Beyond Individual Personas: Aligning Synthetic Dialogue to Population-Level Behavior Distributions

Xinyi Liu[*]
Amazon
University of Illinois
Urbana–Champaign
liu323@illinois.edu

Rinat Khaziev[†]
Semantic Code Lab
rinat@semantic.sh

Hooshang Nayyeri
Amazon
hooshang@amazon.com

Emine Yilmaz
Amazon
University College London
eminey@amazon.co.uk

Charith Peris
Amazon
perisc@amazon.com

Hari Thadakamalla
Amazon
thadakah@amazon.com

Abstract

Synthetic dialogue corpora are increasingly used as proxies for target data, yet persona-grounded generators optimize individual conversations rather than corpus composition, yielding plausible dialogues with distorted population-level behavior mixes. We introduce **GroupPersona**, a framework that aligns synthetic corpora with a reference behavior distribution. GroupPersona turns population statistics into generation controls by discovering recurring behavioral groups in the reference corpus and using their prevalence and interaction patterns to condition user agents.

We evaluate corpus-level fidelity through behavioral alignment and structural preservation, measured by average Jensen–Shannon distance over 12 behavior attributes and 5 dialogue structure attributes, respectively. Across four corpora spanning two source families and two construction regimes, GroupPersona reduces behavioral Jensen–Shannon distance from 0.234 to 0.177 over the best-average baseline, a 24.4% reduction. It is best or tied-best on all four corpora. It maintains comparable structural distance at 0.167 versus 0.169. Together, these results show that GroupPersona improves population-level behavioral fidelity while preserving dialogue structure.

1 Introduction

Persona-grounded dialogue synthesis has become a common way to create realistic conversational data. By assigning a persona or user profile to a generator, prior work can produce dialogues that are fluent, coherent, and locally consistent with a user description (Zhang et al., 2018; Jandaghi et al., 2024; Cheng et al., 2024; Wang et al., 2025). These synthetic dialogues are increasingly used beyond individual examples to expand scarce training data, support user-simulator evaluation, stress-test assistants, and build personalization benchmarks (Suresh et al., 2025; Li et al., 2025a; Dou et al., 2025; Zhao et al., 2025). In these settings, generated dialogues function as corpus-level proxies for target user populations.

This corpus-level use exposes a failure mode that per-dialogue persona checks cannot capture. A generator may produce dialogues that are individually fluent and persona-consistent while still producing the wrong mix of user behaviors. For example, a synthetic population may over-represent exploratory and polite users while under-representing terse, persistent, or task-abandoning users. An assistant evaluated on such data is tested against a population different from the one it is intended to serve, even if every generated dialogue is individually plausible. We call this mismatch between synthetic and reference behavior distributions *behavior-distribution misalignment*.

Addressing this problem requires turning corpus-level behavior statistics into generation-time controls. The control unit must be frequent enough to estimate reliably, expressive enough to represent recurring behavior mixtures, and interpretable enough to guide an LLM user agent. Existing alternatives satisfy these requirements only partially. Individual personas provide local conditioning but do not specify corpus composition (Zhang et al., 2018; Jandaghi et al., 2024; Cheng et al., 2024; Wang et al., 2025). Exact behavior tuples make composition explicit but can fragment the data

[*]Work done during an internship at Amazon.
[†]Work done while at Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 13903–13927
October 24-29, 2026 ©2026 Association for Computational Linguistics

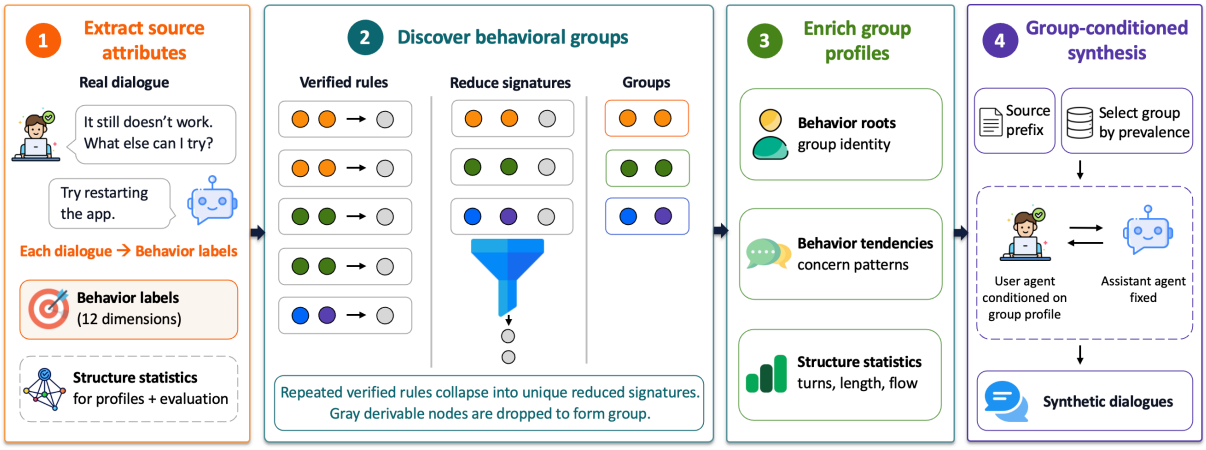

Figure 1: **GroupPersona overview.** The pipeline extracts source attributes, removes rule-derivable behavior labels to form behavioral groups, enriches group profiles, and uses prevalence-weighted groups for dialogue generation.

into sparse joint cases; classic association-rule and frequent-pattern mining instead target recurring co-occurrences without requiring every complete tuple to recur (Agrawal et al., 1993; Agrawal and Srikant, 1994; Han et al., 2007). Latent clusters reduce sparsity, but their identifiers are difficult to translate into generation instructions.

To address these limitations, we introduce **GroupPersona**, a framework that turns corpus-level behavior patterns into explicit generation-time controls. Figure 1 gives an overview. GroupPersona is built around two design choices. First, it separates behavioral identity from dialogue form: categorical behavior labels define groups, while continuous structural statistics such as turn count and utterance length are reserved for profile enrichment and evaluation. Second, it prevents redundant behavior labels from disproportionately shaping group formation by treating labels predictable from others as consequences rather than independent group-defining signals.

These choices yield a compact behavioral signature for each dialogue that retains the information most useful for distinguishing interaction patterns. Dialogues with similar signatures are then organized into recurring behavioral groups.

To turn these behavioral groups into generation controls, each group is converted into a user-agent profile containing behavior roots that define group identity, behavior tendencies that capture common accompanying patterns, and structural statistics that guide dialogue shape. During synthesis, groups are sampled according to their prevalence in the reference corpus, and the corresponding profiles condition the user agent. The assistant remains fixed and group-blind, ensuring that GroupPersona's population-level control is carried by the user side.

We evaluate corpus-level fidelity through behavioral alignment and structural preservation, using distributional distance over 12 behavior attributes and 5 dialogue structure attributes. As a secondary diagnostic, we also assess whether synthetic conversations remain calibrated to the quality characteristics of the reference data. Because fidelity to the reference corpus is the target, higher quality scores are not necessarily better; we therefore treat quality as a calibration diagnostic rather than an objective to maximize. GroupPersona improves behavioral alignment by 24.4% over the best-average baseline while preserving comparable structural fidelity. It also more closely matches reference-conversation quality than the next-best baseline. Together, these results show improved population-level behavioral fidelity without sacrificing dialogue structure or reference-like quality.

We make three contributions:

- **Population alignment target.** We identify behavior-distribution misalignment as a corpus-level failure mode of synthetic dialogue generation and formulate fidelity as matching reference behavior distributions while preserving dialogue structure.
- **Group-conditioned control.** We propose **GroupPersona**, which distills dialogues into core behavioral signatures and converts recurring groups into prevalence-weighted controls for user-agent synthesis.
- **Alignment gains.** Across four corpora, GroupPersona improves behavioral alignment by

24.4% while preserving structural fidelity, with additional quality-calibration results showing closer agreement with reference conversations.

2 Related Work

Persona-grounded dialogue synthesis and user simulation. Personalized dialogue generation conditions models on explicit personas, user profiles, or interaction histories to improve consistency with individual speakers (Li et al., 2016b; Mazaré et al., 2018; Zhang et al., 2018; Wolf et al., 2019). Recent work extends this paradigm through richer persona construction, faithful persona-based generation, user adaptation, implicit or long-term profiles, and customizable role-playing agents (Kim et al., 2023; Jandaghi et al., 2024; Cheng et al., 2024; Wang et al., 2025; Li et al., 2025b; Yang et al., 2025). Synthetic dialogues are also widely used for low-resource data generation, multi-agent simulation, user-simulator evaluation, and personalization benchmarks (Suresh et al., 2025; Li et al., 2025a; Dou et al., 2025; Zhao et al., 2025). These approaches primarily control or evaluate individual conversations. GroupPersona instead targets corpus composition by asking whether synthetic dialogues collectively preserve the behavior distribution of a reference population.

Corpus-level fidelity and population-level control. Prior work on language generation shows that individually plausible outputs can still diverge from target data in aggregate properties (Hashimoto et al., 2019; Li et al., 2016a; Holtzman et al., 2020; Welleck et al., 2020), while dialogue evaluation motivates behavior-aware criteria beyond scalar quality scores (Mehri and Eskenazi, 2020; Finch and Choi, 2020). These lines of work provide tools for measuring distributional or behavioral differences, but do not directly translate a reference population's behavior distribution into generation-time controls. GroupPersona addresses this gap by constructing interpretable behavioral groups from recurring dependencies in the reference corpus and using their prevalence to control synthesis. Its group discovery combines association-rule mining for transparent dependencies (Agrawal et al., 1993; Agrawal and Srikant, 1994; Han et al., 2007) with LLM verification for semantic filtering (Zheng et al., 2023; Gu et al., 2024).

3 Population-Level Alignment Protocol

We formulate population-level alignment as preserving the behavioral composition of a reference corpus while maintaining its basic dialogue structure. Let $X = \{x_n\}_{n=1}^{N}$ denote the reference corpus and $\hat{X} = \{\hat{x}_m\}_{m=1}^{M}$ the synthetic corpus, where each element is a complete dialogue. A feature extractor ϕ maps each dialogue to D dialogue-level attributes,

$$\phi(x) = (\phi_1(x), \ldots, \phi_D(x)).$$

We partition these attributes into behavioral attributes B and structural attributes S. Behavioral attributes describe how users interact, including primary intent, interaction mode, and politeness strategy. Structural attributes describe dialogue form, including turn count and utterance length. This separation makes behavioral fidelity the primary target while using structure to check that improved alignment does not distort basic dialogue form.

Categorical attributes are counted directly, while continuous structural attributes are discretized using quantile bins computed from the reference corpus and applied to both corpora.

For each attribute d, we quantify the mismatch between the reference distribution p_d and the synthetic distribution $\hat{p}_d$ using Jensen–Shannon (JS) distance. Let $m_d = (p_d + \hat{p}_d)/2$:

$$d_{\mathrm{JS}}(p_d, \hat{p}_d) = \sqrt{\tfrac{1}{2}\mathrm{KL}(p_d \| m_d) + \tfrac{1}{2}\mathrm{KL}(\hat{p}_d \| m_d)}.$$

JS distance is symmetric and bounded, and remains well-defined when an attribute value appears in only one of the two corpora.

We then average these per-attribute distances separately over behavioral and structural attributes:

$$\text{Behav-JS} = \frac{1}{|B|} \sum_{d \in B} d_{\mathrm{JS}}(p_d, \hat{p}_d), \quad (1)$$

$$\text{Struct-JS} = \frac{1}{|S|} \sum_{d \in S} d_{\mathrm{JS}}(p_d, \hat{p}_d). \quad (2)$$

Lower values indicate closer agreement with the reference corpus. We use Behav-JS as the primary measure of population-level behavioral alignment and Struct-JS as a preservation check on dialogue form.

4 GroupPersona: Group-Conditioned Population Alignment

GroupPersona turns behavior patterns in a reference corpus into generation controls that preserve

its population composition. As shown in Figure 1, it extracts behavioral and structural attributes, reduces redundant behavior labels to compact signatures, organizes recurring signatures into behavioral groups, enriches each group into a user-agent profile, and samples groups according to their reference-corpus prevalence during synthesis.

4.1 Stage 1: Extracting Source Attributes

GroupPersona represents each reference dialogue with two types of attributes that play different roles in the pipeline. *Behavior attributes* capture how the user interacts and provide the categorical signals used for group discovery. *Structural attributes*, such as turn count and utterance length, describe dialogue form and are used for profile enrichment and structural evaluation. Keeping these roles separate prevents group discovery from being driven by surface properties such as dialogue length.

For example, a dialogue with brief, direct requests can be labelled as command-style, task-focused, and short in its behavioral representation, while its turn count and utterance lengths are recorded separately as structural attributes. The *Stage 1 output* row of Table 1 carries this example forward through the subsequent grouping and profile-enrichment stages.

The behavioral schema is grounded in prior work on intent modeling (Sarikaya, 2017), user adaptation and dialogue evaluation (Cheng et al., 2024; Finch and Choi, 2020; Mehri and Eskenazi, 2020), and spoken-dialogue error handling (Bohus, 2004; Bohus and Rudnicky, 2005). Based on these perspectives, we define 12 behavioral dimensions covering intent and interaction mode, user type and goal, brevity and persistence, topic cohesion and politeness, flexibility and adaptability, and repair and recovery behavior. Each dialogue receives one categorical value for each dimension. Full definitions, value vocabularies, and annotation rules are provided in Appendix A.

4.2 Stage 2: Discovering Behavioral Groups

The behavioral labels extracted in §4.1 provide a rich description of each dialogue, but using the full 12-dimensional combination directly would fragment the corpus into many sparse behavior tuples. Some labels are also redundant because they are predictable from other behaviors in the same dialogue. GroupPersona addresses both issues by removing locally predictable labels and grouping dialogues using the resulting compact behavioral signatures.

Table 1: **Running GroupPersona trace from labels to profile.**

Pipeline step	Trace
Stage 1 output	command-style, task-focused, short, direct, ...
Verified rule	command-style + task-focused ⇒ short
Reduced signature	command-style, task-focused, direct, ... short removed
Behavioral group	**Roots:** command-style, task-focused
Enriched profile	**Roots:** command-style, task-focused **Tendency:** short responses

Verified rules. We represent each dialogue x as a set P_x of 12 behavior attribute–value pairs and mine candidate dependencies of the form

$$A \Rightarrow \ell_t,$$

where A contains one to three behavior pairs and $\ell_t = (a = v)$ is another behavior pair. Such a rule indicates that the behaviors in A reliably predict the target behavior ℓ_t in the group-discovery data. In the running example, the *Verified rule* row of Table 1 marks short responses as predictable from command-style and task-focused behavior.

Candidate rules are screened by support, confidence, and lift and pruned to minimal antecedents. Because statistical associations may reflect duplicated labels, corpus templates, or incidental correlations, an LLM verifier assesses each minimal rule using its statistics and matched group-discovery examples. Only accepted rules can mark a target as *rule-derivable*. Full mining criteria, pruning details, verifier prompts, and diagnostics are provided in Appendices J and K.

Reduced signatures. Accepted rules are applied locally to each dialogue. Starting from the full label set P_x, GroupPersona removes a target ℓ_t when an accepted rule $A^\star \Rightarrow \ell_t$ applies. Rules are applied iteratively until no further target can be removed, yielding the reduced signature R_x.

The reduction is dialogue-specific, so the same behavior label may be removed in one dialogue and retained in another. In the *Reduced signature* row of Table 1, short is removed from group identity while the remaining labels continue to characterize the dialogue. Removed labels are retained for later profile enrichment.

Behavioral groups. GroupPersona forms groups from similar reduced signatures. At each step, the most frequent signature among the remaining dialogues is selected as a seed R_s. An unassigned dialogue joins the group when its signature has sufficient Jaccard similarity to the seed:

$$J(R_x, R_s) = \frac{|R_x \cap R_s|}{|R_x \cup R_s|} \geq \tau_{\text{jacc}}.$$

The procedure repeats until all dialogues have been assigned.

After grouping, we select up to $K_{\max}$ *behavior roots* to give each valid group an interpretable identity. A retained behavior pair $\ell = (a = v)$ qualifies as a root when it is both common within group C and distinctive relative to the group-discovery population:

$$\Pr[\ell \mid C] \geq \tau_{\text{hom}}, \qquad \frac{\Pr[\ell \mid C]}{\Pr[\ell]} \geq \tau_{\text{lift}}.$$

The *Behavioral group* row of Table 1 shows how the retained command-style and task-focused labels become roots of the resulting group. Selecting roots from reduced signatures prevents rule-derivable consequences from defining group identity. Small clusters or clusters without an admissible root are treated as residuals and merged into the nearest valid group. The locked grouping configuration and residual-handling details are provided in Appendix J.

4.3 Stage 3: Enriching Group Profiles

The groups discovered in §4.2 provide compact population units, but synthesis requires a richer description of how their members typically interact. GroupPersona therefore converts each valid group into a user-agent profile with three complementary fields.

Behavior roots define the group's core behavioral identity. *Behavior tendencies* summarize frequent non-root patterns, including labels removed during signature reduction. *Structural statistics* summarize dialogue form, including turn count, word count, and utterance length.

Together, these fields separate group identity from its typical realization. The *Enriched profile* row of Table 1 completes the running trace: command-style and task-focused remain roots, while short responses return as a tendency without defining group membership. Structural statistics then provide the expected dialogue shape. A complete example profile is provided in Appendix M.

4.4 Stage 4: Prevalence-Weighted Group-Conditioned Synthesis

The final stage turns behavioral groups into corpus-level generation controls. Let C_g denote group g after residual assignment. Its sampling probability is determined by its prevalence in the group-discovery data:

$$\pi_g = \frac{|C_g|}{\sum_h |C_h|}.$$

For each synthetic dialogue, GroupPersona samples a group according to π_g and conditions the user agent on the corresponding profile together with the dialogue prefix.

Prevalence-weighted sampling controls the composition of the synthetic corpus, while the selected profile controls how the user behaves within an individual dialogue. The assistant remains fixed and group-blind, receiving only the visible dialogue history and following a generic assistant prompt. Thus, population-level information enters synthesis through the user side without exposing the assistant to group metadata. Full user-agent and assistant-agent prompts are provided in Appendix O.

5 Evaluation Setup: PopAlign-Bench

Benchmark design. No single existing source simultaneously provides a controlled dialogue setting, broad behavioral diversity, and a unified user–assistant evaluation format, motivating the construction of PopAlign-Bench. PopAlign-Bench is designed to test whether population alignment holds across different interaction populations and remains robust to changes in how the same underlying content is expressed. To cover different interaction settings, we combine structured task-oriented and knowledge-grounded dialogues with broader, less scripted Reddit discussions. Specifically, PopAlign-Bench is constructed from MultiWOZ 2.2 (Zang et al., 2020), Wizard of Wikipedia (Dinan et al., 2019), and Pushshift Reddit (Baumgartner et al., 2020).

We vary these two factors in a 2×2 design. The source-family axis contains a Dialogue family, combining MultiWOZ 2.2 and Wizard of Wikipedia, and a Reddit family derived from Pushshift discussions. The construction-regime axis contains Basic, which retains the source interaction structure with minimal standardization, and Advanced, which recasts the same underlying content into a controlled user–assistant format with greater linguistic and interaction-style variation.

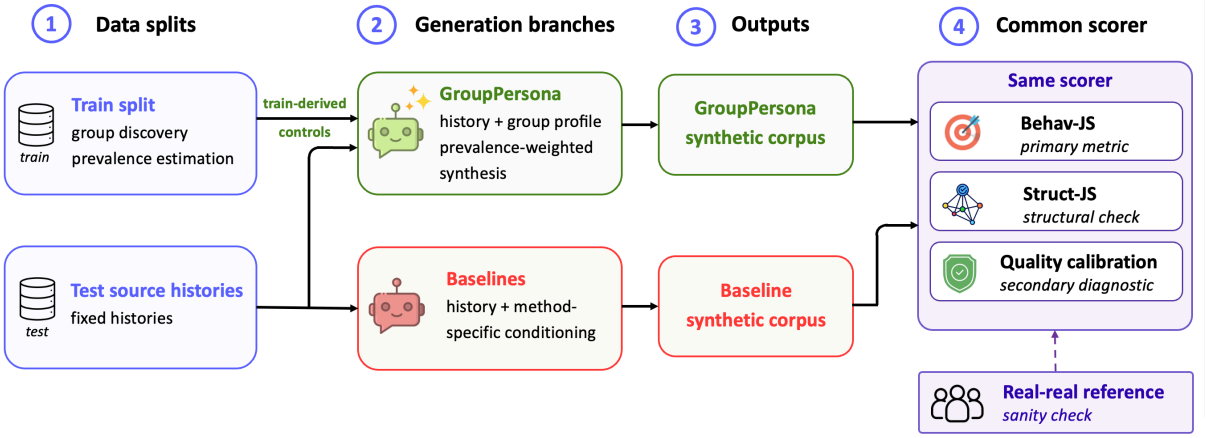

Figure 2: **Evaluation protocol.** All methods continue the same fixed test histories and differ only in conditioning. GroupPersona uses train-derived behavioral groups and prevalence. Behav-JS and Struct-JS are the main metrics, with reference–reference and quality calibration as diagnostics.

This yields Dialogue-Basic, Dialogue-Advanced, Reddit-Basic, and Reddit-Advanced, each with 1,000 training and 500 test conversations.

Evaluation protocol. Training data are used only for method-specific conditioning. For GroupPersona, 200 conversations from the training split are held out for configuration selection. The remaining 800 training conversations are used for rule mining, verification, signature reduction, behavioral-group discovery, and prevalence estimation. The 500-conversation test split is used only for reported evaluation and is never used for configuration selection.

For each test dialogue, all methods start from the same observed history and differ only in the additional conditioning provided during generation. This provides a controlled comparison of the different conditioning strategies. Figure 2 summarizes the protocol.

Scoring, validation, and diagnostics. All generated corpora are scored with the same labelling and extraction pipeline. Behav-JS and Struct-JS are defined in §3. We validate the behavior labeller on 200 independently human-labelled conversations, where it achieves 0.85 macro-F1 and $\kappa = 0.77$ against the adjudicated human reference. Full per-dimension results and annotation details are provided in Appendix L.

We also report two diagnostics. Reference–reference distance is computed between the reference marginals of the training and test splits of the same corpus. Quality calibration measures how closely synthetic conversations match the quality characteristics of reference conversations, rather than rewarding higher judge scores alone. Full calibration details are provided in Appendix P. Unless otherwise noted, main-text results use Claude Sonnet 4 for behavior labelling. Bootstrap confidence intervals are reported in Appendix G.

Table 2: **Method-specific conditioning.** All methods share the same fixed test history; baseline details are in Appendix C.

Method(s)	Additional conditioning
GroupPersona	Train-derived behavioral groups and prevalence
ConceptPersona FaithfulPersona PersonaLens	Per-record persona or profile
AutoPAL Interlocutor	User or persona adaptation
DiaSynth, ICL	Planning or demonstrations

Compared methods. All methods start from the same fixed test histories and differ only in their additional conditioning, summarized in Table 2. GroupPersona estimates population-level behavioral groups and their prevalence from the remaining 800 training conversations, while baselines adapt their respective persona-, profile-, adaptation-, planning-, or demonstration-based conditioning strategies to the shared benchmark format.

We compare against seven baselines spanning these conditioning strategies: AutoPAL (Cheng et al., 2024), ConceptPersona (Kim et al., 2023), DiaSynth (Suresh et al., 2025), FaithfulPersona (Jandaghi et al., 2024), ICL (Brown et al., 2020), Interlocutor (Occhipinti et al., 2025), and

PersonaLens (Zhao et al., 2025). Baseline adaptations and prompts are detailed in Appendix C.

6 Experiments

We organize the experiments around four questions: whether GroupPersona improves population-level alignment, whether the discovered groups explain the gain, which profile signals drive synthesis, and whether the resulting conversations remain calibrated to reference quality. All experiments follow the setup in §5.

6.1 Q1. Does GroupPersona improve population-level alignment?

Q1 asks whether population-derived group controls improve behavioral alignment without sacrificing dialogue form.

GroupPersona improves behavioral alignment across settings. GroupPersona achieves the lowest average Behav-JS in Table 3, reducing it from 0.234 for the best-average baseline to 0.177, a 24.4% reduction. It is best or tied-best on all four corpora, with the largest gains on Dialogue-Basic and Reddit-Advanced, a smaller gain on Dialogue-Advanced, and a tie on Reddit-Basic. The improvement therefore holds across both source families and construction regimes rather than being driven by a single benchmark setting.

Local plausibility does not ensure population fidelity. Table 4 gives a concrete view of the distributional difference. The reference continuation is command-style, task-focused, short, and direct. GroupPersona preserves this interaction pattern, whereas PersonaLens shifts toward a longer, more exploratory exchange from the same history. Both continuations are locally plausible, but only one preserves the reference interaction pattern.

Behavioral gains preserve form and extend beyond the behavior schema. GroupPersona maintains structural alignment comparable to the best-average baseline, with average Struct-JS of 0.167 versus 0.169. The gain also appears in label spaces never used for group construction. Ext-Act, averaged over MIDAS and DailyDialog dialogue acts (Yu and Yu, 2021; Li et al., 2017), improves by 23% over the best-average baseline, while Ext-E/T remains near the best baseline at 0.068 versus 0.066. These checks indicate that the behavioral gain is not obtained by distorting dialogue form or narrowly overfitting the 12-dimensional behavior schema. External label spaces are described in Appendix H, with per-corpus results in Appendix Table 10.

The gain is not tied to one synthesis or labelling family. Because both dialogue synthesis and behavior labelling rely on LLMs, we test whether the observed gain is specific to either the synthesizer or labeller family. For this robustness analysis, we use ConceptPersona, a strong baseline that, in our adaptation, shares GroupPersona's behavior-label vocabulary but represents each record independently rather than through population-level behavioral groups. Table 5 evaluates three settings while keeping the benchmark and rewriting procedure fixed: Claude Sonnet 4 for both synthesis and labelling, DeepSeek-R1 for both, and Claude synthesis with independent DeepSeek-R1 relabelling. The first two test whether the gain transfers across synthesis families, while the third isolates sensitivity to the labeller. GroupPersona improves over ConceptPersona in all three settings, reducing Behav-JS by 21.4%–24.7%. Thus, the alignment gain persists across model families and under independent relabelling.

6.2 Q2. Do the discovered behavioral groups explain the gain?

Q2 tests whether the improvement comes from GroupPersona's specific group-discovery procedure rather than from group conditioning alone. We also test whether LLM verification contributes beyond statistical rule mining. Table 6 compares these alternatives, with full per-corpus results in Appendix Table 11.

Behaviorally structured groups drive the gain. Replacing GroupPersona's groups while keeping the synthesis pipeline unchanged substantially weakens behavioral alignment. Average Behav-JS rises from 0.177 to 0.207 with k-means groups and to 0.245 with random groups. The benefit therefore does not come from adding an arbitrary group identifier, but from organizing dialogues around recurring behavioral structure.

The resulting grouping also remains practically sized, producing 74–94 groups across corpora with residual rates of 8.9%–9.2%. This preserves population heterogeneity without leaving a large fraction of dialogues outside the discovered groups. Exact per-corpus diagnostics are reported in Appendix Table 13.

Table 3: **Distribution alignment on PopAlign-Bench under Claude Sonnet 4 labelling.** Dialogue-B/A and Reddit-B/A denote Basic/Advanced. Columns report Behav-JS by corpus and average Behav-JS, Struct-JS, Ext-Act, and Ext-E/T. Lower is better; bold and underline mark the best and second-best non-reference results.

	Behav-JS by corpus				Average checks			
Method	Dialogue-B	Dialogue-A	Reddit-B	Reddit-A	Behav	Struct	Ext-Act	Ext-E/T
AutoPAL	0.244	0.256	0.187	0.274	0.240	0.170	0.167	0.073
ConceptPersona	0.236	<u>0.252</u>	0.181	<u>0.269</u>	0.235	0.177	0.154	0.070
DiaSynth	0.247	0.259	<u>0.162</u>	0.288	0.239	0.176	0.155	0.070
FaithfulPersona	0.235	0.263	**0.156**	0.280	<u>0.234</u>	<u>0.169</u>	<u>0.145</u>	0.072
ICL	0.327	0.323	0.310	0.316	0.319	0.231	0.248	0.085
Interlocutor	<u>0.232</u>	0.254	0.265	0.284	0.258	0.183	0.186	0.081
PersonaLens	0.245	0.271	0.210	0.284	0.253	0.176	0.168	**0.066**
GroupPersona	**0.148**	**0.228**	**0.156**	**0.174**	**0.177**	**0.167**	**0.112**	<u>0.068</u>
Reference	*0.030*	*0.024*	*0.029*	*0.023*	*0.026*	*0.085*	—	—

Table 4: **Same-history user continuations.** All methods continue the same restaurant-search history; assistant turns are omitted. Full continuations and conditioning details are in Appendix D.

Source	**User continuation**	**Interaction pattern**
Reference	"Italian works." → "What's the address?" → "Phone number?"	Command-style, task-focused, short, direct
GroupPersona	"Italian." → "Address." → "Phone?"	Command-style, task-focused, short, direct
PersonaLens	"I'd like Italian, something not too expensive." → "Can you tell me where it is and how to reach them?" → "Do they take reservations for tonight…"	QA-style, exploratory, medium, polite

Table 5: Cross-family robustness on Behav-JS (↓). R/S/L denotes rewriter/synthesizer/labeller: C = Claude Sonnet 4, D = DeepSeek-R1. Reduction is relative to ConceptPersona (↑). GroupPersona results are bolded.

Setting	**R/S/L**	**GroupP.**	**ConceptP.**	**Red.**
In-family synth. + label				
Claude Sonnet 4	C/C/C	**0.177**	0.235	24.7%
DeepSeek-R1	C/D/D	**0.187**	0.238	21.4%
Out-of-family relabel				
C→D	C/C/D	**0.177**	0.234	24.4%

Semantic verification improves signature reduction. The LLM verifier filters statistically strong dependencies that may reflect duplicated labels, corpus-specific templates, or incidental co-occurrence rather than meaningful behavioral relationships. For example, it rejects a high-scoring rule linking *response brevity = short* to *topic cohesion = fragmented* when the association reflects a construction-specific co-occurrence, so the rule removes no label during signature reduction. Removing this step increases average Behav-JS from 0.177 to 0.189, showing that statistical rule mining alone is less effective. A human audit of 320 rules rated the verifier decisions $3.83/5$ on average (Appendix L). Full diagnostics and implementation details are provided in Appendix J.

6.3 Q3. Which profile signals matter for synthesis?

Q3 asks what information is needed to realize a selected behavioral group during synthesis. We keep the discovered groups and the selected group for each dialogue fixed, and remove one conditioning component at a time: the source prefix, natural-language root guidance, behavior tendencies, or structural statistics. The group's explicit root identifier is retained in every ablation, so differences reflect how each profile component guides generation rather than changes in group selection.

Root guidance and structure matter most. Removing structural statistics increases average Behav-JS by 0.066, while removing the natural-language root guidance increases it by 0.041. This shows that selecting the right group is not enough. The user agent also needs explicit cues about how the group behaves and how its dialogues are typically structured. Source context and behavior tendencies also help, but their effects are smaller.

Table 6: GroupPersona ablations under Claude Sonnet 4 labelling (Jensen–Shannon distance, ↓). Per-corpus columns report Behav-JS. Average columns report Behav-JS, its change from GroupPersona-full (Δ; positive is worse), Struct-JS, Ext-Act, and Ext-E/T. Within each ablation block, bold and underline mark the largest and second-largest Behav-JS degradations.

	Behav-JS by corpus				**Average checks**				
Variant	Dialogue-B	Dialogue-A	Reddit-B	Reddit-A	Behav	Δ	Struct	Ext-Act	Ext-E/T
GroupPersona-full	0.148	0.228	0.156	0.174	0.177	–	0.167	0.112	0.068
Group-discovery ablations									
no LLM rule verifier	0.160	0.230	0.174	0.191	0.189	+0.012	0.167	0.128	0.066
k-means groups	0.182	0.248	0.194	0.205	0.207	<u>+0.030</u>	0.167	0.143	0.074
random groups	0.233	0.264	0.235	0.246	0.245	**+0.068**	0.167	0.174	0.079
Profile-conditioning ablations									
− source prefix	0.129	0.265	0.231	0.176	0.200	+0.023	0.177	0.146	0.068
− root-guidance paragraph	0.223	0.281	0.190	0.178	0.218	<u>+0.041</u>	0.194	0.166	0.072
− behavior tendencies	0.149	0.265	0.174	0.155	0.186	+0.009	0.154	0.135	0.069
− structural statistics	0.256	0.276	0.216	0.224	0.243	**+0.066**	0.221	0.173	0.067

Table 7: Quality calibration to reference conversations. Avg. averages the eight judge dimensions; MAD measures deviation from the reference-conversation quality profile. Full scores are in Table 16.

Method	**Avg. score**	**MAD↓**
Reference	*6.92*	*0.00*
AutoPAL	5.03	1.89
ConceptPersona	8.31	1.39
DiaSynth	8.69	1.77
FaithfulPersona	8.07	1.15
ICL	6.01	<u>0.91</u>
Interlocutor	5.44	1.49
PersonaLens	8.49	1.57
GroupPersona	7.45	**0.63**

The full profile is most consistent. No single ablation improves performance consistently across corpora or evaluation axes. Each component contributes a different signal, and using them together produces the most reliable alignment overall.

6.4 Q4. Does GroupPersona preserve reference-like conversation quality?

Q4 asks whether improved behavioral alignment comes at the cost of shifting conversation quality away from the reference corpus. Higher LLM-as-Judge scores are not necessarily better because synthetic conversations can become overly polished. We therefore compare each method's 8-dimensional quality profile with that of the reference conversations using mean absolute deviation (MAD). Lower MAD indicates closer agreement with the reference. Full rubric and per-corpus results are provided in Appendix P.

GroupPersona is closest to the reference quality profile. GroupPersona achieves the lowest MAD at 0.63, compared with 0.91 for ICL, the closest baseline. Several persona-based baselines score above the reference range, while AutoPAL and Interlocutor fall below it. GroupPersona therefore improves behavioral fidelity without making the synthetic corpus systematically more polished or lower quality than the reference conversations.

7 Conclusion

We introduced GroupPersona, a framework for addressing behavior-distribution misalignment in synthetic dialogue corpora. While persona-grounded generators can produce plausible individual conversations, they do not necessarily preserve the behavior mix of a reference population. GroupPersona addresses this gap by turning recurring population behavior patterns into prevalence-weighted group controls for user-agent synthesis.

Across four PopAlign-Bench corpora, GroupPersona reduces average Behav-JS by 24.4% relative to the best-average baseline while maintaining comparable structural alignment. Additional analyses show that the gain persists across synthesis and labelling families and that the resulting conversations remain closer to the reference quality profile. This suggests that corpus-level fidelity should be evaluated alongside the plausibility of individual synthetic conversations. Together, these results show that population-level behavioral fidelity can be explicitly measured and improved through group-conditioned generation.

Limitations

GroupPersona has several limitations. First, group discovery uses fixed thresholds, which may not transfer optimally across corpora with different levels of behavioral diversity. Second, behavior labelling and rule verification rely on strong LLMs, although we validate the labeller against human annotations and test robustness across model families. Third, PopAlign-Bench covers English task-oriented, knowledge-grounded, and social-discussion data, so other languages, domains, and modalities may require schema adaptation. Finally, the quality-calibration analysis in §6.4 is a secondary diagnostic based on a single judge family and does not establish downstream training or deployment utility.

Ethical Considerations

We use the source datasets under their existing licenses. Reddit discussions are de-identified before labelling by removing URLs and user mentions, and the synthetic dialogues do not use private user data. LLM-based rule verification and evaluation are performed offline.

Because population alignment can preserve undesirable patterns present in the reference corpus, GroupPersona may also reproduce source-corpus biases. We report behavior dimensions separately so that such distributional patterns remain inspectable.

References

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., USA. Association for Computing Machinery.

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile. Morgan Kaufmann.

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. Ph.d. thesis proposal, Carnegie Mellon University, Pittsburgh, PA.

Dan Bohus and Alexander I. 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, Lisbon, Portugal. Special Interest Group on Discourse and Dialogue (SIGdial).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yi Cheng, Wenge Liu, Kaishuai Xu, Wenjun Hou, Yi Ouyang, Chak Tou Leong, 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. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35212–35290, Suzhou, China. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2020. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15245–15270, Bangkok, Thailand. Association for Computational Linguistics.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Mingjin Li, Yu Liu, Huayi Liu, Xiang Ye, Chao Jiang, Hongguang Zhang, and Yu Ruan. 2025a. MADS: Multi-Agent Dialogue Simulation for Diverse Persuasion Data Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 399–415, Suzhou (China). Association for Computational Linguistics.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. 2025b. Toward Multi-Session Personalized Conversation: A Large-Scale Dataset and Hierarchical Tree Framework for Implicit Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11506, Suzhou, China. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Daniela Occhipinti, Marco Guerini, and Malvina Nissim. 2025. When Harry Meets Superman: The Role of The Interlocutor in Persona-Based Dialogue Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17964–17985, Vienna, Austria. Association for Computational Linguistics.

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 Eng Siong 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, Albuquerque, New Mexico. Association for Computational Linguistics.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know You First and Be You Better: Modeling Human-Like User Simulators via Implicit Profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

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*.

Bohao Yang, Dong Liu, Chenghao Xiao, Kun Zhao, Chen Tang, Chao Li, Lin Yuan, Guang Yang, and Chenghua Lin. 2025. Crafting Customisable Characters with LLMs: A Persona-Driven Role-Playing Agent Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20216–20240, Suzhou, China. Association for Computational Linguistics.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18023–18055, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Behavioral schema: provenance and value vocabulary

The 12 behavioral dimensions span three coverage axes informed by prior work. *Intent dynamics* (primary intent type, interaction mode, user type) are motivated by intent-oriented personal-assistant architectures (Sarikaya, 2017); *interaction management* (response brevity, user goal profile, persistence level, topic cohesion, politeness strategy, interaction flexibility) is informed by work on adaptive personal agents and multi-dimensional dialogue evaluation (Cheng et al., 2024; Finch and Choi, 2020; Mehri and Eskenazi, 2020); *user–system dynamics* (persona adaptability, repair behavior, error recovery style) is informed by spoken-dialogue error awareness and recovery (Bohus, 2004; Bohus and Rudnicky, 2005). Each dimension is observable at the conversation level, discrete with a 2–7-value vocabulary, and domain-agnostic. The full per-dimension definitions, value vocabularies, and tie-break rules used by the LLM labeller are reproduced below.

12-dimension per-record labelling prompt (full)

Role. You are an expert in conversation behaviour analysis. Given one complete user–assistant dialogue, assign exactly one categorical label per dimension. The labels are aggregated across a large corpus for distributional analysis, so consistency of interpretation matters more than per-dialogue expressiveness.

Input. A single dialogue with turns alternating between `user:` and `assistant:`.

Critical constraints. (1) Output *only* a single valid JSON object; no Markdown, no prose, no trailing commas. (2) Use exactly the twelve keys listed below; no missing or extra keys. (3) Each value must be one of the allowed options for that key, or the sentinel `"unknown"` permitted by (4). (4) If a label cannot be reliably inferred from the conversation alone, output `"unknown"`. (5) Rely on observable behaviour; do not hallucinate user intent or demographics.

Dimension definitions and value vocabularies.
1. `primary_intent_type` ∈ {`General_Task`, `Info-seeking`, `Chitchat`, `Music`, `Reminder`, `Alarm`, `Smart_Home`}. The dominant intent category across the user's turns; ties are resolved by the opening request.
2. `interaction_mode` ∈ {Command-style, QA-style, Multi-turn_Dialog-style}. Command = imperative one-shot; QA = question → answer pattern; Multi-turn = ≥ 3 back-and-forth turns building on prior context.
3. `user_type` ∈ {Power_User, One-shot_User, Tasker, Casual}. Power = exploits advanced options; One-shot = ends after one exchange; Tasker = follows a task plan; Casual = informal, low engagement.
4. `response_brevity` ∈ {Short, Medium, Long}. Median user-utterance word count: Short ≤ 6, Medium 7–20, Long ≥ 21.
5. `user_goal_profile` ∈ {Tasker, Explorer, Goal-switcher, Chatter}. Tasker = single goal pursued to completion; Explorer = probing/learning; Goal-switcher = ≥ 2 goal changes; Chatter = social, no goal.
6. `persistence_level` ∈ {High, Medium, Low}. After a failed/unhelpful system reply: High = retries $\geq 3\times$; Medium = 1–2 retries; Low = abandons.
7. `topic_cohesion` ∈ {High, Medium, Fragmented}. High = one topic; Medium = related topics; Fragmented = ≥ 2 unrelated topics.
8. `politeness_strategy` ∈ {Direct, Polite, Formal, Friendly}. Direct = imperative, no marker; Polite = "please/thanks"; Formal = third-person/honorifics; Friendly = casual greetings, emotive language.
9. `interaction_flexibility` ∈ {Adaptive, Static}. Adaptive = reformulates after misunderstanding; Static = repeats verbatim or gives up.
10. `persona_adaptability` ∈ {High, Low}. High = tone/style shifts to match system; Low = constant tone regardless.
11. `repair_behavior` ∈ {Rephrase, Clarify, Ignore, Retry}. After a breakdown: Rephrase = restates differently; Clarify = adds info / asks back; Ignore = moves on; Retry = repeats identically.
12. `error_recovery_style` ∈ {Clarification_Request, Repetition, Restart, Task_Abandonment}. Dominant strategy when the system errs: ask for clarification, repeat, restart conversation, or abandon the task.

Output format.

```
{
  "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": "..."
}
```

Tie-breaking. If two values are equally supported, pick the one earlier in the value vocabulary. If the dialogue is too short (< 4 turns) to infer a dimension, return `"unknown"` for that key rather than guess.

B PopAlign-Bench corpus construction and reference floors

The benchmark follows the $2{\times}2$ design in §5: source family (dialogue datasets vs. Reddit discussions) crossed with construction regime (Basic vs. Advanced). The source-family axis changes the user population; the regime axis changes the surface realisation of that population. All four cells share the same schema, scoring code, reference-floor estimator, and released source-history identifiers.

Source-family axis. The **Dialogue** family combines MultiWOZ-2.2 (Zang et al., 2020) task-oriented dialogues with Wizard-of-Wikipedia (Dinan et al., 2019) knowledge-grounded chitchat in a fixed 70/30 split. This mixture covers goal-driven information seeking and casual knowledge-grounded exchange. The **Reddit** family is built from Pushshift Reddit (Baumgartner et al., 2020) threads in *AskReddit*, *askphilosophy*, *advice*, *NoStupidQuestions*, and *relationship_advice*. Each conversation is one top-level post followed by the three-deepest top-rated comment chain, with handles, URLs, and subreddit names stripped. This family covers longer, less task-scripted opinion exchange and advice-seeking. The train-train cross-family Behav-JS is 0.15–0.16 (Tab. 8), roughly five to seven times the within-corpus reference floor, so averaging across the two families does not hide family-specific failures.

Construction regime. The **Basic** regime keeps lightly normalised source dialogues and uses them directly as the reference. The **Advanced** regime re-renders every turn of the same source dialogue with a Claude Sonnet 4 diversity-promoting rewrite prompt. The prompt preserves speaker identity, intent, and information content per turn, while varying phrasing, vocabulary, politeness markers, and turn length within reasonable bounds. Empirically, the rewriter changes the structural distribution (Struct-JS Basic–Advanced $= 0.31$) but leaves the behavioral distribution close to the within-corpus floor (Behav-JS Basic–Advanced $= 0.022$). We therefore use Advanced variants as controlled stress tests for surface-form drift, not as claims about naturally occurring population samples. The dataset-transformation prompt is reproduced below.

Dataset-to-TOD transformation prompt (full)

This prompt converts a source dialogue (MultiWOZ / Wizard-of-Wikipedia / Pushshift Reddit thread) into a task-oriented user–assistant dialogue suitable for behavioral labelling. It runs in two sequential steps. Step 1 extracts the semantic core; Step 2 regenerates the dialogue under a target behavioural profile (one profile for the *Basic* variants, four profiles for the *Advanced* variants).

Step 1 — Semantic core extraction.
Role. You are an expert conversation analyst. Given a raw dialogue (or a Reddit thread that has already been pre-processed to remove URLs, user mentions, and subreddit references), extract the semantic core that any new conversation on the same topic must preserve.
Output (strict JSON).

```
{
  "core_user_need": "...",
  "key_information": ["..."],
  "task_sequence": ["..."],
  "domain": "...",
  "critical_details": ["..."]
}
```

- `core_user_need`: the fundamental goal the user is trying to accomplish.
- `key_information`: essential facts, named entities, and context.
- `task_sequence`: ordered list of main sub-objectives.
- `domain`: subject area (e.g. travel, cooking, technical support).
- `critical_details`: specific constraints, preferences, or numeric values.

Step 2 — Behavioural profile instantiation.
Role. You are a conversation generator. Given the semantic core from Step 1 and a target behavioural profile, produce a new user–assistant dialogue that faithfully instantiates the profile.
Constraints (apply to every variant). (1) Generate 10–40 alternating turns, labelled `user:` and `assistant:`. (2) Preserve every fact in the semantic core. (3) Do not copy more than 5 contiguous words from the source dialogue. (4) Each turn is natural-language, no system tokens.
Profiles.

Basic — short-request baseline. User turns are concise (1–2 sentences), task-focused, direct, no courtesy markers. Assistant responds context-appropriately.
Advanced (four profiles, one dialogue per profile per semantic core).

- *(a) Short-request, unhelpful user.* User repeatedly expresses dissatisfaction and resists assistant suggestions before grudgingly accepting one.
- *(b) Short-request, topic shift.* User makes at least two abrupt unrelated topic changes mid-conversation, returning briefly to the main task only at the end.
- *(c) Short-request, early termination.* User abandons the task before completion with an explicit closing signal (e.g. "nevermind, forget it").
- *(d) Short-request, baseline.* Standard short-request behaviour as in Basic.

Reddit specifics. For Reddit-Basic / Reddit-Advanced, Step 1 runs on the pre-processed thread (post body + top-3 comment chain with URLs, handles, and subreddit names stripped). Step 2 is identical to the Dialogue variants.

Output for both steps is strict JSON; no surrounding text.

Per-cell sizing, splits, and reference floors. Each corpus contains 1,000 training conversations and 500 test conversations with fixed source-history identifiers. The train split supplies persona, example, cluster, and pattern information to generators; the test split is the JS reference. Reference–reference distance is computed between the reference marginals of the training and test splits of the same corpus. The Behav-JS floors are $0.030/0.024/0.029/0.023$, and the Struct-JS floors are $0.076/0.087/0.094/0.086$ for Dialogue-Basic / Dialogue-Advanced / Reddit-Basic / Reddit-Advanced, respectively.

Cross-corpus JS as scale anchors. Table 8 reports JS between the train splits of the four corpora. Behavioral JS between Basic and Advanced variants of the same source is 0.022, close to the within-corpus floor. Behavioral JS between the dialogue datasets and Reddit discussions is 0.15–0.16, which gives the natural scale of two distinct source populations. The cross-population values provide a scale anchor rather than a mathematical ceiling; GroupPersona's headline Behav-JS values are reported relative to both this anchor and the within-corpus floor.

Table 8: Cross-corpus JS between train splits. Only the upper triangle is shown because JS distance is symmetric. Behav-JS stays near the reference floor for same-source Basic/Advanced variants and is much larger across the dialogue datasets and Reddit discussions; Struct-JS is more sensitive to surface-form changes. Bold marks Basic/Advanced comparisons within the same source family.

	Dia-B	**Dia-A**	**Red-B**	**Red-A**
Behav-JS, train↔train				
Dia-B	—	**0.022**	0.160	0.158
Dia-A	–	—	0.150	0.151
Red-B	–	–	—	**0.022**
Red-A	–	–	–	—
Struct-JS, train↔train				
Dia-B	—	**0.313**	0.716	0.739
Dia-A	–	—	0.718	0.699
Red-B	–	–	—	**0.308**
Red-A	–	–	–	—
Reference	*Behav 0.023–0.030; Struct 0.076–0.094*			

C Baseline implementation details

All baselines are adapted to the same PopAlignBench input-output format. The synthesis LLM is held fixed to Claude Sonnet 4 across baselines and GroupPersona; the same per-corpus train split supplies persona, example, or profile information, and the same test split is used as the JS reference. Thus differences in JS reflect how each method represents and conditions on the source population rather than which backbone or evaluator is used. The baseline conditioning setups are described below; the GroupPersona synthesis prompt is reported in App. O.

AutoPAL-inspired (Cheng et al., 2024). We adapt AutoPAL's idea of conditioning persona adaptation on user interactions into a static persona-conditioned synthesis baseline. Each synthesis record is conditioned on an automatically generated long-form user persona. In our adaptation, the synthesis LLM first writes a one-paragraph free-text profile covering preferences, topics of interest, and conversational habits, then uses it as the user-agent system prompt; we use Claude Sonnet 4 for parity.

ConceptPersona-inspired (Kim et al., 2023). We adapt CPX's concept-based persona-expansion idea to the benchmark's behavior vocabulary. In our adaptation, each user persona is constructed by retrieving conceptually related attributes from an attribute pool, where the pool is the union of distinct values observed in the train split's 12-dimensional behavioral labels. This gives the baseline access to the same label vocabulary as GroupPersona, but it produces per-record attribute bags rather than

per-group root signatures.

DiaSynth-inspired (Suresh et al., 2025). We adapt DiaSynth's topic–subtopic–persona conditioning pipeline. In our adaptation, a topic-and-subtopic taxonomy is mined from the train split by clustering free-text topic labels into 30 topic and 90 subtopic cells. Synthesis is conditioned on a sampled topic-subtopic pair and a one-line user persona, with the topic seed drawn proportionally to the train-set topic marginal.

FaithfulPersona-inspired (Jandaghi et al., 2024). We adapt its generator–critic filtering principle to persona construction. In our adaptation, a generation pass writes a candidate short-text persona for one source prefix, and a verifier pass accepts only personas above a fixed consistency threshold. Both generator and verifier use Claude Sonnet 4 for parity.

ICL (Brown et al., 2020). Four randomly drawn train conversations are placed in the synthesis prompt as in-context examples. No explicit persona or structured conditioning is provided, so this serves as a lower-information example-based baseline.

Interlocutor-inspired (Occhipinti et al., 2025). We adapt Interlocutor's explicit conditioning on both the target speaker's persona and the interlocutor's profile. The user agent is conditioned on a single user persona, and the assistant agent is conditioned on a separately mined assistant persona; the two agents then run a multi-turn dialogue without a shared group identifier. Both personas are mined from the train split for this benchmark adaptation.

PersonaLens-inspired (Zhao et al., 2025). We adapt PersonaLens's profile-conditioned user-simulation setup to PopAlign-Bench. In our adaptation, each synthesis profile uses five fields—occupation, domain, expertise, tone, and intent—drawn from train marginals, and synthesis is conditioned on this profile plus a sampled task slot. These five fields are our benchmark adaptation rather than PersonaLens's original profile schema.

D Conditioning differences on a shared test history

This appendix expands the qualitative example in §6.1. It traces one fixed test history through GroupPersona and every baseline, showing what each method attaches to that history before synthesis and how the resulting user behavior compares with the reference continuation. Every method receives the identical opening:

```
user: Need cheap restaurant downtown.
assistant: I can help you find an
affordable restaurant downtown. What
type of cuisine are you interested in?
```

The reference continuation for this history is labelled command-style interaction, a task-focused user goal, short responses, and direct phrasing, with a median user utterance of about two words.

Table 9 summarizes what each method passes to the user agent and which behavior dimensions then diverge from the reference. GroupPersona alone carries a training-population behavior mix; the other methods condition on per-record attributes, personas, plans, or examples, allowing interaction style to drift.

Full continuations. The three continuations below are taken from our experiment records and relabelled by the scoring pipeline used in the main results.

Continuations of the same fixed history

Reference.

```
user: Italian works.
assistant: I found Mama's Kitchen, an Italian
restaurant downtown with budget-friendly
prices. They serve classic pasta dishes and
pizzas at reasonable rates.
user: What's the address?
assistant: Mama's Kitchen is located at 412
Pine Street in the downtown area.
user: Phone number?
assistant: Their phone number is (555)
847-2901.
user: Thanks.
assistant: You're welcome! Enjoy your meal at
Mama's Kitchen.
```

Relabelled: command-style, task-focused, short responses, direct phrasing; median user utterance $\approx$ 2 words.

GroupPersona.

```
user: Italian.
assistant: Mama's Kitchen on Pine Street is
an affordable Italian spot downtown. Want the
address or phone?
user: Address.
assistant: 412 Pine Street, downtown.
user: Phone?
assistant: (555) 847-2901.
user: Thanks.
assistant: You're welcome — enjoy your meal!
```

Relabelled: command-style, task-focused, short responses, direct phrasing; median user utterance $\approx$ 1–2 words.

PersonaLens.

```
user: I'd like Italian, something not too
expensive.
assistant: Mama's Kitchen downtown fits —
budget-friendly Italian.
```

Table 9: **What each method attaches to the same fixed test history.** The reference continuation for this history is command-style, task-focused, short, and direct. Only GroupPersona conditions on a behavior mix estimated from the training population; the remaining methods condition on signals derived from the history or sampled per record.

Method	Conditioning passed to the user agent	Dimensions that diverge from the reference	User-turn length
GroupPersona	The labels read off the history are linked to a train-derived group profile: command-style interaction and a task-focused goal as roots; short responses, high topic cohesion, and direct phrasing as tendencies; about five turns with two- to four-word user utterances as structural guidance.	**none**	**stays short**
ConceptPersona	The same history labels are expanded into a per-record attribute bag: task-focused goal, medium response length, low persistence, high topic cohesion, adaptive interaction.	brevity, politeness	longer
FaithfulPersona	The history labels are rewritten as a short free-text profile: asks for things directly and briefly, wants quick practical answers, does not chat beyond the task.	brevity, politeness	longer
PersonaLens	A structured identity-and-task record is sampled: office worker, restaurant domain, novice expertise, direct tone, and the task of finding a cheap nearby place.	interaction mode, goal, brevity, politeness	longer
AutoPAL	A long-form persona is written before synthesis: prefers quick local food help, likes trying new places, may ask about dishes, ambience, and options.	interaction mode, goal, politeness	longer, unstable
Interlocutor	Biography-style personas are created for both agents: a chatty, exploratory user interested in discovering local restaurants, and an enthusiastic local-food guide.	interaction mode, goal, brevity, politeness	longer
DiaSynth	A topic-driven plan is built: dining out, budget restaurant search, an inexpensive-restaurant seeker, with casual, curious, and extended traits.	interaction mode, goal, brevity, politeness	longer
ICL	Four raw training conversations are placed in the prompt, with no labels, persona, or length target.	interaction mode, politeness	varies with the drawn examples

```
user: Can you tell me where it is and how to
reach them?
assistant: 412 Pine Street, downtown; phone
(555) 847-2901.
user: Do they take reservations for tonight,
and is there parking nearby?
assistant: I'd recommend calling them directly
to confirm both.
user: Okay, I'll do that. Thanks for the help.
assistant: You're welcome!
```

Relabelled: QA-style, exploratory goal, medium responses, polite phrasing; median user utterance ≈ 9.5 words.

The history is identical in all three cases; the difference comes from the signal attached to it, a prevalence-selected population profile for GroupPersona against an independently sampled per-record profile for PersonaLens.

E Full Q1 distribution-alignment results

Table 10 expands Table 3 by reporting per-corpus Struct-JS, Ext-Act, and Ext-E/T in addition to Behav-JS. The main text keeps per-corpus Behav-JS and average checks in Table 3 to make the primary comparison readable.

F Full Q2–Q3 ablation results

Table 11 expands Table 6 by reporting per-corpus Struct-JS, Ext-Act, and Ext-E/T for each ablation. The variant names follow Table 6: group-discovery ablations test the discovered grouping mechanism, and profile-conditioning ablations test synthesis-time inputs.

G Bootstrap confidence intervals

Per-corpus bootstrap 95% CIs on Behav-JS are computed by resampling synthesis and test records at the conversation level (200 iterations, percentile method). Table 12 reports GroupPersona-full versus the best baseline per (corpus, labelling-family) cell. The GroupPersona CI is disjoint from the best-baseline CI on Dialogue-Basic and Reddit-Advanced under both families. On Dialogue-Advanced under Claude Sonnet 4 the displayed intervals meet at the rounded endpoint 0.241; the remaining Dialogue-Advanced and Reddit-Basic intervals overlap, consistent with the smaller absolute gaps in §6.1.

Table 10: Full per-corpus distribution-alignment results under Claude Sonnet 4 labelling (Jensen–Shannon distance, ↓). Average results are reported in Table 3; this table expands the per-corpus structural and external checks. Bold and underline mark best and second-best non-reference values in each column.

	Dialogue-Basic				Dialogue-Advanced				Reddit-Basic				Reddit-Advanced			
Method	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T
AutoPAL	0.244	0.172	0.157	0.074	0.256	**0.179**	0.183	0.052	0.187	**0.148**	0.163	0.084	0.274	0.182	0.165	0.081
ConceptPersona	0.236	**0.162**	0.137	0.083	0.252	0.200	0.163	0.053	0.181	0.164	0.172	**0.063**	0.269	0.182	0.144	0.082
DiaSynth	0.247	0.169	0.138	0.081	0.259	0.211	0.161	0.053	0.162	0.156	0.171	0.069	0.288	0.169	0.152	0.075
FaithfulPersona	0.235	0.164	0.140	**0.069**	0.263	0.188	0.153	0.053	**0.156**	0.150	0.146	0.069	0.280	0.176	0.140	0.095
ICL	0.327	0.250	0.238	0.071	0.323	0.215	0.264	0.067	0.310	0.238	0.262	0.093	0.316	0.220	0.229	0.109
Interlocutor	0.232	0.175	0.171	0.089	0.254	0.190	0.209	0.074	0.265	0.197	0.178	0.080	0.284	0.169	0.188	0.080
PersonaLens	0.245	0.175	0.159	0.076	0.271	0.187	0.182	0.047	0.210	0.176	0.170	0.067	0.284	0.166	0.160	0.075
GroupPersona (ours)	**0.148**	0.177	**0.115**	0.076	**0.228**	0.194	**0.120**	**0.043**	**0.156**	0.170	**0.109**	0.086	**0.174**	**0.125**	**0.103**	**0.067**
Reference (reference–reference)	*0.030*	*0.076*	—	—	*0.024*	*0.087*	—	—	*0.029*	*0.094*	—	—	*0.023*	*0.086*	—	—

Table 11: Full per-corpus GroupPersona ablation results under Claude Sonnet 4 labelling (Jensen–Shannon distance, ↓). Average results are reported in Table 6; this table expands the per-corpus structural and external checks. Lower is better.

	Dialogue-B				Dialogue-A				Reddit-B				Reddit-A			
Variant	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T	Behav	Struct	Ext-Act	Ext-E/T
GroupPersona-full	0.148	0.177	0.115	0.076	0.228	0.194	0.120	0.043	0.156	0.170	0.109	0.086	0.174	0.125	0.103	0.067
Group-discovery ablations																
no LLM rule verifier	0.160	0.177	0.125	0.066	0.230	0.194	0.143	0.042	0.174	0.170	0.132	0.076	0.191	0.125	0.111	0.082
k-means groups	0.182	0.177	0.149	0.088	0.248	0.194	0.142	0.045	0.194	0.170	0.153	0.086	0.205	0.125	0.128	0.076
random groups	0.233	0.177	0.160	0.091	0.264	0.194	0.185	0.051	0.235	0.170	0.185	0.080	0.246	0.125	0.164	0.093
Profile-conditioning ablations																
− source prefix	0.129	0.126	0.127	0.072	0.265	0.197	0.155	0.051	0.231	0.219	0.166	0.082	0.176	0.165	0.137	0.067
− root-guidance paragraph	0.223	0.191	0.153	0.057	0.281	0.229	0.185	0.057	0.190	0.209	0.169	0.093	0.178	0.148	0.158	0.080
− behavior tendencies	0.149	0.140	0.127	0.072	0.265	0.204	0.152	0.046	0.174	0.161	0.143	0.063	0.155	0.110	0.117	0.094
− structural statistics	0.256	0.232	0.169	0.069	0.276	0.238	0.191	0.031	0.216	0.210	0.184	0.086	0.224	0.204	0.146	0.080

Table 12: Conversation-level bootstrap 95% CIs for Behav-JS (200 resamples). *GA*: GroupPersona-full. *Best BL*: best baseline per (corpus, family) cell. ◇ marks (corpus, family) cells where the GA and best-baseline CIs are disjoint (separation at the 95% level).

Corpus	Family	GA 95% CI	Best BL 95% CI
Dialogue-Basic	Claude Sonnet 4	[0.140, 0.166]◇	[0.223, 0.246]
	DeepSeek-R1	[0.149, 0.178]◇	[0.225, 0.246]
Dialogue-Advanced	Claude Sonnet 4	[0.213, 0.241]	[0.241, 0.264]
	DeepSeek-R1	[0.229, 0.260]	[0.243, 0.263]
Reddit-Basic	Claude Sonnet 4	[0.147, 0.172]	[0.149, 0.170]
	DeepSeek-R1	[0.154, 0.178]	[0.161, 0.184]
Reddit-Advanced	Claude Sonnet 4	[0.167, 0.191]◇	[0.260, 0.279]
	DeepSeek-R1	[0.168, 0.192]◇	[0.259, 0.283]

H External scoring axes

Metric definition. The two external columns in Tables 3 and 6 summarize JS over four axes from prior dialogue-annotation schemes. **Ext-Act** averages MIDAS and DailyDialog dialogue-act JS; because these act spaces overlap our schema through intent, interaction mode, and politeness-related behavior, it serves as an external interaction-behavior check. **Ext-E/T** averages DailyDialog emotion and topic JS, which are outside our behavior schema and serve as drift checks. None of these axes is used for group construction or synthesis.

Source taxonomies. MIDAS (Yu and Yu, 2021) defines 23 per-user-turn functional and social acts, including statements, opinions, complaints, commands, open- and yes/no questions, thanks, apologies, and closings; it overlaps our interaction-mode, primary-intent, and politeness dimensions. GroupPersona's cross-corpus average on MIDAS is 0.132 versus 0.159 for the best-average baseline. DailyDialog (Li et al., 2017) provides a 4-way act axis (inform, question, directive, commissive), a 7-way emotion axis, and a 10-way topic axis. Its act labels partially overlap intent and interaction mode, while emotion and topic are non-degradation checks that no benchmark baseline directly conditions on.

Extraction protocol. For every corpus-method cell, both synthetic and test conversations are labelled under each external schema using a Claude Sonnet 4 labeller with the published rubric and value vocabulary. MIDAS labels are assigned per user turn; DailyDialog labels are assigned as dominant conversation-level act, emotion, and topic. We compute empirical marginals over the relevant unit and JS between synthetic and test marginals, then average the two act distances for Ext-Act and the DailyDialog emotion and topic distances for

Ext-E/T. As a calibration check, relabelling a 200-conversation subset with DeepSeek-R1 changes per-axis JS by at most 0.012.

I Per-dimension behavioral JS radar

Figure 3 disaggregates the Behav, Struct, Ext-Act, and Ext-E/T columns of Table 3. Relative to FaithfulPersona and averaged over corpora, GroupPersona's largest behavioral improvements occur on `response_brevity`, `politeness_strategy`, and `error_recovery_style`; `persona_adaptability` is approximately tied and slightly worse for GroupPersona. The identity-style dimensions are already approximately matched by strong baselines. On structural axes, methods sit in a tight band. On external axes, GroupPersona improves the partially overlapping MIDAS and DailyDialog act axes, while DailyDialog emotion and topic remain close across methods.

J Group-discovery diagnostics and locked configuration

This appendix reports the implementation details behind Stage 2, including rule thresholds, minimal-rule pruning, LLM verification, reduced-signature construction, greedy Jaccard clustering, root admission, residual handling, and diagnostics.

Group-discovery normalization. Before rule mining, the group-discovery pipeline merges a small number of semantically adjacent long-tail values: Reminder/Alarm/Smart_Home become Device-Action; Casual/One-shot_User become Casual_or_Oneshot; Explorer/Goal-switcher become Explorer_or_Switcher; Polite/Formal become Polite_or_Formal; Ignore/Retry become Ignore_or_Retry; and Restart/Task_Abandonment become Restart_or_Abandon. This normalization is used only for group discovery; all reported Behav-JS results use the original vocabularies in App. A.

Verified rules. Each dialogue starts as a set P_x of 12 behavior attribute–value pairs. We mine association rules over the 800-conversation group-discovery set with antecedent arity 1–3. A rule predicts a target behavior pair $\ell_t = (a = v)$ from antecedent pairs A, written as $A \Rightarrow \ell_t$. Candidate rules must satisfy confidence ≥ 0.80, lift ≥ 1.3, and support ≥ 0.03. They are scored by

$$\mathrm{score}(r) = \mathrm{conf}(r) \log_2 \mathrm{lift}(r) \sqrt{\mathrm{support}(r)}.$$

For each multi-pair antecedent, we backward-prune to a minimal sufficient subset $A^\star$. An antecedent pair ℓ is removed from A when its deletion reduces confidence by at most $\delta = 0.05$:

$$\mathrm{conf}(A \Rightarrow \ell_t) - \mathrm{conf}(A \setminus \{\ell\} \Rightarrow \ell_t) \leq \delta.$$

Pruning that leaves confidence unchanged or higher therefore also qualifies. Rules that collapse to the same $(A^\star, \ell_t)$ are deduplicated. Each deduplicated minimal rule is then judged by the LLM verifier described in App. K. Only accepted rules can mark a target pair as derivable.

Reduced signatures. Accepted rules are applied independently within each dialogue. Rules are scanned in increasing antecedent arity and applied to a fixed point. Starting from $R_x^{(0)} = P_x$, a target ℓ_t is removed only when $\ell_t \in R_x^{(t)}$ and $A^\star \subseteq R_x^{(t)} \setminus \{\ell_t\}$. The scan then restarts on the reduced set, and the fixed point is the clustering signature R_x. The size of R_x can vary across dialogues because different rules fire in different local contexts.

Verifier reliability is evaluated in Appendix L.

Greedy Jaccard clustering. Dialogues are clustered in the reduced-signature space using a deterministic single-pass procedure. Among unassigned dialogues, we select the most frequent remaining reduced signature as the seed R_s. The cluster absorbs every unassigned dialogue x whose reduced signature satisfies

$$J(R_x, R_s) = \frac{|R_x \cap R_s|}{|R_x \cup R_s|} \geq \tau_{\text{jacc}}.$$

Absorbed dialogues are removed from the pool, and the process repeats until all dialogues are assigned. Ties are resolved using a fixed ordering.

Root admission and residual handling. After clustering, each valid cluster is named by up to $K_{\max}$ root pairs. A root pair is a retained behavior pair $\ell = (a = v)$ that is both common inside cluster C and distinctive relative to the group-discovery split:

$$\Pr[\ell \mid C] \geq \tau_{\text{hom}}, \qquad \frac{\Pr[\ell \mid C]}{\Pr[\ell]} \geq \tau_{\text{lift}}.$$

Clusters smaller than τ_{size} or without any admissible root pair are placed in the residual set. At synthesis time, each residual cluster is merged into the nearest valid group by average Jaccard similarity.

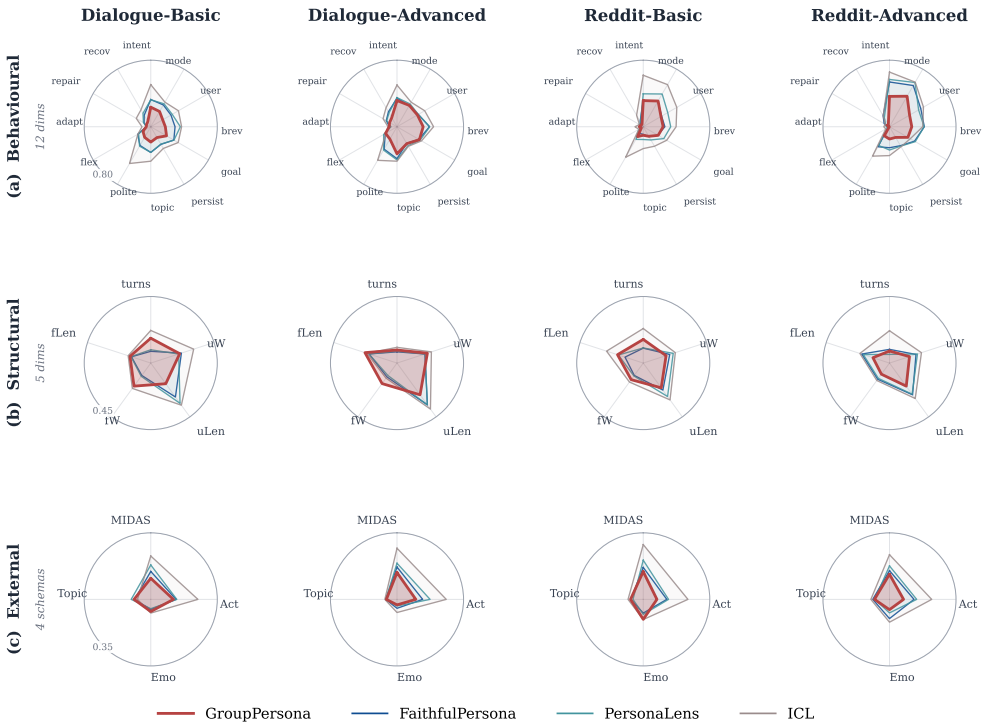

Figure 3: Per-dimension JS radar for GroupPersona-full and three representative baselines under Claude Sonnet 4. Rows show Behav-JS, Struct-JS, Ext-Act, and Ext-E/T; columns show the four corpora. Lower is better. The three rows use different radial scales (0.80, 0.45, and 0.35), so shapes should be compared within a row rather than across rows. GroupPersona's largest average behavioral gains over FaithfulPersona are on brevity, politeness, and error recovery.

Locked configuration and diagnostics. The locked configuration is

$$(\tau_{\text{hom}}, \tau_{\text{lift}}, \tau_{\text{jacc}}) = (0.60, 1.15, 0.50),$$
$$(\tau_{\text{size}}, K_{\max}) = (3, 4).$$

For GroupPersona, 200 conversations from the training split are held out for configuration selection. The remaining 800 training conversations are used for rule mining, verification, signature reduction, behavioral-group discovery, and prevalence estimation. The 500-conversation test split is used only for reported evaluation and is never used for configuration selection.

Table 13 reports the exact group counts and residual rates from these 800-conversation group-discovery sets.

Table 14 shows two actual groups from the Dialogue-Basic group-discovery set. It illustrates how compact root signatures coexist with heterogeneous but strongly skewed non-root behavior tendencies.

Table 13: Group-discovery diagnostics for each corpus's remaining 800 training conversations. Residual rate is the percentage of those conversations placed in the residual set before nearest-group reassignment at synthesis time.

Corpus	Groups	Residual (%)
Dialogue-Basic	90	9.2
Dialogue-Advanced	74	8.9
Reddit-Basic	94	9.1
Reddit-Advanced	82	9.0

K LLM judge rubric for minimal-rule verification

This appendix details the LLM verification step used in Stage 2. After backward pruning, each candidate minimal rule is written as $A^\star \Rightarrow \ell_t$, where $A^\star$ is a minimal set of antecedent behavior pairs and $\ell_t = (a = v)$ is the target behavior pair. The verifier is used only to filter mined rules before signature reduction. It does not generate rules, cluster dialogues, or score synthetic conversations.

For each minimal rule, the judge receives four inputs: the rule in plain text, its support, confidence, lift, and pattern score, three matched training ex-

Table 14: Dialogue-Basic case study of two groups discovered from the remaining 800 training conversations under the locked configuration $(\tau_{\text{hom}}, \tau_{\text{lift}}, \tau_{\text{size}}, K_{\max}) = (0.60, 1.15, 3, 4)$. Group A is the largest 4-root group; Group B is the smallest 2-root group admitted under the size floor. Roots are the (dimension, value) pairs that satisfy in-group homogeneity $\geq \tau_{\text{hom}}$ and lift $\geq \tau_{\text{lift}}$ (§4.2); the LLM judge verifies reduction rules, not roots. Non-root rows show conditional marginals $\Pr[D' \mid g]$, which act as characteristic patterns at synthesis time.

Field	**Group A (casual chitchat)**	**Group B (one-shot impatient)**
Cluster size	53 records	5 records
# of roots admitted	4	2
Root signature — the group identifier		
primary_intent_type	**Chitchat** (root)	non-root: Info-seeking 60%, Music 20%, Smart_Home 20%
interaction_mode	**Multi-turn_Dialog-style** (root)	non-root: QA-style 100%
user_type	**Casual** (root)	**One-shot_User** (root)
user_goal_profile	**Chatter** (root)	non-root: Explorer 60%, Goal-switcher 20%, Tasker 20%
response_brevity	non-root: Long 81%, Medium 11%, Short 8%	**Short** (root)
Non-root marginals — characteristic patterns at synthesis		
persistence_level	Low 53%, Medium 34%, High 13%	Medium 80%, High 20%
topic_cohesion	Medium 77%, Fragmented 17%, High 6%	High 80%, Fragmented 20%
politeness_strategy	Formal 36%, Polite 25%, Friendly 23%	Direct 60%, Polite 40%
interaction_flexibility	Adaptive 74%, Static 26%	Static 80%, Adaptive 20%
persona_adaptability	High 83%, Low 17%	Low 60%, High 40%
repair_behavior	Clarify 47%, Rephrase 23%, Ignore 19%	Rephrase 80%, Retry 20%
error_recovery_style	Restart 45%, Clarification_Req. 23%, Task_Abandon. 17%	Repetition 80%, Restart 20%

amples whose behavior-pair sets contain $A^\star$, and the parent rule before pruning. The matched examples mark whether ℓ_t is also present, allowing the judge to check whether the rule reflects a meaningful behavior dependency rather than a template artifact, near-duplicate label, or unsupported correlation. The parent rule is included to verify that pruning did not remove a semantically necessary antecedent.

The judge returns `is_reasonable` $\in \{0, 1\}$, `confidence` $\in [0, 1]$, and a brief justification. Only rules with `is_reasonable=1` can mark a target pair as derivable during signature reduction in §4.2. The full prompt, accept/reject rubric, and output format are reproduced below.

LLM pattern-judge prompt (full)

Role. You are a senior dialogue researcher reviewing one *minimal* association rule for inclusion in a group-conditioned synthesis pipeline. The rule has been backward-pruned to the smallest pair-subset whose confidence on the consequent is at most δ=0.05 below that of the parent rule. Accepted rules will be used to drop the consequent pair from a dialogue whenever the dialogue's other 11 pairs already contain the minimal antecedent. Your job is to decide whether the minimal rule still reflects a *semantically* plausible user-behaviour regularity, not merely a statistical artefact, given its corpus statistics, the parent rule, and three example conversations that match the antecedent.

Input.

- `minimal_rule`: the minimal antecedent $A^\star$ and consequent (D=d), e.g. "(`primary_intent_type=Chitchat`) AND (`user_type=Casual`) $\Rightarrow$ (`user_goal_profile=Chatter`)".
- `parent_rule`: the rule from which $A^\star$ was pruned (or the same rule if already minimal), so you can sanity-check that minimisation did not strip a semantically necessary atom.
- `stats`: `support`, `confidence`, `lift`, `pattern_score` computed on $A^\star$.
- `examples`: 3 example conversations (verbatim turns) whose 12 attribute-value pairs contain $A^\star$; for each example, also show whether the pair-set also contains the consequent.

Decision criterion. The minimal rule is *reasonable* iff (i) a reader can articulate *why* the minimal antecedent would imply the consequent (causal, definitional, or strong behavioural correlation) independent of the numbers, (ii) the minimal antecedent is not missing a semantically necessary atom that the parent carried, (iii) the statistics meet the pre-filter floors (`confidence` ≥ 0.80, `lift` ≥ 1.3, `support` ≥ 0.03; near-threshold rules deserve extra scrutiny), *and* (iv) at least 2 of the 3 example conversations visibly contain both the antecedent and the consequent.

Reject if any of the following.

- *Sampling artefact*: lift barely above 1; consequent is the marginal mode anyway.
- *Tautology*: consequent is a rephrasing of the antecedent on the same dimension.
- *Example contradiction*: conversations matching the antecedent do not exhibit the consequent.
- *Generic-prior consequent*: consequent is `unknown` or a default category and the lift is driven by that prior.
- *Template-induced co-occurrence*: the link is mechanically forced by the corpus generation pipeline (e.g. "`interaction_mode=QA-style` $\Rightarrow$ `turn_count=2`" in a synthesised QA dataset).

Output (strict JSON; no surrounding text).

```
{
  "is_reasonable": true|false,
  "confidence": 0.0,
  "reasoning": "..."
}
```

The `confidence` field is your self-reported confidence in the binary decision; it is recorded for analysis and is not used for rule filtering. The `reasoning` field must be one short sentence that cites the concrete signal driving the decision (e.g. "all three examples show short imperative turns, consistent with the consequent").

L Human validation protocol

Two LLM-driven steps are validated against human judgment: per-dialogue behavior labelling and per-rule accept/reject decisions. The audit was conducted by project-internal expert annotators for research validation; no external participants or crowdworkers were recruited or paid, and no new personal data were collected. Annotators worked only from observable conversational behavior, without inferring demographics, private attributes, or unstated intent. Rule ratings use a 1–5 scale, where higher scores indicate stronger agreement; both verifier reliability and labelling agreement are reported below.

Rule-verification audit. For each of the four corpora and two generation settings (Claude Sonnet 4 and DeepSeek-R1), we sample 40 minimal rules, giving $4 \times 2 \times 40 = 320$ rules in total. Each rule is rated for whether the minimal antecedent still captures a plausible user-behavior regularity for the consequent rather than a template artifact, duplicate label, or unsupported correlation, and, where applicable, whether the parent rule's pruning kept the semantics. The reported human rating of 3.83 ± 0.12 is pooled across these rules; Claude Sonnet 4 is used as the verifier throughout.

Agreement with an adjudicated human reference. To measure the labeller directly, two annotators independently labelled all 12 dimensions for 50 conversations from each corpus, 200 in total. Human–human κ is computed before discussion; the remaining disagreements are then resolved into an adjudicated human reference, against which the default labeller is scored.

Table 15 reports per-dimension agreement. The default labeller agrees strongly with the adjudicated reference (macro-F1 0.85, $\kappa = 0.77$), close to the $\kappa = 0.82$ the two annotators reach with each other. Agreement is lowest on persistence, adaptability, and repair or recovery, which depend on failure and adjustment events that occur in only part of the data, but remains substantial on every dimension.

LLM-as-Judge prompt for reference-conversation-anchored quality rating

Role. You are a senior dialogue-systems researcher rating one candidate conversation on standard conversation-quality dimensions. Use the target-corpus examples to calibrate what score levels are typical for this corpus rather than silently replacing them with an idealised standard. Do not directly score similarity to the examples; assign the candidate's level on each quality dimension.

Input.

- `candidate_conversation`: the full multi-turn transcript to rate.
- `corpus_description`: a one-paragraph description of the reference corpus (domain, register, expected user behavior, interaction style).
- `reference_examples`: three reference conversations randomly sampled from the reference corpus, included verbatim so the judge can anchor against "what the reference corpus looks like" rather than an idealised prior.

Calibration note in the prompt. The prompt explicitly states that reference conversations are imperfect — they contain hesitation, mid-stream topic changes, false starts, repair, persistence, mood shifts, and varied turn lengths — so the judge should not automatically assign maximum scores to unusually polished conversations. Candidate and reference conversations are rated with the same descriptive scale; calibration is assessed only after scoring.

Dimensions (1–10, higher = more of the named quality property).

- **Flow**: the degree of smooth, natural progression and plausible pacing.
- **H-Con (User Consistency)**: the degree to which the user maintains a coherent behavior pattern (verbosity, politeness, persistence) within the conversation.
- **A-Con (Assistant Consistency)**: the consistency of the assistant's style and capability.
- **Ctx (Context Utilisation)**: the degree to which later turns use information from earlier turns.
- **Turn**: the naturalness and balance of turn lengths and turn-taking rhythm.
- **Topic**: the coherence of topic introduction, maintenance, and transitions.
- **Use (Usefulness)**: the usefulness of the reached outcome, judged on the same scale for candidates and references.
- **Ovrl (Overall Quality)**: a holistic assessment of conversation quality.

Output (strict JSON).

```
{
  "flow":   <float 1.0-10.0>,
  "h_con":  <float 1.0-10.0>,
```

Table 15: **Per-dimension agreement with an adjudicated human reference.** LLM columns refer to the default labeller, Claude Sonnet 4.

Behavioral dimension	LLM–human F1	Human–human κ	LLM–human κ
Primary intent type	0.85	0.87	0.82
Interaction mode	0.89	0.86	0.83
User type	0.83	0.82	0.75
Response brevity	0.96	0.92	0.92
User goal profile	0.85	0.83	0.77
Persistence level	0.79	0.76	0.69
Topic cohesion	0.86	0.84	0.79
Politeness strategy	0.89	0.88	0.84
Interaction flexibility	0.82	0.79	0.72
Persona adaptability	0.80	0.76	0.69
Repair behavior	0.82	0.80	0.73
Error recovery style	0.79	0.74	0.67
Macro average	**0.85**	**0.82**	**0.77**

```
  "a_con":  <float 1.0-10.0>,
  "ctx":    <float 1.0-10.0>,
  "turn":   <float 1.0-10.0>,
  "topic":  <float 1.0-10.0>,
  "use":    <float 1.0-10.0>,
  "ovrl":   <float 1.0-10.0>,
  "rationale": "..."
}
```

The rationale field cites the concrete signal driving the lowest-scoring dimension.

Aggregation. For each (method, corpus) cell we sample and score 500 generated conversations, then average per dimension to obtain that cell's row in Table 7. The reference-conversation row uses the same rubric applied to reference-corpus conversations. Mean absolute deviation (MAD) from the reference-conversation row is then computed across the 8 dimensions and averaged over the four corpora.

M Example group profile

The following block shows an example enriched profile in the format passed to the user agent. It separates group identity, accompanying behavior tendencies, and structural guidance.

Example enriched group profile

Behavior roots.

- `interaction_mode = QA-style`
- `user_goal_profile = Explorer`
- `topic_cohesion = High`

Behavior tendencies.

- Users usually ask follow-up questions rather than issuing one-shot commands.
- User turns are usually medium length and remain on the same topic.
- When the assistant response is incomplete, users tend to clarify the request instead of abandoning the task.
- Politeness markers such as "please" or "thanks" appear in a minority but recurring subset of dialogues.

Group metadata enrichment prompt

This prompt produces the three conditioning ingredients consumed by the user agent at synthesis time — (ii) group_persona, (iii) conversation_patterns, and (iv) group_statistics — from one cluster of reference dialogues.

Role. You are a conversation analyst. Given a set of reference dialogues assigned to the same behavioural group g, together with g's root signature (a small tuple of (dimension, value) pairs shared by the cluster's records), produce a compact metadata object that conditions downstream synthesis.

Input. (a) Root signature for g: e.g. {`user_type=Casual`, `primary_intent=Chitchat`, `user_goal=Chatter`, `interaction_mode=Multi-turn_Dialog-style`}. (b) Up to $n \leq 30$ reference dialogues belonging to g (full text + per-record 12-dim labels + structural fields).

Output (strict JSON, three fields).

- `group_persona` (2–3 sentences) — typical goals, tone, and interaction style of users in g. Anchor every claim to evidence in the provided dialogues. Do not invent demographics.
- `conversation_patterns` (3–5 bullets) — canonical discourse flows, intent transitions, and the top-3 most informative *non-root* dimensions with their in-cluster marginal (e.g. "`response_brevity`: Long 81%, Medium 11%, Short 8%"). Patterns are transition-style, not content-specific.
- `group_statistics` — numeric summary of the structural fields (mean and standard deviation of turn count, user/full word count, user/full utterance length). Computed on the provided dialogues, not estimated.

Guidelines. Prefer concrete observed behaviours over abstract descriptors. For high-homogeneity clusters (top-4 dim agreement ≥ 0.75), the persona should foreground the root dimensions verbatim; for lower-homogeneity clusters, mark unstable dimensions with their marginal ("mostly Casual; sometimes One-shot users"). Never claim a behaviour that does not appear in $\geq 1/3$ of the provided dialogues.

Structural guidance.

- Expected dialogue length: medium, typically 6–10 turns.
- Expected user utterance length: medium, typically one or two sentences.
- Expected interaction shape: initial information request followed by clarification or refinement turns.

Source prefix.

```
user: I am trying to decide which
train route is better for a weekend
trip.
assistant: Sure. What cities are you
```

Table 16: Full quality calibration to reference conversations. Scores use an 8-dimension LLM-as-Judge rubric; MAD is mean absolute deviation from the reference-conversation profile.

Method	Flow	H-Con	A-Con	Ctx	Turn	Topic	Use	Ovrl	MAD↓
Reference	*6.66*	*7.80*	*6.42*	*6.14*	*7.58*	*6.86*	*7.28*	*6.66*	*0.00*
AutoPAL	4.80	7.08	3.84	4.10	6.52	6.04	3.76	4.12	1.89
ConceptPersona	8.24	8.32	8.10	8.34	8.48	8.74	8.16	8.10	1.39
DiaSynth	8.64	8.96	8.52	8.58	8.66	9.12	8.48	8.56	1.77
FaithfulPersona	8.10	8.38	7.92	7.86	8.24	8.30	7.92	7.84	1.15
ICL	6.02	6.72	5.74	5.68	6.36	6.18	5.62	5.78	<u>0.91</u>
Interlocutor	5.58	6.00	5.10	5.16	5.88	5.70	4.78	5.30	1.49
PersonaLens	8.44	8.78	8.58	8.24	8.48	8.66	8.32	8.42	1.57
GroupPersona	7.34	8.60	6.22	7.18	8.20	7.96	7.06	7.02	**0.63**

```
travelling between, and do you care
more about price or travel time?
```

Instruction to user agent. Generate the next user turns as a member of this group. Preserve the source prefix, follow the behavior roots as the main identity, use the tendencies as likely but not mandatory behavior, and keep the dialogue within the structural guidance.

N Metadata enrichment prompt

The metadata-enrichment step converts each discovered behavioral group into the three profile fields shown in Figure 1: behavior roots, behavior tendencies, and structural statistics. In the implementation prompt, these fields are serialized as `group_persona`, `conversation_patterns`, and `group_statistics`, respectively. The naming differs only at the prompt-template level: `group_persona` is the natural-language rendering of the group's behavior roots, `conversation_patterns` summarizes behavior tendencies, and `group_statistics` stores structural statistics.

For the group-discovery ablations in §6.2, structural guidance is held fixed to isolate the effect of group identity, while the group-conditioned behavioral metadata is recomputed for the alternative clustering.

Group-conditioned multi-agent synthesis prompt

A two-agent loop synthesises one dialogue per group g: a *user agent* conditioned on g's metadata, and a *fixed assistant agent* held identical across methods and ablations.

User-agent system prompt (full configuration; ablations remove one bracketed clause, see below).

```
You are a user in a population that interacts with a
voice assistant.
Your behaviour comes from the behavioural group
described below; stay in character throughout the
conversation.

[root] Group identifier: {root_dim1}={root_val1},
. . . , {root_dim4}={root_val4}.
[persona] {group_persona_paragraph}
[patterns] Characteristic patterns (top-3 non-root
marginals): {dim_a=val%, . . . ; dim_b=val%, . . . ;
dim_c=val%, . . . }.
[stats] Typical turn count μ ± σ, user-utterance
length μ ± σ, word frequencies as in training.
[history] Conversation seed (reference prefix):
{history_text}. Continue as the user, matching the
seed's style.

Output: only the next user turn; no role token. Stop
when the conversation is naturally finished or the
typical turn count is reached.
```

Assistant-agent system prompt (identical across all methods).

```
You are a generic voice assistant. Be helpful,
concise, and grounded in the conversation. You do
not access any external metadata about the user.
```

The explicit *[root]* values are retained in all conditioning ablations to keep group selection fixed. For the group-discovery ablations, a random or k-means cluster identifier replaces the GroupPersona identifier while structural guidance is held fixed.

Table 17: Per-corpus quality calibration to reference conversations. Values are MAD from the reference-conversation quality profile across 8 LLM-as-Judge dimensions (↓). Bold and underline mark best and second-best non-reference values.

Method	Dia-B	Dia-A	Red-B	Red-A	Avg.
AutoPAL	1.76	1.88	1.93	2.00	1.89
ConceptPersona	1.21	1.41	1.39	1.53	1.39
DiaSynth	1.54	1.78	1.75	1.99	1.77
FaithfulPersona	1.07	1.08	1.16	1.28	1.15
ICL	<u>0.74</u>	<u>0.86</u>	<u>1.02</u>	<u>1.03</u>	<u>0.91</u>
Interlocutor	1.43	1.51	1.49	1.53	1.49
PersonaLens	1.38	1.58	1.56	1.75	1.57
GroupPersona	**0.57**	**0.62**	**0.63**	**0.69**	**0.63**

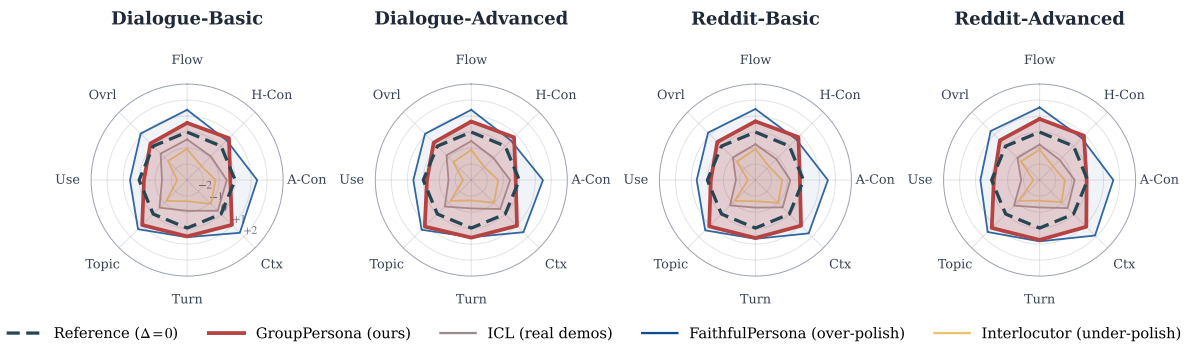

Figure 4: Per-corpus quality calibration as signed deviation from reference conversations (Δ = Method − Reference) across 8 LLM-as-Judge dimensions. The dashed zero circle marks the reference-conversation profile ($\Delta = 0$); positive values indicate over-polishing and negative values indicate under-polishing on that dimension. Unlike MAD, the signed view exposes the direction of each method's deviation from the reference profile.

O Synthesis prompt

The user-agent prompt is assembled from a fixed group index and four synthesis inputs. The group index identifies which group was selected, while the profile inputs correspond to the Figure 1 fields used to condition generation: **source prefix**, **behavior roots**, **behavior tendencies**, and **structural statistics**. In the prompt template, these are serialized as `[history]`, `[persona]`, `[patterns]`, and `[stats]`, respectively. Here, `[persona]` is the natural-language behavior-root guidance derived from the group's root signature; `[patterns]` contains non-root behavior tendencies; and `[stats]` contains structural statistics.

The conditioning-component ablations in §6.3 remove one synthesis input at a time: NO-HISTORY removes `[history]` and corresponds to − source prefix; NO-PERSONA removes `[persona]` and corresponds to − natural-language root guidance; NO-PATTERNS removes `[patterns]` and corresponds to − behavior tendencies; NO-STATS removes `[stats]` and corresponds to − structural statistics. The minimal `[root]` group identifier is retained in all profile-conditioning ablations so that group selection remains fixed; it serves only as the selected group index, not as a descriptive profile field. For the group-discovery ablations in §6.2, the selected group is replaced by a random or k-means group, while structural guidance is held fixed.

The assistant-agent system prompt is held fixed across all methods and ablations. The full prompt and ablation-to-block mapping are reproduced below.

P Conversation-quality calibration setup

This appendix documents the LLM-as-Judge rubric, prompt, and aggregation protocol used for the quality-calibration audit in §6.4.

Why anchor against reference conversations. Standard LLM-as-Judge quality rubrics can reward synthetic dialogues that are overly polished relative to reference conversations. In our audit, reference test conversations occupy a mid-range quality band rather than the top of the 1–10 scale. We therefore evaluate calibration by measuring how close each method's quality-score profile is to the reference-conversation profile under the same rubric and judge.

Dimensions. We use 8 conversation-level dimensions: Flow, Human Consistency (H-Con), Assistant Consistency (A-Con), Context utilisation (Ctx), Turn balance (Turn), Topic coherence (Topic), Usefulness (Use), and Overall conversation quality (Ovrl). Each dimension is rated on a 1–10 descriptive quality scale, where higher indicates more of the named quality property. Reference examples calibrate what quality levels are typical for the corpus; closeness is computed afterward as MAD between the method and reference score profiles.

Prompt design. The judge receives the candidate conversation, a short corpus description, and three reference conversations from the same corpus as anchors. The prompt explicitly notes that reference conversations can include hesitation, topic shifts, false starts, repair, persistence, and varied turn lengths. This prevents the judge from treating smoothness alone as realism. The full prompt is reproduced in this appendix.

Full per-dimension quality scores. Table 16 reports the full 8-dimension LLM-as-Judge scores summarized in Table 7. MAD is computed against the reference-conversation quality profile across the eight dimensions.

Per-corpus scoring and aggregation. For the quality-calibration audit, we sample 500 generated conversations from each method–corpus cell. We score this sample with the same Claude Sonnet 4 judge and average scores along each quality dimension. The reference-conversation row uses the same rubric applied to test conversations from the same corpus. Table 7 reports the average 8-dimensional score profile across corpora, while Table 17 reports the per-corpus MAD summary. Figure 4 complements these magnitude-only summaries by showing the signed deviation from the reference profile.