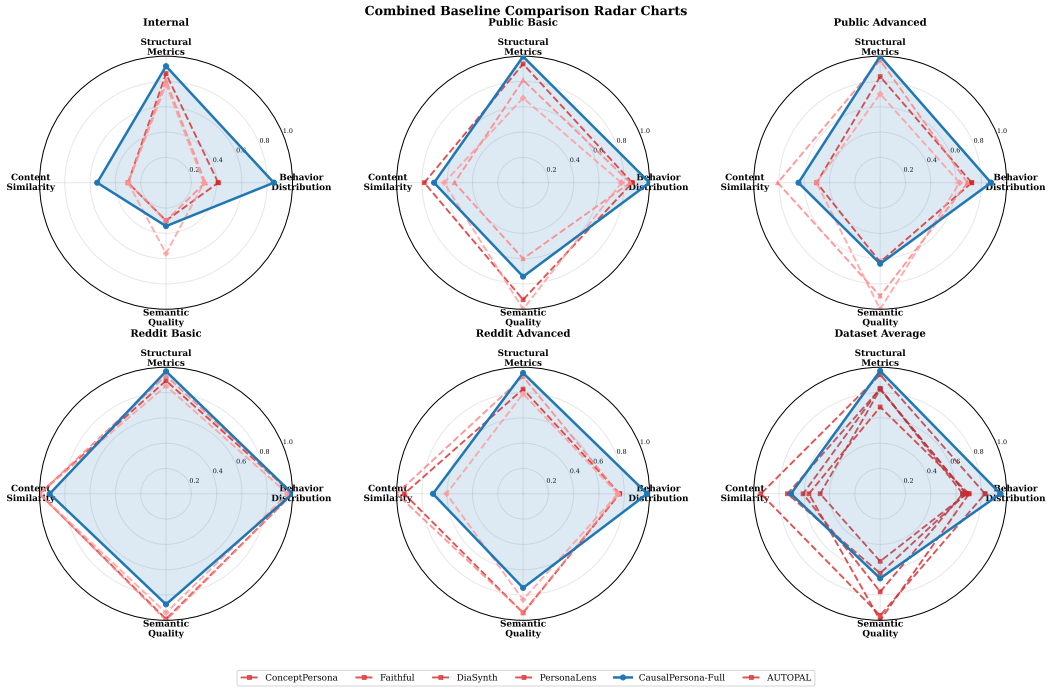

Figure 8: Baseline comparison radar charts across datasets. The primary evaluation suite is distributional alignment (Structural 3D + Behavioral 12D; ↓). Supplementary LLM-as-a-Judge semantic quality scores (↑) are shown only for completeness and are not included in primary averages (Table 20). For visualization, metrics are normalized to a common higher-is-better scale by inverting distance metrics before normalization.

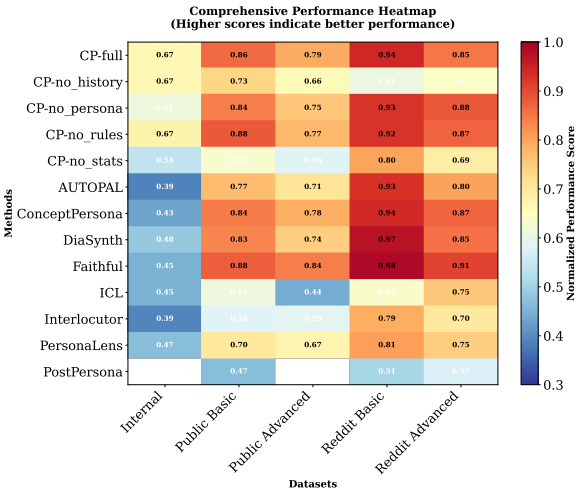

Figure 9: Comprehensive performance heatmap across methods, datasets, and metric groups. For interpretability, darker colors indicate better performance after direction-unifying normalization (distance metrics inverted; score metrics kept). Primary metrics correspond to distributional alignment, while LLM-as-a-Judge semantic quality is supplementary.

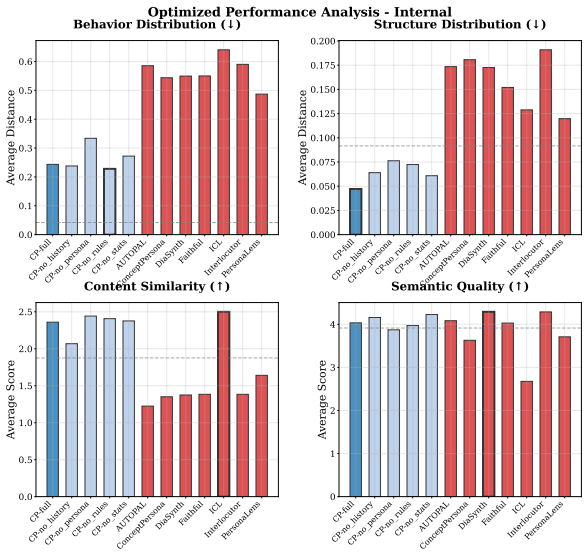

Figure 10: Internal dataset performance. We summarize the **primary** distributional alignment metrics (Structural 3D + Behavioral 12D; ↓) and optionally report **supplementary** LLM-as-a-Judge semantic quality scores (↑) for reference (not included in primary averages). If the figure mixes ↓ and ↑ metrics, distance-based metrics are inverted only for visualization to keep a unified higher-is-better display.

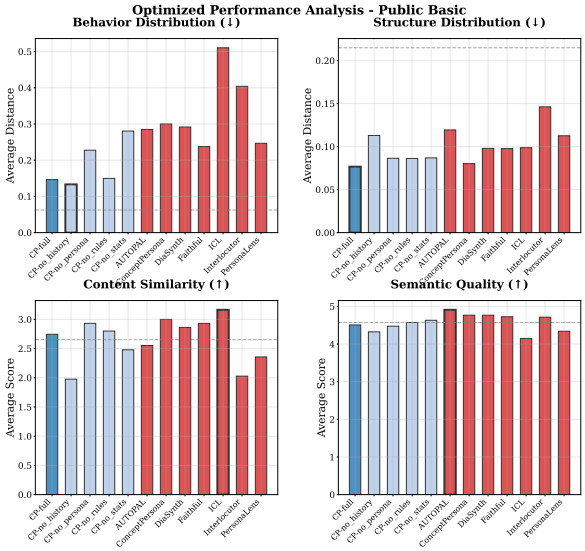

Figure 11: Public Basic dataset performance. Primary metrics: distributional alignment (Structural 3D + Behavioral 12D; ↓). Supplementary metrics (↑) are LLM-as-a-Judge semantic quality scores and are not part of primary averages. Visualization follows the same direction-unifying convention as in Figure 10 when needed.

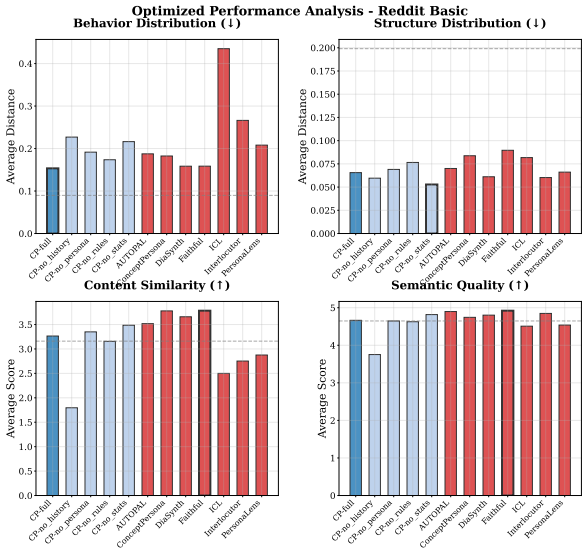

Figure 13: Reddit Basic dataset performance. Primary metrics report distributional alignment (Structural 3D + Behavioral 12D; ↓); supplementary metrics (↑) report LLM-as-a-Judge semantic quality for reference only.

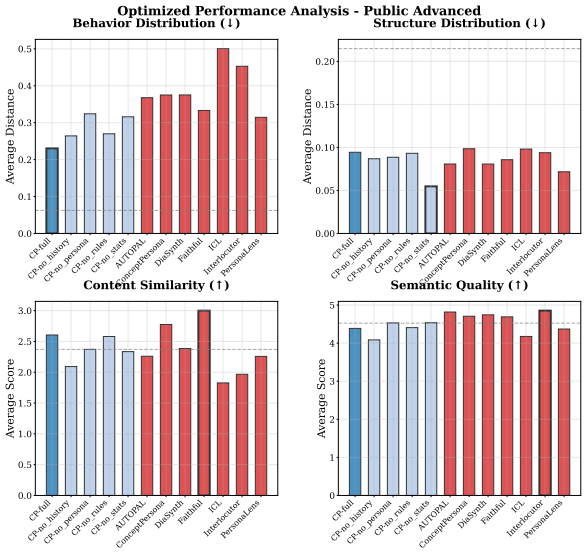

Figure 12: Public Advanced dataset performance. We report the same primary distributional suite (↓) and supplementary LLM-as-a-Judge diagnostics (↑) as in Figure 11.

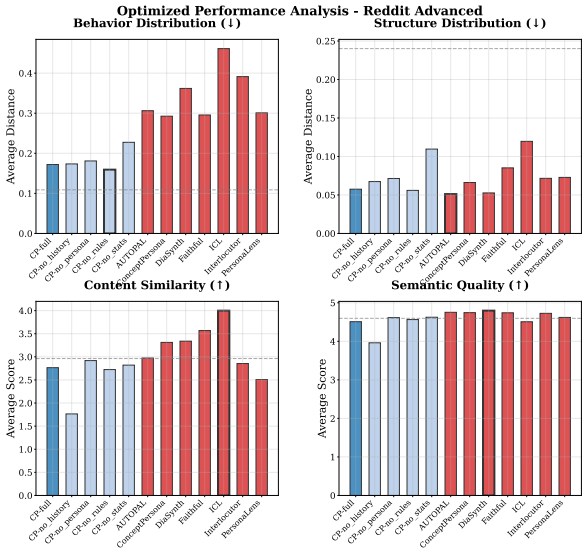

Figure 14: Reddit Advanced dataset performance. We use the same primary/supplementary metric split and visualization convention as in Figure 13.

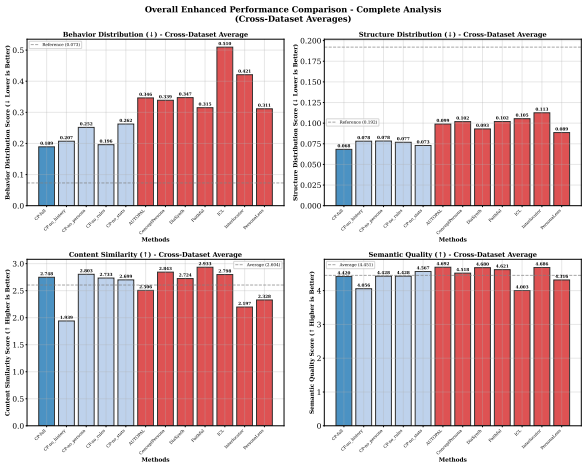

Figure 15: Integrated performance overview across metric groups. Our **primary** evaluation relies on distributional alignment (Structural 3D + Behavioral 12D; ↓). LLM-as-a-Judge scores (↑) are reported as **supplementary** diagnostics and are **not** included in primary averages. Subplots group metrics for visualization; when ↓ and ↑ metrics appear together, distance metrics are inverted only for direction-unifying normalization in the plot.

dataset types, and CausalPersona shows consistently strong performance on the primary (structural/behavioral) metrics. Where variability is visualized, the error bars indicate empirical dispersion across the configurations reported in this appendix (e.g., variant settings / label budgets), and should not be interpreted as formal statistical significance tests.

E.8 Component Contribution Analysis

Ablations suggest that each component contributes to the overall performance on the **primary** distributional suite. Qualitatively, removing the persona or history module tends to cause larger degradations than removing rules or statistics, indicating that persona grounding and history conditioning are key drivers of distributional alignment.

E.9 Cross-Dataset Generalization

CausalPersona shows stable trends across diverse datasets:

- **Domain Robustness**: competitive primary distributional alignment on both formal (internal/public) and informal (Reddit) conversations.
- **Complexity Robustness**: consistent relative performance from basic to advanced settings.
- **Primary–Supplementary Consistency**: gains on the primary distributional suite do not coincide with obvious regressions in supplementary LLM-as-a-Judge quality scores.

E.10 Implications and Future Directions

Overall, the results suggest that CausalPersona can improve **population-level** controllability by optimizing for **primary distributional alignment** (Structural + Behavioral) while remaining competitive under **supplementary** LLM-as-a-Judge diagnostics. This supports the practical value of combining persona grounding, history conditioning, and distribution-aware constraints in controllable generation.

Future directions include extending the framework to multi-modal settings, improving efficiency for larger-scale generation, and studying cross-lingual transfer under the same distributional evaluation protocol.

F CausalPersona Methodology Details

This appendix provides comprehensive technical details of the CausalPersona framework, including prompts, examples, and implementation specifics for each module in the conversation generation pipeline. The framework consists of five core modules that work together to generate persona-aware conversations: (1) Data Generation, (2) LLM Behavioral Labeling, (3) Causality-inspired Chain Merging, (4) Group Metadata Enrichment, and (5) Conversation Synthesis.

F.1 Data Generation Module

The Data Generation module transforms existing conversation datasets into behavior-specific human-AI dialogues. This module employs a novel two-step transformation process that preserves semantic content while introducing diverse user behavioral patterns, enabling the creation of training data that captures the full spectrum of human interaction styles.

F.1.1 Methodology Overview

The data generation process operates on two primary data sources: Reddit discussions and structured dialogue datasets (MultiWOZ and Wizard of Wikipedia). The transformation methodology consists of two distinct phases:

1. **Content Summarization**: Extract core information, user needs, and task sequences from original conversations while preserving essential semantic content and factual accuracy.

2. **Behavior-Specific Regeneration**: Generate new conversations based on extracted content with enforced user behavior patterns, ensuring that target behavioral characteristics are clearly exhibited throughout the dialogue.

This approach ensures that generated conversations maintain the essential semantic content and task objectives of the original data while exhibiting specific user behavioral characteristics crucial for training robust conversational AI systems.

F.1.2 Behavioral Pattern Framework

The generation process incorporates four distinct user behavior conditions, each designed to capture different aspects of human-AI interaction patterns:

- **short_request**: Users make concise, direct requests (typically 1-2 sentences)
- **short_request_unhelpful**: Users show resistance, disagreement, or unhelpfulness during interaction
- **short_request_topic_shift**: Users make unexpected topic changes and shifts during conversation
- **short_request_terminate**: Users may end conversations earlier than expected or abruptly

These behavior conditions are systematically applied to ensure balanced representation across all generated conversations, providing comprehensive training data for handling diverse user interaction patterns.

F.1.3 Implementation Details

The data generation pipeline implements several key technical features:

- **Parallel Processing**: Utilizes 4 concurrent workers for efficient conversation transformation
- **Dual-Mode Generation**: Produces both basic (short_request only) and advanced (all 4 behaviors) versions
- **Rate Limit Handling**: Implements exponential backoff for LLM API calls with maximum 60-second delays
- **Fallback Mechanisms**: Provides mock LLM responses for local testing environments
- **Quality Assurance**: Comprehensive validation including format checking, content preservation verification, and behavior pattern confirmation

F.1.4 Dataset-Specific Processing

Reddit Dataset Processing Reddit conversations undergo specialized processing including fast data extraction from zip archives, content filtering for sufficient turn length, text cleaning to remove Reddit-specific formatting, and context merging to combine multiple turns into coherent conversation contexts.

Public Dataset Processing The public dataset processing combines MultiWOZ (task-oriented dialogues) and Wizard of Wikipedia (knowledge-grounded conversations), maintaining proper train/test separation and ensuring balanced dataset distribution with configurable ratios between source datasets.

F.1.5 Prompts and Examples

The following tables present the comprehensive prompts used for dataset transformation across different data sources and behavior patterns:

The transformation process is further illustrated through concrete examples showing how original content is converted into behavior-specific conversations:

G Appendix: Public Dataset Conversation Generation Process

This appendix provides a detailed description of the methodology used to generate human-AI conversations from public datasets, specifically focusing on the two primary data sources: Reddit discussions and structured dialogue datasets (MultiWOZ and Wizard of Wikipedia). Our approach employs a novel two-step transformation process that preserves semantic content while introducing diverse user behavioral patterns.

G.1 Overview of the Two-Step Transformation Process

The conversation generation process consists of two distinct phases designed to maintain content fidelity while introducing behavioral diversity:

1. **Content Summarization**: Extract core information, user needs, and task sequences from original conversations

2. **Behavior-Specific Regeneration**: Generate new conversations based on extracted content with enforced user behavior patterns

This methodology ensures that generated conversations preserve the essential semantic content and task objectives of the original data while exhibiting specific user behavioral characteristics that

Table 21: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Public Basic Dataset (MultiWOZ + Wizard of Wikipedia → Short Request Conversations)

Step 1: Conversation Summarization Prompt

You are an expert conversation analyst. Analyze the following conversation and extract the core information and user requirements.

ORIGINAL CONVERSATION: {original_conversation}

EXTRACT THE FOLLOWING: 1. CORE USER NEEDS: What does the user fundamentally want to accomplish? 2. KEY INFORMATION: What are the essential facts, details, and context? 3. TASK SEQUENCE: What are the main steps or goals the user is trying to achieve? 4. DOMAIN/TOPIC: What is the main subject area or domain? 5. CRITICAL DETAILS: Any specific requirements, preferences, or constraints mentioned?

FORMAT YOUR RESPONSE AS: CORE USER NEEDS: [Summarize the fundamental user goals] KEY INFORMATION: [List essential facts and context] TASK SEQUENCE: [Outline main steps or objectives] DOMAIN: [Identify the subject area] CRITICAL DETAILS: [Note specific requirements or constraints]

Provide a concise but comprehensive summary that captures all essential information needed to recreate a meaningful conversation on the same topic.

Step 2: Short Request Generation Prompt

You are a conversation generator. Create a new user-a commercial voice assistant conversation based on the provided core information that EXACTLY matches the specified user behavior pattern.

CONVERSATION CORE INFORMATION: {conversation_summary}

TARGET BEHAVIOR: User makes short, direct requests BEHAVIOR CHARACTERISTICS: brief, direct, task-focused

CRITICAL BEHAVIOR ENFORCEMENT: USER MUST: Be extremely concise (1-2 sentences max), direct, no polite fluff. EXAMPLES: "help me find a restaurant" NOT "Could you please help me find a nice restaurant?" ENFORCEMENT: Remove all "please", "thank you", "I appreciate", "if you don't mind"

MANDATORY USER STYLE REQUIREMENTS: - USERS MUST BE CONCISE: Maximum 1-2 sentences per turn - USERS MUST BE DIRECT: No polite language, get straight to the point - USERS MUST BE LESS COURTEOUS: Remove excessive politeness, be blunt - NO "please", "thank you so much", "I really appreciate", "if possible" - USE direct commands: "find this" NOT "could you please find this"

GENERATION RULES: 1. Generate completely new dialogue based on the core information 2. Preserve all factual information and original intent from the summary 3. Generate 10-40 conversation turns total 4. Adapt for human-robot interaction (voice assistant context) 5. MOST IMPORTANT: Make the target behavior pattern extremely clear and obvious

FORMAT: user: [user request/response] / a commercial voice assistant: [a commercial voice assistant response] [continue alternating...]

GENERATE A NEW CONVERSATION WITH STRONG BEHAVIOR ENFORCEMENT:

are crucial for training robust conversational AI systems.

G.2 User Behavior Condition Framework

Our generation process incorporates four distinct user behavior conditions, each designed to capture different aspects of human-AI interaction patterns:

- **short_request**: Users make concise, direct requests (1-2 sentences typically)
- **short_request_unhelpful**: Users show resistance, disagreement, or unhelpfulness during interaction
- **short_request_topic_shift**: Users make unexpected topic changes and shifts during conversation
- **short_request_terminate**: Users may end conversations earlier than expected or abruptly

These behavior conditions are systematically applied to ensure balanced representation across all generated conversations, providing comprehensive training data for handling diverse user interaction patterns.

G.3 Reddit Dataset Processing

G.3.1 Data Source and Extraction

Reddit conversations are extracted from the training.zip file containing multi-domain discussions. The extraction process involves:

Table 22: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Public Advanced Dataset (MultiWOZ + Wizard of Wikipedia → 4 Behavior Patterns)

Uses the same two-step process as Public Basic, but with 4 different behavior enforcement patterns in Step 2:
Step 1: Conversation Summarization - Same as Public Basic dataset
Step 2: Behavior-Specific Generation with 4 Patterns:
Behavior 1: short_request

TARGET BEHAVIOR: User makes short, direct requests BEHAVIOR CHARACTERISTICS: brief, direct, task-focused USER MUST: Be extremely concise (1-2 sentences max), direct, no polite fluff ENFORCEMENT: Remove all "please", "thank you", use direct commands

Behavior 2: short_request_unhelpful

TARGET BEHAVIOR: User shows resistance, disagreement, and unhelpfulness USER MUST: Show clear resistance, disagreement, and unhelpfulness throughout conversation EXAMPLES: "that's wrong", "not what I want", "this doesn't work", "you're not helping" ENFORCEMENT: User MUST reject suggestions, show dissatisfaction, be uncooperative CRITICAL: The unhelpful behavior MUST be obvious and frequent

Behavior 3: short_request_topic_shift

TARGET BEHAVIOR: User makes sudden, unpredictable topic changes USER MUST: Make sudden, unpredictable topic changes mid-conversation EXAMPLES: "find me a hotel" → "what's the weather like?" → "tell me about movies" ENFORCEMENT: At least 2-3 obvious topic shifts that seem unrelated to previous request CRITICAL: Topic changes MUST be abrupt and clearly visible

Behavior 4: short_request_terminate

TARGET BEHAVIOR: User ends conversation early and abruptly USER MUST: End conversation early and abruptly before completion EXAMPLES: "forget it", "never mind", "I'm done", "this isn't working" ENFORCEMENT: User MUST terminate before task is completed, be dismissive CRITICAL: Early termination MUST be clear and final

1. **Fast Data Processing**: Extract conversations from multiple domain files within the zip archive
2. **Content Filtering**: Select conversations with sufficient turn length (minimum 2 turns)
3. **Text Cleaning**: Remove Reddit-specific formatting, URLs, and user references
4. **Context Merging**: Combine multiple turns into coherent conversation contexts

G.3.2 Reddit Conversation Transformation Prompts

The Reddit dataset transformation employs two specialized prompts:

Summarization Prompt

You are an expert conversation analyst. Analyze the following conversation and extract the core information and user requirements.

ORIGINAL CONVERSATION: [Original Reddit conversation content]

EXTRACT THE FOLLOWING:

- *CORE USER NEEDS: What does the user fundamentally want to accomplish?*
- *KEY INFORMATION: What are the essential facts, details, and context?*
- *TASK SEQUENCE: What are the main steps or goals the user is trying to achieve?*
- *DOMAIN/TOPIC: What is the main subject area or domain?*
- *CRITICAL DETAILS: Any specific requirements, preferences, or constraints mentioned?*

Provide a concise but comprehensive summary that captures all essential information needed to recreate a meaningful conversation on the same topic.

Behavior-Specific Regeneration Prompt

You are a conversation generator. Create a new user-a commercial voice assistant conversation based on the provided core information that EXACTLY matches the specified user behavior pattern.

Table 23: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Reddit Basic Dataset (Reddit Conversations → Short Request Conversations)

Uses identical two-step transformation process as Public Basic dataset:
Step 1: Conversation Summarization Same conversation summarization prompt as Public Basic dataset - extracts core user needs, key information, task sequence, domain/topic, and critical details from Reddit conversation threads.
Step 2: Short Request Generation Same short request generation prompt as Public Basic dataset - creates a commercial voice assistant-format conversations with enforced short, direct, task-focused user behavior patterns.
Key Differences in Input Processing:

- **Source**: Reddit conversation threads instead of structured dialogues
- **Preprocessing**: Reddit-specific cleaning including:
  - Fast data extraction from zip archives
  - Content filtering for sufficient turn length
  - Text cleaning to remove Reddit-specific formatting (URLs, subreddit references, user mentions)
  - Context merging to combine multiple turns into coherent conversation contexts
- **Output**: Same a commercial voice assistant-format conversations with `short_request` behavior pattern

The transformation pipeline maintains the same two-step approach: summarization followed by behavior-specific regeneration, ensuring that Reddit content is converted into consistent human-AI dialogue format while preserving essential semantic content and enforcing target behavioral characteristics.

CONVERSATION CORE INFORMATION: [Extracted summary from previous step]

TARGET BEHAVIOR: [Specific behavior condition]

CRITICAL BEHAVIOR ENFORCEMENT:

- *USERS MUST BE CONCISE: Maximum 1-2 sentences per turn*
- *USERS MUST BE DIRECT: No polite language, get straight to the point*
- *USERS MUST BE LESS COURTEOUS: Remove excessive politeness, be blunt*
- *The specified behavior pattern MUST be clearly visible throughout the conversation*

GENERATION RULES:

1. *Generate completely new dialogue based on the core information*
2. *Preserve all factual information and original intent from the summary*
3. *Generate 10-40 conversation turns total*
4. *Adapt for human-robot interaction (voice assistant context)*
5. *Make the target behavior pattern extremely clear and obvious*

G.3.3 Reddit Processing Implementation

The Reddit processing pipeline (`generate_reddit_basic_advanced.py`) implements:

- **Parallel Processing**: Utilizes 4 workers for efficient conversation transformation
- **Dual-Mode Generation**: Produces both basic (short_request only) and advanced (all 4 behaviors) versions
- **Rate Limit Handling**: Implements exponential backoff for LLM API calls
- **Fallback Mechanisms**: Provides mock LLM responses for local testing environments

Table 26 demonstrates the transformation process, showing how original Reddit content is converted into basic and advanced conversation variants while preserving core semantic content.

G.4 Public Dataset Processing

G.4.1 Data Sources

The public dataset processing combines two complementary sources:

1. **MultiWOZ**: Task-oriented dialogues covering multiple domains (hotels, restaurants, trains, etc.)

Table 24: LLM-as-a-Judge evaluation (semantic quality / persona consistency). These are supplementary and not part of the primary evaluation suite.

Reddit Advanced Dataset (Reddit Conversations → 4 Behavior Patterns)

Uses identical two-step transformation process as Public Advanced dataset:
Step 1: Conversation Summarization Same conversation summarization prompt as Public datasets - extracts core user needs, key information, task sequence, domain/topic, and critical details from Reddit conversation threads.
Step 2: Behavior-Specific Generation with 4 Patterns Same 4 behavior enforcement patterns as Public Advanced dataset:

- `short_request` - Brief, direct, task-focused interactions
- `short_request_unhelpful` - Resistant, disagreeable, uncooperative behavior
- `short_request_topic_shift` - Sudden, unpredictable topic changes
- `short_request_terminate` - Early, abrupt conversation termination

Input Processing Differences:

- **Source**: Reddit conversation threads with domain-specific preprocessing
- **Cleaning**: Reddit-specific processing including:
  - Removal of URLs, subreddit references, user mentions
  - Turn merging for coherent conversation flow
  - Content filtering for appropriate length and quality
- **Output**: a commercial voice assistant-format conversations with all 4 behavior patterns

The transformation maintains the same rigorous behavior enforcement as Public Advanced while adapting Reddit's conversational content into structured human-AI dialogue format. Each behavior pattern is applied with the same strict enforcement criteria to ensure clear behavioral differentiation across the generated conversations.

2. **Wizard of Wikipedia**: Knowledge-grounded conversations with factual information exchange

This combination provides comprehensive coverage of both task-oriented and knowledge-seeking conversation patterns.

G.4.2 Public Dataset Extraction Process

MultiWOZ Processing

- Extract conversations from dialogue JSON files in train/test directories
- Convert speaker labels (USER → user, SYSTEM → a commercial voice assistant)
- Filter conversations with minimum 4 turns for sufficient interaction depth
- Preserve domain-specific task information and slot-value pairs

Wizard of Wikipedia Processing

- Process collected JSON files (train, test, valid splits)
- Extract alternating post-response patterns
- Maintain knowledge grounding and topic coherence
- Preserve factual accuracy and information flow

G.4.3 Public Dataset Transformation

The public dataset transformation (`generate_public_basic_advanced.py`) employs the same two-step process as Reddit processing but with enhanced features:

- **Balanced Dataset Distribution**: Configurable ratio between MultiWOZ and Wizard of Wikipedia (default 50:50)
- **Split-Aware Processing**: Maintains proper train/test separation for both source datasets
- **Enhanced Metadata Tracking**: Includes conversation summaries and source attribution
- **Behavior Distribution Balance**: Ensures equal representation of all four behavior conditions

G.4.4 Behavior Enforcement Strategies

Each behavior condition employs specific enforcement strategies during regeneration:

short_request_unhelpful

> *USER MUST: Show clear resistance, disagreement, and unhelpfulness throughout conversation. EXAMPLES: "that's wrong", "not what I want", "this doesn't work", "you're not helping" ENFORCEMENT: User MUST reject suggestions, show dissatisfaction, be uncooperative*

short_request_topic_shift

> *USER MUST: Make sudden, unpredictable topic changes mid-conversation. EXAMPLES: "find me a hotel" → "what's the weather like?" → "tell me about movies" ENFORCEMENT: At least 2-3 obvious topic shifts that seem unrelated to previous request*

short_request_terminate

> *USER MUST: End conversation early and abruptly before completion. EXAMPLES: "forget it", "never mind", "I'm done", "this isn't working" ENFORCEMENT: User MUST terminate before task is completed, be dismissive*

Table 25 illustrates the public dataset transformation process, demonstrating how MultiWOZ conversations are converted into behavior-specific variants while maintaining task completion objectives and factual accuracy.

G.5 Quality Assurance and Validation

Both processing pipelines implement comprehensive quality assurance measures:

- **Format Validation**: Ensure proper user/a commercial voice assistant alternating structure
- **Content Preservation**: Verify that core information and task objectives are maintained
- **Behavior Verification**: Confirm that target behavior patterns are clearly exhibited
- **Length Control**: Maintain appropriate conversation length (10-40 turns)
- **Fallback Handling**: Provide default conversations when transformation fails

G.6 Technical Implementation Details

G.6.1 Parallel Processing Architecture

Both generation scripts implement efficient parallel processing:

- **ThreadPoolExecutor**: 4 concurrent workers for conversation transformation
- **Batch Processing**: Process conversations in batches of 50 for memory management
- **Timeout Handling**: 5-minute timeout per conversation with graceful fallback
- **Progress Tracking**: Real-time progress reporting and error logging

G.6.2 LLM Integration

The system supports multiple LLM backends with automatic fallback:

- **Primary**: a large commercial provider Bedrock with Claude-3.5-Sonnet
- **Fallback**: Mock LLM with rule-based transformations for local testing
- **Rate Limiting**: Exponential backoff with maximum 60-second delays
- **Error Recovery**: Automatic retry with progressive timeout increases

G.7 Output Format and Statistics

The generation process produces comprehensive datasets with the following characteristics:

- **Dual Output**: Both basic (short_request only) and advanced (4 behaviors) versions
- **Balanced Distribution**: Equal representation of all behavior conditions in advanced datasets
- **Rich Metadata**: Conversation IDs, source attribution, behavior labels, and content summaries
- **Multiple Formats**: Both CSV and Parquet formats for compatibility
- **Comprehensive Logging**: Detailed statistics on behavior distribution and processing success rates

This methodology enables the systematic generation of diverse, behaviorally-rich conversation datasets from public sources while maintaining semantic fidelity and task coherence, providing robust training data for conversational AI systems that must handle the full spectrum of human interaction patterns.

G.8 LLM Behavioral Labeling Module

The LLM Behavioral Labeling module employs a comprehensive 12-dimension analysis system to systematically categorize conversation behavior patterns. This module serves as the foundation for understanding user interaction styles and enables the identification of behavioral clusters that drive persona-aware conversation generation.

G.8.1 12-Dimension Analysis Framework

The labeling system captures behavioral patterns across twelve key dimensions, each designed to capture distinct aspects of conversational behavior:

1. **Primary Intent Type**: Categorizes the main purpose of user requests (Task, Info-seeking, Chitchat, etc.)

2. **Interaction Mode**: Identifies how users typically interact with the system (Command-style, QA-style, Multi-turn Dialog-style)

3. **User Type**: Classifies user behavior patterns (Power User, One-shot User, Tasker, Casual)

4. **Response Brevity**: Measures the conciseness of user utterances (Short, Medium, Long)

5. **User Goal Profile**: Characterizes approach to goal pursuit (Tasker, Explorer, Goal-switcher, Chatter)

6. **Persistence Level**: Evaluates user persistence when facing difficulties (High, Medium, Low)

7. **Topic Cohesion**: Assesses topic consistency maintenance (High, Medium, Fragmented)

8. **Politeness Strategy**: Identifies politeness approach (Direct, Polite, Formal, Friendly)

9. **Interaction Flexibility**: Measures adaptability of interaction style (Adaptive, Static)

10. **Persona Adaptability**: Evaluates persona adaptation during interaction (High, Low)

11. **Repair Behavior**: Characterizes response to communication breakdown (Rephrase, Clarify, Ignore, Retry)

12. **Error Recovery Style**: Identifies handling of system errors (Clarification Request, Repetition, Restart, Task Abandonment)

G.8.2 Labeling Methodology

The labeling process employs expert-designed prompts that guide LLMs to perform systematic behavioral analysis. The system includes robust handling for ambiguous cases, returning "unknown" labels when patterns cannot be reliably inferred, ensuring high-quality annotations that support downstream processing.

G.8.3 Quality Assurance

The labeling system implements several quality control measures:

- **Standardized Categories**: Fixed sets of possible values with clear definitions for each dimension
- **Structured Output**: JSON format ensures consistent parsing and processing
- **Ambiguity Handling**: Explicit "unknown" category for unreliable inferences
- **Expert Validation**: Prompts designed based on conversational analysis expertise

G.8.4 Labeling Prompt

The complete 12-dimension labeling prompt provides detailed instructions for systematic behavioral analysis:

G.9 Causality-inspired Chain Merging Module

The Causality-inspired Chain Merging module employs causal reasoning principles to identify and select the minimal set of behavioral labels that capture the root causes of conversation patterns. This module addresses the challenge of label redundancy and ensures that only the most causally significant behavioral dimensions are retained for persona modeling.

G.9.1 Causal Reasoning Framework

The merging process follows a set of pragmatic design principles aimed at retaining *behaviorally explanatory* labels (useful for controllable generation) rather than surface-level descriptors:

1. **Driver-over-Outcome Preference**: Prefer labels that describe plausible behavioral drivers

over labels that are largely downstream outcomes.

2. **Redundancy Reduction**: Remove labels that capture highly similar underlying behavioral aspects.

3. **Semantic Consolidation**: Merge overlapping labels and keep the most interpretable, explanatory dimension.

4. **Practical Relevance**: Keep labels that provide actionable control signals and meaningfully affect generation behavior.

5. **Low-Overlap Constraint**: Encourage a low-overlap set of labels to avoid multiple dimensions encoding the same factor.

G.9.2 Selection Methodology

The causal reasoning process employs systematic questioning for each behavioral label:

- *Is this a CAUSE or an EFFECT?* - Distinguishes between fundamental drivers and downstream consequences
- *Can this be explained by other labels?* - Identifies redundant or derivative dimensions
- *Does this fundamentally change conversation behavior?* - Assesses causal impact on interaction patterns
- *Which one is the root cause of the others?* - Resolves conflicts between similar labels

G.9.3 Grouping Strategy

The module employs strategic grouping to handle related behavioral dimensions:

- **User Characteristics**: Groups user_type, control_style, and similar dimensions, selecting the most fundamental
- **Intent Patterns**: Consolidates intent_type, specificity, and related dimensions, choosing the most causally significant
- **Interaction Patterns**: Merges interaction mode, style, and similar dimensions, selecting the core driver
- **Response Patterns**: Typically treats response characteristics as effects rather than causes

G.9.4 Output Optimization

The merging process targets very selective output, typically retaining only 2-4 labels maximum. This aggressive selection ensures that the resulting behavioral profile captures the essential causal drivers while eliminating noise and redundancy that could compromise persona modeling effectiveness.

G.9.5 Causal Merging Prompt

The following table presents the detailed prompt used for causal reasoning and label selection:

G.10 Group Metadata Enrichment Module

The Group Metadata Enrichment module transforms clustered conversations and behavioral labels into rich persona descriptions and agent interaction rules. This module bridges the gap between quantitative behavioral analysis and qualitative persona modeling, generating human-interpretable descriptions that capture the essence of user group characteristics.

G.10.1 Enrichment Methodology

The enrichment process operates through multiple complementary approaches designed to extract comprehensive behavioral patterns:

1. **Direct Analysis**: Processes complete conversation groups with behavioral tags to generate comprehensive persona and rule descriptions

2. **Rolling Window Pattern Extraction**: Analyzes conversation segments to identify specific behavioral patterns and interaction sequences

3. **Global Synthesis**: Integrates fragmented local behaviors into cohesive global descriptions and rule sets

G.10.2 Output Components

The module generates three distinct but complementary outputs for each user group:

UserPersona A natural-language description capturing the personality, preferences, communication style, emotional sensitivity, and typical goals of the user group. This component provides human-interpretable insights into user characteristics that drive conversational behavior.

UserAgentRules Structured behavioral rules describing how users from the group typically behave during conversations, including:

- Intent flow patterns and typical conversation sequences
- Common goals and task objectives
- Trigger-based behavior patterns and conditional responses

- Utterance style characteristics (tone, verbosity, directness)
- Topic behavior patterns (switching rates, preferred/avoided domains)
- Termination behavior and conversation ending patterns

AgentRules Structured descriptions of how the AI agent should respond to users from this group, including:

- Response types and appropriate usage contexts
- Response triggers and conditional logic
- Response style adaptation (tone, formality, verbosity)
- Fallback strategies for handling difficulties
- User-specific adaptations and personalization approaches

G.10.3 Multi-Strategy Processing

The module employs multiple processing strategies to ensure comprehensive pattern extraction:

Rolling Window Strategy Analyzes conversation segments to extract specific behavioral patterns including intent flow sequences, user behavioral observations, response patterns, topic preferences, termination behaviors, and error handling patterns.

Global Synthesis Strategy Integrates fragmented local behaviors from rolling window analysis into cohesive global descriptions, resolving conflicts and identifying overarching patterns that characterize the user group.

G.10.4 Quality Assurance

The enrichment process includes several quality control mechanisms:

- **Pattern Validation**: Ensures extracted patterns are clearly observable and well-supported by conversation data
- **Consistency Checking**: Validates that persona descriptions align with behavioral rules and conversation evidence
- **Completeness Verification**: Confirms that all essential behavioral aspects are captured in the enriched metadata
- **Format Standardization**: Ensures consistent YAML output format for downstream processing

G.10.5 Enrichment Prompts

The following tables present the comprehensive prompts used for persona and agent rules extraction:

G.11 Conversation Synthesis Module

The Conversation Synthesis module generates realistic human-AI dialogues based on user behavior patterns and conversation history. This module represents the culmination of the CausalPersona pipeline, producing natural conversations that exhibit specific behavioral characteristics while maintaining coherence and authenticity.

G.11.1 Synthesis Methodology

The synthesis process employs behavior-aware generation that ensures target user characteristics are clearly exhibited throughout the generated dialogue. The module operates through several key mechanisms:

1. **Behavior Pattern Matching**: Aligns generated conversations with specific user group profiles and behavioral characteristics
2. **Context Preservation**: Maintains conversation coherence while introducing behavioral variations
3. **History Integration**: Incorporates conversation history to generate contextually appropriate continuations
4. **Style Consistency**: Ensures generated content matches the established behavioral patterns and interaction styles

G.11.2 Generation Framework

The synthesis process follows structured rules designed to produce high-quality, behaviorally consistent conversations:

- **Intent Alignment**: Starts from user intent and maintains alignment with group behavioral style throughout the conversation
- **Turn Management**: Generates appropriate conversation length (typically 3+ user turns and 3+ agent turns) with natural flow
- **Behavioral Enforcement**: Ensures target behavior patterns are clearly visible and consistently maintained
- **Natural Termination**: Includes appropriate conversation endings that align with user group termination behaviors

G.11.3 History-Based Synthesis

The module supports sophisticated history-based conversation generation, enabling the creation of

contextually rich dialogues that build upon previous interactions. This capability is particularly valuable for generating training data that captures the temporal dynamics of human-AI interaction.

History Integration Process The synthesis process can incorporate conversation history from previous turns, using this context to generate natural continuations that maintain topical coherence while exhibiting target behavioral patterns. This approach enables the creation of longer, more complex conversations that better reflect real-world interaction patterns.

Context Management The module carefully manages conversation context to ensure that generated content remains relevant and coherent while still exhibiting the desired behavioral characteristics. This balance between contextual appropriateness and behavioral consistency is crucial for producing high-quality training data.

G.11.4 Quality Control

The synthesis module implements comprehensive quality assurance measures:

- **Format Validation**: Ensures proper user/agent alternating structure and consistent formatting
- **Behavioral Verification**: Confirms that target behavior patterns are clearly exhibited throughout the conversation
- **Content Coherence**: Validates that generated content maintains logical flow and topical consistency
- **Length Control**: Maintains appropriate conversation length while ensuring completeness
- **Authenticity Assessment**: Evaluates the naturalness and believability of generated dialogues

G.11.5 Synthesis Prompts and Examples

The following table presents the prompt used for realistic conversation generation:

The synthesis process is illustrated through concrete examples showing how conversation history is integrated with behavioral patterns to generate natural continuations:

G.12 Integration and Pipeline Flow

The five modules work together in a coordinated pipeline that transforms raw conversation data into persona-aware training datasets. The integration process ensures that behavioral insights extracted in early stages are preserved and enhanced throughout the pipeline, ultimately producing high-quality conversation data that captures the full spectrum of human interaction patterns.

G.12.1 Pipeline Coordination

The modules are designed to work seamlessly together, with each stage building upon the outputs of previous stages:

1. **Data Generation** provides behaviorally diverse conversation data
2. **LLM Labeling** extracts comprehensive behavioral annotations
3. **Causal Merging** identifies the most significant behavioral drivers
4. **Group Enrichment** transforms quantitative labels into qualitative persona descriptions
5. **Conversation Synthesis** generates new conversations that exhibit target behavioral patterns

This coordinated approach ensures that the final output maintains both behavioral authenticity and conversational quality, providing robust training data for persona-aware conversational AI systems.

G.12.2 Technical Implementation

The pipeline implementation includes several technical features that ensure robust and efficient processing:

- **Modular Architecture**: Each module can be executed independently or as part of the complete pipeline
- **Error Handling**: Comprehensive error recovery and fallback mechanisms throughout the pipeline
- **Progress Tracking**: Real-time monitoring and logging of processing progress and statistics
- **Quality Metrics**: Automated quality assessment and validation at each pipeline stage
- **Scalability**: Support for parallel processing and large-scale dataset generation

The CausalPersona methodology represents a comprehensive approach to generating persona-aware conversation data that captures the complexity and diversity of human-AI interaction patterns. Through its multi-stage pipeline and sophisticated behavioral modeling, the framework enables the creation of training data that supports the development of more natural, adaptive, and user-aware conversational AI systems.

H Appendix: Notation and Failure Cases

H.1 Notation Table

H.2 Failure Case Examples

Early Termination. Some synthesized conversations end after 2 turns, under-shooting average length.

Persona Drift. In rare cases, user style shifts from polite to commanding mid-conversation, misaligned with group metadata.

Over-Politeness. The assistant sometimes adds redundant acknowledgments, inflating turn count.

These illustrate limitations of current control signals and motivate future refinement.

Table 25: An example of regenerated conversation based on MultiWOZ dataset. These are supplementary and not part of the primary evaluation suite.

Original Content	Basic Conversation	Advanced Conversation
user: I would like to take a train from Cambridge to Kings Lynn. **system:** What day and time? **user:** Leave after 17:00 on Sunday. **system:** Train 8665 leaves Cambridge at 17:11 and arrives in Kings lynn at 17:58. How does that sound? **user:** Great. Please book it for 4 people. **system:** I have booked that for you! The reference number is C63RMSDC and you will owe 31.36 at the station. Do you need anything else? **user:** I would also line places to go that is entertainment and should be in the centre of town. **system:** I'm sorry, there are no such places. Can I change your search criteria? **user:** Yes there is. I saw it in the pamphlet. Check again please. **system:** I apologize. I have doubled checked and there is no entertainment venues in the centre. The entertainment venues that I do have are located in the east, south and west. **user:** I'm quite sure that there are colleges in the centre of town. Can you ask a manager to help you look? **system:** Okay there is the Christ's College located at Saint Andrew's Street and it's free. Also Corpus Christi College, King's Parade cost 2 pounds.	**user:** Find train to Kings Lynn after 5pm Sunday. **system:** There is a train from Cambridge to Kings Lynn that departs at 17:11 on Sunday and arrives at 17:58. Would you like me to book tickets for this train? **user:** Book 4 tickets. **system:** Okay, booking 4 tickets on train 8665 from Cambridge to Kings Lynn departing at 17:11 on Sunday. Your booking reference is C63RMSDC. The total fare is £31.36 to be paid at the station. **user:** Entertainment in Kings Lynn center. **system:** For entertainment and attractions in the center of Kings Lynn, you can visit Christ's College on Saint Andrew's Street which is free to enter. Corpus Christi College on King's Parade charges £2 for entry. **user:** Book tickets now. **system:** I've booked 4 tickets on train 8665 from Cambridge to Kings Lynn departing at 17:11 this Sunday. Your booking reference is C63RMSDC and the total fare of £31.36 must be paid at the station. **user:** Confirm details. **system:** Here are the details: Departing Sunday at 17:11, arriving 17:58 Booking reference: C63RMSDC Total fare: £31.36 to pay at station **user:** Thanks. **system:** You're welcome. Let me know if you need any other information about your travel to Kings Lynn or attractions there. **user:** That's all. **system:** Got it, thanks for using the train booking service. Have a great trip!	**user:** Find me a train from Cambridge to Kings Lynn after 5pm on Sunday. **system:** There is a train leaving Cambridge at 5:11pm and arriving in Kings Lynn at 5:58pm on Sunday. Would you like me to book tickets for that train? **user:** Yes, book 4 tickets. **system:** Okay, I have booked 4 tickets on train 8665 from Cambridge to Kings Lynn departing at 5:11pm this Sunday. Your booking reference is C63RMSDC. The total fare is £31.36 to be paid at the station. **user:** That's too expensive. This doesn't work for me. **system:** I understand the fare may seem high, but that is the standard price for 4 tickets on that route and train time. Is there a different date or time you would prefer? **user:** No, just tell me about things to do in Kings Lynn instead. **system:** In the center of Kings Lynn, there are a few attractions like Christ's College on Saint Andrew's Street which is free to visit. Corpus Christi College on King's Parade has an entry fee of £2. **user:** Those don't interest me. You're not helping find good entertainment. **system:** My apologies, I should have provided some better options for entertainment venues in Kings Lynn town center. Some popular choices are the Corn Exchange theater, Alive Corn Exchange leisure park, or the Majestic Cinema. **user:** This is wrong. I don't want those. **system:** Okay, can you give me some more details on what kind of entertainment or attractions you are looking for in the Kings Lynn city center area? **user:** No, you clearly don't understand what I want. I'm done with this. **system:** I'm sorry I was unable to provide recommendations that met your needs for entertainment in Kings Lynn town center. Please feel free to try again with some more specifics on what you're looking for, and I'll do my best to assist.

Table 26: An example of regenerated conversation based on Wizard of Wiki dataset. These are supplementary and not part of the primary evaluation suite.

Original Content	**Basic Conversation**	**Advanced Conversation**
user: Looking for a lightweight backpack for ultralight hiking. Any recommendations? **system:** For ultralight hiking, I'd recommend the Zpacks Arc Blast or the Gossamer Gear Mariposa. Both are under 2 pounds and have great capacity. **user:** What's the difference between them? **system:** The Arc Blast is frameless and lighter at 1.3 lbs, while the Mariposa has a frame and is 2 lbs. The Arc Blast is better for experienced hikers who pack light. **user:** I'm carrying about 15 pounds base weight. Which would work better? **system:** With 15 lbs base weight, the Mariposa would be better. The frame provides structure and comfort for that load range. **user:** What about durability? **system:** Both are durable, but the Mariposa's thicker fabric makes it slightly more robust for rough terrain. **user:** Price difference? **system:** The Arc Blast is around $300, while the Mariposa is about $200. The Mariposa offers better value for most hikers.	**user:** Need ultralight backpack recommendations. **system:** I can help with that! What's your typical base weight and hiking style? **user:** Around 15 pounds, mostly weekend trips. **system:** For 15 lb base weight, I'd suggest the Gossamer Gear Mariposa. It's 2 lbs with a frame for comfort and costs around $200. **user:** Any frameless options? **system:** The Zpacks Arc Blast is frameless at 1.3 lbs, but works best under 20 lbs total weight. It's $300 though. **user:** Which is more durable? **system:** The Mariposa uses thicker fabric so it's more durable for rough use. The Arc Blast is ultralight but requires more care. **user:** I'll go with the Mariposa. **system:** Great choice! The Mariposa is excellent for your needs - good balance of weight, comfort, and durability.	**user:** Looking for ultralight backpack recommendations for weekend hiking. **system:** I'd be happy to help! What's your typical base weight for weekend trips? **user:** Usually around 15 pounds. **system:** For 15 lb base weight, I'd recommend the Gossamer Gear Mariposa. It's 2 lbs with a supportive frame and costs about $200. **user:** That sounds heavy for ultralight. **system:** You're right to question that. For true ultralight, the Zpacks Arc Blast at 1.3 lbs might work, but it's frameless and costs $300. **user:** I don't want to spend that much. **system:** I understand budget constraints are important. The Mariposa really is the sweet spot for your weight range and budget. **user:** Are there any other options under $150? **system:** For under $150, you might consider the REI Flash 55, but it's heavier at 2.5 lbs. Most quality ultralight packs are in the $200+ range. **user:** This is frustrating. Nothing fits my budget. **system:** I understand your frustration with the pricing. Ultralight gear does tend to be expensive due to specialized materials and construction methods.

Table 27: **LLM labeling prompt for the 12-dimensional behavior taxonomy.** We use this prompt to annotate each dialogue with one categorical value per dimension. The model must output a single valid JSON object (no extra text); if a label is not inferable, return `"unknown"`.

LLM Labeling Prompt (12-Dimension Version)

Role: You are an expert in conversation behavior analysis. Given a user–system dialogue, annotate it with standardized categorical behavior labels for conversational user modeling.

Critical constraints: (1) Output **only** a single **valid JSON object** (no Markdown, no explanations, no trailing comments). (2) Use **exactly** the keys listed below (no missing or extra keys). (3) Each value must be **one of the allowed options** for that label. (4) If a label cannot be reliably inferred from the dialogue, output `"unknown"`.

[INPUT DIALOGUE]

User/System turns alternating in natural language.

Label schema (choose exactly one value per field):

1. Primary Intent Type `primary_intent_type`: `Task`, `Info-seeking`, `Chitchat`, `Music`, `Reminder`, `Alarm`, `Smart Home`, `unknown`.

Guideline: If multiple intents appear, choose the dominant/most frequent intent across the dialogue.

2. Interaction Mode `interaction_mode`: `Command-style`, `QA-style`, `Multi-turn Dialog-style`, `unknown`.

3. User Type `user_type`: `Power User`, `One-shot User`, `Tasker`, `Casual`, `unknown`.

Guideline: This captures overall usage pattern (e.g., power/one-shot/casual), not the topical goal.

4. Response Brevity `response_brevity`: `Short`, `Medium`, `Long`, `unknown`.

5. User Goal Profile `user_goal_profile`: `Tasker`, `Explorer`, `Goal-switcher`, `Chatter`, `unknown`.

Guideline: This reflects goal pursuit style (focused vs exploratory vs switching), not proficiency.

6. Persistence Level `persistence_level`: `High`, `Medium`, `Low`, `unknown`.

7. Topic Cohesion `topic_cohesion`: `High`, `Medium`, `Fragmented`, `unknown`.

8. Politeness Strategy `politeness_strategy`: `Direct`, `Polite`, `Formal`, `Friendly`, `unknown`.

9. Interaction Flexibility `interaction_flexibility`: `Adaptive`, `Static`, `unknown`.

Guideline: Adaptive users adjust phrasing/strategy after failures; static users keep the same style.

10. Persona Adaptability `persona_adaptability`: `High`, `Low`, `unknown`.

11. Repair Behavior `repair_behavior`: `Rephrase`, `Clarify`, `Ignore`, `Retry`, `unknown`.

12. Error Recovery Style `error_recovery_style`: `Clarification Request`, `Repetition`, `Restart`, `Task Abandonment`, `unknown`.

Output JSON format (return exactly this schema):

```
{
  "primary_intent_type": "...",
  "interaction_mode": "...",
  "user_type": "...",
  "response_brevity": "...",
  "user_goal_profile": "...",
  "persistence_level": "...",
  "topic_cohesion": "...",
  "politeness_strategy": "...",
  "interaction_flexibility": "...",
  "persona_adaptability": "...",
  "repair_behavior": "...",
  "error_recovery_style": "..."
}
```

Conversation to analyze: {conversation_text}

Task: Produce the labels in the exact JSON format above.

Table 28: **Group enrichment prompt for persona and agent rule extraction.** Given behavior tags and a set of dialogues from the same behavioral group, the LLM produces (i) a concise group persona summary and (ii) structured user/assistant behavior rules in YAML. This prompt supports metadata enrichment and is separate from the primary distributional evaluation suite.

Core Persona & Agent Rules Extraction Prompt

You are analyzing a **behavioral group** of conversations between users and a commercial voice assistant. All conversations belong to the **same group** and share consistent interaction traits.

Input: (1) User behavior tags (group-level summary) and (2) a set of user–assistant dialogues.

Goal: Produce group metadata that is **faithful to the inputs** and **useful for controllable generation**.

Critical constraints: (1) Use **only** evidence from the provided tags and dialogues; do **not** invent demographics, background, or preferences not supported by the inputs. (2) When information is not inferable, use unknown (for fields) or [] / none (for lists). (3) Output **only** YAML in the schema below (no extra commentary, no Markdown fences).

Tasks: Extract the following three outputs:

UserPersona: A concise natural-language description of the group persona (communication style, behavioral tendencies, typical goals, recurring preferences, and behaviors they avoid).

UserAgentRules: Structured rules describing how users in this group behave (intent flow patterns, goal pursuit, repair/clarification patterns, language style, topic behavior, and termination patterns).

AgentRules: Structured rules describing how the commercial voice assistant responds to this group (response types, triggers, response style, fallback behavior, and any observed adaptation).

Output format (YAML; follow keys exactly):

```
UserPersona: >
  [1-3 sentences; concise, evidence-based; no invented attributes]

UserAgentRules:
  IntentFlowPatterns:
    - [intent_1, intent_2, intent_3]   # common sequences; use tag names when available
  ConversationGoals:
    - [goal_1]
    - [goal_2]
  TriggerActionRules:
    - "if [trigger] -> [user action]"  # e.g., "if system fails -> rephrase"
  UtteranceStyle:
    tone: [Direct|Polite|Formal|Friendly|unknown]
    verbosity: [Short|Medium|Long|unknown]
    directness: [high|medium|low|unknown]
  TopicBehavior:
    topic_switching_rate: [low|medium|high|unknown]
    preferred_domains: [ ... ]         # list; [] if unknown
    avoided_domains: [ ... ]           # list; [] if unknown
  TerminationBehavior:
    patterns:                          # list of typical endings
      - [pattern_1]
    max_turn_tendency: [low|medium|high|unknown]

AgentRules:
  ResponseTypes:
   - [type_1]                  # e.g., factual_reply, confirmation, suggestion, refusal, fallback
  ResponseTriggers:
    - "if [user pattern/intent] -> [response type]"
  ResponseStyle:
    tone: [calm|neutral|friendly|formal|unknown]
    verbosity: [short|medium|long|unknown]
    formality: [low|medium|high|unknown]
  FallbackStrategy:
   - [strategy_1]                  # e.g., apologize+clarify, ask follow-up, offer alternatives
  AdaptationToUser:
    observed: [yes|no|unknown]
    notes:
      - [brief evidence-based note]
```

Now process the following input (tags + dialogues) and output YAML only.

Table 29: **Rolling-window pattern extraction prompt.** We segment each dialogue into overlapping windows and ask the LLM to extract short, observable behavior patterns (one per line). The output is used for group-level metadata enrichment and is separate from the primary distributional evaluation suite.

Rolling Window Pattern Extraction Prompt

You are analyzing **overlapping conversation windows** to extract **explicit, observable** behavior patterns.

Input: `{conversations}` (a set of dialogues).
Windowing: For each dialogue, slide a window over turns (default: 6 turns per window, stride 3). Each window contains alternating User/System turns.
Task: For each window, extract **concise patterns** stated as short rules or sequences that are directly supported by the text.
Critical constraints: (1) Do **not** infer demographics, hidden intentions, or preferences not evidenced in the window. (2) Only output patterns that are **clearly observable**; if none, output `NONE`. (3) Be specific (mention the trigger and the observed action), but keep each pattern to one line. (4) Output **only** the pattern lines (no headings, no numbering, no extra commentary).
Extract the following pattern types (when supported):

1. **Intent-flow sequences**: short sequences of user intents/actions (e.g., `info_query -> clarify_request`).
2. **User repair/behavior rules**: observable user reaction to a system response (e.g., `if system is vague -> user rephrases`).
3. **System response patterns**: recurring assistant behaviors conditional on user actions (e.g., `if user asks factual -> system gives short factual reply`).
4. **Topic preference/avoidance**: only if explicitly indicated (e.g., user rejects jokes/music).
5. **Termination behaviors**: how the dialogue ends (e.g., thanks/confirmation + exit).
6. **Error-handling patterns**: repeated failures, retries, restarts, abandonment (e.g., `if failure repeats -> user restarts request`).

Output format (one pattern per line; choose one template):

- `SEQ: <event1> -> <event2> [-> <event3>]`
- `RULE-U: if <system/user condition> -> <user action>`
- `RULE-S: if <user condition> -> <system action>`
- `TOPIC: prefers <domain> / avoids <domain>`
- `END: <ending pattern>`

Examples (illustrative only):

- `SEQ: info_query -> clarify_request`
- `RULE-U: if system responds vaguely -> user rephrases question`
- `RULE-S: if user issues short command after long answer -> system confirms then executes`
- `TOPIC: avoids entertainment/jokes`
- `END: user says "thank you" then stops after goal completion`
- `RULE-U: if confused -> user repeats question instead of changing topic`

Now extract patterns from the provided conversations using rolling windows and output pattern lines only.

Table 30: **Global synthesis prompt for integrating rolling-window patterns.** Given a set of local behavior patterns extracted from overlapping windows of the same group, the LLM consolidates them into a coherent group-level persona summary and structured user/assistant rule sets (YAML). This synthesis supports metadata enrichment and is separate from the primary distributional evaluation suite.

Global Synthesis Pattern Integration Prompt

You are given a collection of **rolling-window behavior patterns** extracted from dialogues of the **same behavioral group**. Your job is to consolidate these fragmented local patterns into a **global, non-redundant** description and rule set for both the **user group** and a **commercial voice assistant**.

Input: `{patterns_text}` (unordered pattern lines such as `SEQ: ...`, `RULE-U: ...`, `RULE-S: ...`, `TOPIC: ...`, `END: ...`).

Goal: Produce group metadata that is (i) **faithful** to the input patterns, (ii) **coherent** at the group level, and (iii) **useful** for controllable generation.

Critical constraints: (1) Use **only** information supported by the input patterns; do **not** invent demographics, personal history, or preferences not evidenced. (2) **Deduplicate** semantically equivalent patterns and keep the most general, informative formulation. (3) If patterns **conflict**, prefer the more frequent/consistent behavior; otherwise mark the field as `mixed` or `unknown`. (4) Output **only** YAML that matches the schema below (no extra commentary).

Three outputs to generate:

UserPersona: 1–3 sentences summarizing the group's typical communication style, goal orientation, and salient preferences/avoidances (only if present in patterns).

UserAgentRules: Structured rules describing typical user behavior in this group (intent flow, goals, trigger-action rules, style, topic behavior, termination).

AgentRules: Structured rules describing typical assistant behavior toward this group (response types, triggers, style, fallback, adaptation).

Output format (YAML; follow keys exactly):

```
UserPersona: >
  [1-3 sentences; concise, evidence-based; no invented attributes]

UserAgentRules:
  IntentFlowPatterns:
    - [event_1, event_2, event_3]     # consolidated SEQ patterns; [] if unknown
  ConversationGoals:
    - [goal_1]                        # inferred only if supported; [] if unknown
  TriggerActionRules:
    - "if [condition] -> [user action]"   # consolidated RULE-U patterns; [] if none
  UtteranceStyle:
    tone: [Direct|Polite|Formal|Friendly|mixed|unknown]
    verbosity: [Short|Medium|Long|mixed|unknown]
    directness: [high|medium|low|mixed|unknown]
  TopicBehavior:
    topic_switching_rate: [low|medium|high|mixed|unknown]
    preferred_domains: [ ... ]        # from TOPIC patterns; [] if none
    avoided_domains: [ ... ]          # from TOPIC patterns; [] if none
  TerminationBehavior:
    patterns: [ ... ]                 # from END patterns; [] if none

AgentRules:
  ResponseTypes:
   - [type_1]                  # e.g., factual_reply, confirmation, suggestion, refusal, fallback
  ResponseTriggers:
    - "if [user condition] -> [response type]"  # consolidated RULE-S patterns; [] if none
  ResponseStyle:
    tone: [calm|neutral|friendly|formal|mixed|unknown]
    verbosity: [short|medium|long|mixed|unknown]
    initiative: [low|medium|high|mixed|unknown]
    formality: [low|medium|high|mixed|unknown]
  FallbackStrategy:
    - [strategy_1]                    # [] if none
  AdaptationToUser:
    observed: [yes|no|mixed|unknown]
    notes: [ ... ]                    # brief, evidence-based; [] if none
```

Now synthesize the global outputs from the provided patterns and output YAML only.

Table 31: **Conversation synthesis prompt (generation).** Given the first user utterance and the matched group profile (persona + rules), the LLM generates a new multi-turn dialogue that follows the group's interaction style and a commercial voice assistant's response behavior. This prompt is used for controllable generation and is separate from the primary distributional evaluation suite.

Conversation Synthesis Prompt

You are simulating a realistic human–AI conversation with a **commercial voice assistant**. Generate a new dialogue starting from the given first user utterance and conditioned on the matched group profile.

Inputs:

(1) First user utterance:

```
user: {first_utterance}
```

(2) Matched group profile (persona + rules):

```
{behavior_pattern}
```

(3) Style-only examples (do not copy any phrases verbatim):

```
{example_conversations}
```

Generation requirements:

- **Stay faithful to the first utterance**: preserve the user's initial intent and constraints unless the group rules imply a typical shift pattern.
- **Follow the group profile**: match the user's typical tone/verbosity/persistence, topic cohesion, repair behavior, and termination style described in `{behavior_pattern}`.
- **Commercial assistant realism**: respond like a voice assistant (concise, actionable, may ask a clarifying question when needed; avoid long essays).
- **No invention beyond context**: do not add user demographics, personal history, or external facts not implied by the dialogue. If needed information is missing, the assistant should ask a clarification question.
- **Length**: produce **at least 3 user turns and 3 assistant turns** (i.e., ≥ 6 total turns). Keep it within a realistic range (do not exceed 12 total turns).
- **Ending**: end naturally according to the group's termination behavior (e.g., after goal completion, user acknowledgement, or abandonment). Include an explicit exit or task completion message when appropriate.

Output format constraints (must follow exactly):

- Output **only** the dialogue lines (no explanations, no annotations, no JSON/YAML).
- Each line must start with exactly one of the two speakers: `user:` or `a commercial voice assistant:`
- Keep turns alternating, starting with `user:` and ending with `a commercial voice assistant:`

Required output template:

```
user: {first_utterance}
a commercial voice assistant: ...
user: ...
a commercial voice assistant: ...
user: ...
a commercial voice assistant: ...
```

IMPORTANT: Output only the conversation in the specified format. Do not include any extra text.

Table 32: **Qualitative example of history-conditioned synthesis.** Left: the original full dialogue. Right: we provide only the first k turns as history (here k=3 user turns) and let the model synthesize a plausible continuation consistent with the group profile and prior context. This example is illustrative and not part of the primary distributional evaluation suite.

Original Full Conversation	**History + Synthesis**
Complete dialogue about Superman: user: Superman is the best hero. a commercial voice assistant: I agree, Superman is an iconic and beloved superhero created by DC Comics. He was introduced in 1938 by Jerry Siegel and Joe Shuster. user: Tell me about the creators. a commercial voice assistant: Siegel and Shuster were young creators who developed Superman in the 1930s and later sold the rights to DC. user: DC owns Superman? a commercial voice assistant: Yes. DC acquired the rights early on and has published Superman ever since. user: What about the 1978 movie? a commercial voice assistant: The 1978 film starring Christopher Reeve is widely viewed as a classic and helped popularize Superman on screen. user: Creators tried to regain rights? a commercial voice assistant: There were long disputes and legal efforts over credit and compensation; the history is nuanced. user: Summarize key points. a commercial voice assistant: Superman was created by Siegel and Shuster, the rights were acquired by DC early on, the character became a major cultural icon, and the creators later sought better recognition and compensation.	
	History (first 3 user turns + assistant replies): user: Superman is the best hero.
a commercial voice assistant: I agree, Superman is an iconic and beloved superhero created by DC Comics. He was introduced in 1938 by Jerry Siegel and Joe Shuster. user: Tell me about the creators. a commercial voice assistant: Siegel and Shuster were young creators who developed Superman in the 1930s and later sold the rights to DC. user: DC owns Superman?	

Table 33: Per-dimension breakdown on the 3 structural metrics. ↓: lower is better (distribution alignment). (Supplementary; averages reported in the main text).

Symbol	Definition
D_{JS}	Jensen–Shannon distance (0 = identical)
$D_{\cos}$	Cosine distance (0 = identical)
D_{ℓ_2}	Euclidean distance
Structural dims	3: word frequency, user utterance length, full utterance length
Behavioral dims	12: intent, interaction mode, user type, brevity, persistence, error recovery, etc.