PERSONAFACTORY: STYLISTIC USER PROMPT AUGMENTATION WITH SYNTHETIC PERSONA BANK

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly replace real people in the AI development pipeline, as they are used to generate training dialogues at scale, control alignment and evaluation. However, out-of-the-box models produce monolingual and stylistically flat texts as synthetic user prompts, failing to replicate the real user-AI communication mechanics. We present **PersonaFactory**, an end-to-end pipeline to generate detailed user personas, later using them to augment existing user prompts from their point of view. As part of this work, we showcase two major data contributions, facilitating research in LLM-based stylistic chat generation. First, we release a bank of nearly 100,000 structured, fully synthetic personas, focusing on their background and communication behavior with a chatbot. Second, we release a corpus of roughly 236,000 persona-styled synthetic conversations spanning 77 languages, produced by using the generated bank of personas to augment existing user prompts and match personas to them based on original prompt intent and profile background. [1] We also provide an ablation study of the assistant persona design for response augmentation and downstream performance after training on such conversations. Finally, we evaluate the effect of personal features on style adoption and intent preservation, using an LLM-judge and heuristic metrics to assess augmentation efficiency. We find that simulating users' unique behavioral features in the conversation greatly benefits from a structured, attribute-based approach, compared to a single free-text description in existing personalization methods. We show that the produced augmented data is hard to distinguish from human-written texts for real people and that PersonaFactory as a training artifact can improve model adaptive behavior purely from data impact without steering.

1 INTRODUCTION

Conversational datasets are integral to large language model development, shaping how models understand users' intents and behave in diverse settings. Existing open-source datasets largely focus on LLM-generated synthetic setups because human-based data scaling is cost- and time-intensive. Therefore, AI is now used in all stages of training other models on real-life interaction with human users. This includes user questions and assistant responses, judging the quality of such QA pairs, and evaluating model efficiency in chat interaction settings. However, this efficiency comes at the cost of authenticity and diversity - the raw material for most simulated users is the same narrow band of text: instruction-tuning corpora written in a homogeneous, well-punctuated, standard-register "assistant/user" voice. The resulting distribution is averaged to one bland user, who is monolingual, verbose, and stylistically mundane, painting a greatly limited reflection of real-life user distribution spanning beyond numerous cultures, languages, and unique behavior characteristics. Such datasets, widely used to train open models, lack non-native phrasing, code-switching, terse or lowercase typing, dialectal variation, and wide gaps in digital fluency, giving a narrow picture of how users actually communicate to AI in real-life conditions.

Personalization Gap. When simulating conversations, there are two common design pitfalls: (1) existing methods on simulating personalized conversation from scratch with a simple pre-seeded persona information lack focus depth on important parts of persona background which influence

[1] We release an anonymous version of the pipeline codebase and data sample at Anonymous4Open

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

content and style with a greater impact; (2) rewriting existing conversations preserves the original intent, but at the cost of limited visual difference. Grounding conversations between a user and the AI assistant in a pre-generated persona presents an opportunity to address this, with related existing work on synthetic persona generation; however, persona descriptions are usually limited to a 2-5 short sentence summary or lack information about behavior in a chatbot setting. Compressing the user's full background into a short description may change *what* the user is saying, but not *how*, as it does not focus on their writing style.

Our work. We present **PersonaFactory**, a pipeline to target these particular issues in synthetic persona and chat data generation. The key idea is to keep the original dialogue trajectory from an existing real or synthetic conversation, but re-simulate the user side under a structured synthetic persona: the request stays fixed, while the register, typing style, and language behavior match those of a simulated user drawn from a large, controllable population. The persona description covers not only basic information such as age, origin, occupation, and a persona description, but also writing style, linguistic quirks, digital fluency and AI stance, painting a much richer background on *how* the persona would interact with an AI model. Since the intervention is a constrained rewrite of real user turns rather than free generation, it maintains the realism of verified chats while gaining the diversity and authenticity from the user's side.

Contributions. To improve the efficiency of user simulation, we introduce three components, which also represent the principal contributions of this paper:

1. **A structured user-simulator profile schema and a bank of nearly 100,000 profiles** (Section 3). Each profile has a *background* layer (geography, occupation, education, age, native language, AI stance, digital fluency) and a *style* layer (capitalization, punctuation, emoji, typos, abbreviation habits, code-switching, quirks, and a verbatim example message). Profiles are grounded on a manually-curated country, sub-region, and occupation grid with region-decoupled occupations and sector-coherent fluency, generated by three open LLMs and screened by an independent judge.
2. **A grounded simulation pipeline** (Section 4) that (i) conservatively matches a language- and script-compatible persona to each log, refusing to assign a persona rather than risk a language mismatch; (ii) rewrites user turns to preserve the exact request, constraints, and language while adopting persona style; and (iii) regenerates assistant turns with *next-turn continuity supervision*, using the real follow-up user message as weak supervision that the simulated turn stays on-trajectory, followed by a six-class sanity gate.
3. **An intrinsic evaluation of simulation fidelity** (Section 5) separating *style adoption* from *intent preservation*, with position-bias and multi-judge controls, and complemented by judge-free corpus-scale metrics (embedding similarity, edit depth, and rule-based entity and style checks). Our headline finding is that a *structured* per-axis style representation is preferred over a free-text style description on style fidelity, while both preserve intent equally.

As a result of this research, we plan to open-source our generated persona bank and augmented chat corpus together with a deeper look into feature-level effect on personalization and a refined end-to-end pipeline to improve chat logs from scratch, verified by a variety of LLM-as-a-judge and heuristic metrics.

2 Background and Related Work

User and persona simulation. LLM-based user simulation has largely replaced agenda-based simulators (Ni et al., 2026), either by training purpose-built user models (Naous et al., 2026; Kong et al., 2024; Sun et al., 2024; Wang et al., 2025a) or by optimizing simulators with reinforcement learning (Wang et al., 2026; Qian et al., 2025; Zhang et al., 2026); a parallel persona-conditioned line, from PersonaChat (Zhang et al., 2018) to role-play agents (Tseng et al., 2024; Samuel et al., 2025; Wang et al., 2025b), conditions generation on short biographical profiles. Across both lines, the simulated user remains generic: an averaged user served as a live, hard-to-audit model, with personas that say *who* a character is but not *how* they type, and the behavior distribution becomes skewed and hard to control (Li et al., 2026). PersonaFactory instead rewrites existing logs under a quality-gated persona population with a first-class style layer, and releases a curated corpus rather than a simulator.

Persona datasets and persona-conditioned chat data generation. Large persona datasets are largely used to improve synthetic training data (Ge et al., 2025; Meyer & Corneil, 2025; Ma et al., 2026; Ryan et al., 2025), on top of persona-free recipes (Wang et al., 2023; Ding et al., 2023), and real-log sources (Zhao et al., 2024). However, adding a persona influences the content of the data. At the same time, the style remains in the same generic voice, as the persona itself is usually a one- to two-sentence free-text description that underspecifies typing behavior. Methodologically, the closest work is SynthPAI (Yukhymenko et al., 2024), which seeds LLM agents with structured profiles, including a description of writing style, to generate realistic Reddit threads as the privacy-preserving alternative to attribute inference evaluation of Staab et al. (2024). However, SynthPAI targets user-user interaction rather than user-AI chat. We build on this idea of a separate style entity: we decompose the style description into structured per-axis keys and show that this representational change alone improves style fidelity (Section 5), ground generation in existing logs with the original intent fixed rather than generating from scratch, and decouple occupation from region to suppress stereotypes (Section 7).

Sociolinguistic variation in NLP. Dialect and register variation is known to degrade and bias NLP systems (Blodgett et al., 2016; Ziems et al., 2022), and persona prompting alone induces only limited diversity (Kambhatla et al., 2025); yet the synthetic pipelines producing today's conversational training data barely vary register, so the variation these systems handle worst is exactly what their training data lacks. PersonaFactory makes this variation (code-switching, typo tendency, capitalization, digital fluency) the explicit object of simulation, as conditioning variables rather than claims of dialect authenticity.

3 Scaling PersonaFactory Bank with Realistic Synthetic Personas

First, we describe the synthetic profile generation pipeline, which produces a highly diverse bank of nearly 100,000 personas (99,696), focused not only on users' backgrounds but also on their communication style in a human-AI setting.

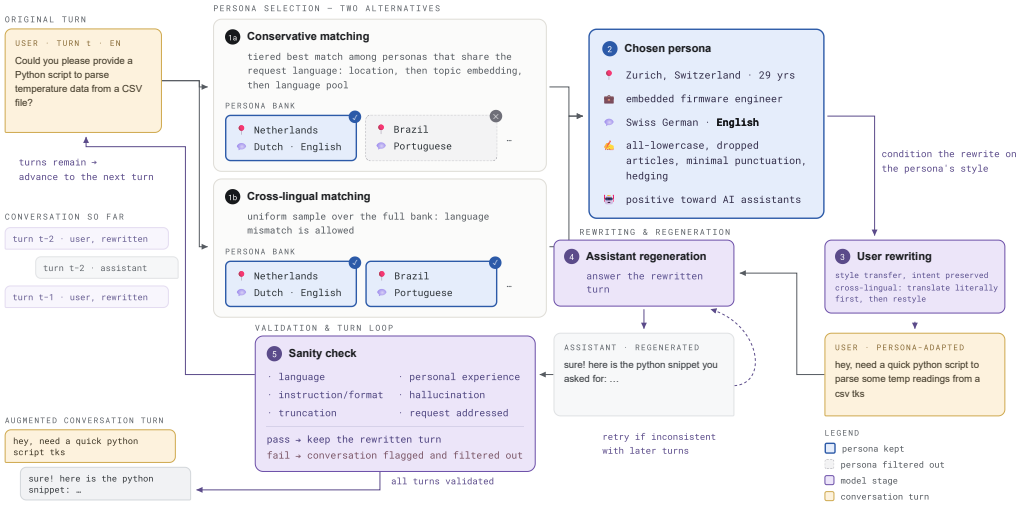


Figure 1: The PersonaFactory pipeline. A real log is matched to a compatible user-simulator profile; user turns are rewritten in that profile's voice with the request held fixed, and assistant turns are regenerated correspondingly.

Structured profile representation. We split the persona representation into two layers: *background* and *style*. The *background* layer provides information on age (adults only, 18–72 in the generated bank, as we find simulating underage personas unethical), gender, country/region of origin, occupation (with sector), education, native language (with regional dialect), hobbies, AI stance and digital fluency level. Based on this background, we condition the generation of *style* layer, covering sentence-length tendency, question framing, native-language error patterns, code-switching

frequency and type, a cultural reference pool, a structured seven-key chatting style (capitalization, punctuation, emoji use, abbreviation habits, typo tendency, message structure, sign-offs), several (2-4) linguistic quirks, and an example message the persona would usually type. Separating them lets us use the first layer for content adjustment and the second layer to change how it will be written (Section 4).

Generating the persona bank. We pass the pre-defined JSON schema into the generator and use it to produce N personas in the specified format. As generator models, we used `Gemma 4 31B` (Team et al., 2026), `Ministral 3 14B` (Liu et al., 2026) and `GLM 4.7 Flash` (GLM-4.5 Team et al., 2025), balancing each model's geographical and multilingual strengths for various regions – the model selection is equally balanced across generated models. We ground the generation in a manually curated set of 81 countries and their sub-regions, combined with over 200 concrete occupations spanning sixteen sectors (details in Appendix D). We sample occupations independently of origin region to avoid biased profiles, which we initially observed (e.g. for Switzerland, the most popular occupation would be watchmaker or software engineer; Eastern Europe and Asia are more agriculture-heavy). Moreover, digital fluency is matched to the occupation sector to ensure realistic backgrounds. The style layer is filled in the same generation pass. Still, it is explicitly conditioned on the already-sampled background: the generation prompt (Appendix A) instructs the generator to derive each style attribute from the persona's age, education, digital fluency, and native language, choosing every structured key from a fixed set of options that must stay consistent with the free-text style sentence and the verbatim example message. Profiles with a degenerate style layer, that is, a missing style description, fewer than two linguistic quirks, no example message, or empty structured keys, are rejected outright before judging.

Judge verification. An independent judge (`MiniMax-M2.5` (Chen et al., 2026), not one of the generators, avoiding self-preference) scores every profile on internal coherence, occupation/location plausibility, LLM-user authenticity, stereotype severity, and realism, with a two-pass repair that corrects implausibly low fluency for technical roles and rewrites age-incoherent job titles before re-judging. The gate is strict but rarely triggered: about 97% of generated profiles are kept, judged realism is high, and only a handful of profiles are flagged for non-trivial stereotype severity. Appendix D (Table 4) provides full coverage, demographic, and per-generator quality statistics.

4 Prompt Augmentation using Synthetic Personas

A grounded user simulator must satisfy four requirements: **(R1)** preserve the real task and its constraints, defaulting to its language; **(R2)** express sociolinguistic diversity beyond the standard register; **(R3)** stay on the logged conversation's trajectory; and **(R4)** remain auditable and leakage-free, so that data trained on it carries no hidden persona shortcut. PersonaFactory's four stages map onto these requirements: conservative matching and constrained rewriting serve R1, cross-lingual selection and structured style layer serve R2, next-turn continuity supervision serves R3, and storing only observable turns serves R4. We illustrate the pipeline in Figure 1.

Persona matching and selection. We first detect the language and script of the initial user turn using deterministic Unicode block classifiers (covering CJK, Cyrillic, Arabic, etc.) coupled with a confidence-thresholded language detector.[2] To serve distinct research needs, the pipeline operates under two selection regimes:

1. **Conservative (Language-Matched) Selection:** A persona is eligible only if it is verified to communicate natively in the prompt's language and script. Non-Latin prompts are mapped strictly to native-speaking personas sharing the target script, non-English Latin prompts to detected-language personas, and English prompts to the broader English-speaking pool. A single best-matching persona is retrieved through prioritized tiers: geographic location matching, top-1 semantic embedding cosine similarity, or deterministic language-pool fallback. If no compatible persona exists, the pipeline returns the original turn verbatim.
2. **Cross-Lingual Diverse Selection:** To reflect authentic multilingual populations who interact with LLMs in global languages, this mode samples a single persona uniformly at random from the en-

[2] `https://github.com/pemistahl/lingua-py`

tire bank (where native-English personas constitute only 19.6%), allowing non-English personas to re-voice prompts with their regional register, cultural framing, and loanwords while relaxing strict script-stability abort gates.

Exact-format prompts (e.g., verbatim reproduction, code generation tasks, direct translation, and strict JSON/quoted output instructions) are automatically detected and bypassed by format guards. In a controlled pilot, conservative matching raises style adoption by more than 20 points over a random persona of the same language (Appendix F).

User-turn rewriting. For eligible turns, the rewriter is instructed to preserve the original intent while introducing only subtle persona traits. Linguistic quirks and colloquialisms are applied probabilistically at a rate of $p = 0.25$ per turn; this prevents over-stylization and caricatures across multi-turn dialogues while preserving subtle demographic and register cues. A length-ratio guard automatically rejects over-compressed rewrites ($< 50\%$ of the original length), falling back to the verbatim seed prompt if any validation gate fails. Quoted text, code, URLs and numbers are masked before the rewrite and restored afterwards, and a rewrite that loses a masked span is retried. In cross-lingual mode, we tried two ways of reaching the persona's language: re-voicing the source turn directly into that language in a single pass, and first translating each turn literally and then restyling the translation with the same-language rewrite. The two-step variant preserves the request and lands in the target language more reliably, at a small cost in style adoption (Appendix F), so the pipeline uses it.

Assistant regeneration with continuity supervision. The restyled user turn is appended to the running context, and the assistant response is regenerated. The persona background/style is provided to the responder as an unsaved system prompt: the model sees the user's register during generation, but the stored training example contains only the observable user and assistant turns, so a downstream learner acquires style from conversational cues, not from hidden persona metadata. For non-final turns, a judge checks whether the regenerated response plausibly leads to the actual next user message from the log; if not, a bounded adaptation loop (up to 2 iterations) rewrites the assistant response so that the real historical follow-up remains a coherent continuation, preserving the logged dialogue trajectory. If the judge still flags the response after the last iteration, the last adapted response is kept and left to the sanity gate. The continuity-judge and adaptation prompts are given in Appendix A.

Sanity checks and judge verification. A final judge screens each conversation for six failure classes: language mismatch (for conservative matching), instruction/format violation, truncation, fabricated personal experience, incoherence/hallucination, and failure to address the request. These sanity verdicts are stored as metadata flags, allowing downstream training pipelines to selectively filter for strictly validated, constraint-intact dialogues.

5 Evaluation and Augmentation Analysis

We target two main tasks with our proposed persona-based augmentation method: it should sound like the seeded user (style adoption) and should still keep the same original topic in the same language (intent preservation). We evaluate both and study the trade-off between these two goals to compare design choices. We define the following next steps to achieve this: we define the two fidelity axes and evaluate them across all 552,654 released turns using both an out-of-family LLM judge and deterministic, judge-free metrics; we answer our central design question with a paired, order-randomized A/B that isolates the effect of style representation by holding generator model, persona, and token budget fixed; and we re-score a stratified subset with four judges to confirm the direction is not a single-judge artifact.

Corpus-scale method evaluation. We score both fidelity axes through two complementary measurement families: an out-of-family reasoning judge (`MiniMax-M2.5`) and deterministic, judge-free heuristics (Appendix F). Across all 552,654 released turns, the LLM judge yields **97.18% intent preservation** and **68.23% style adoption** (66.35% passing both simultaneously), with higher-capacity generator backbones adopting style more faithfully (Gemma-4-31B: 70.73% vs. Ministral-3-14B: 64.06%). We complement this with deterministic metrics from an independent, rule- and embedding-based family, revealing three consistent patterns. First, rewrites introduce substantial surface variation while preserving underlying meaning: same-language adaptations modify a large share

of the surface form while maintaining high multilingual embedding similarity and core-vocabulary retention. Additionally, style steerability stays strong, achieving high adherence on emoji usage and moderate adherence on punctuation and capitalization patterns, whereas continuous attributes like message length show a smooth shift.

Writing style representation ablation. Our central design question concerns representation: to make a user sound a certain way, is it better to provide the rewriter with a single free-text sentence describing the users chatting style in a flat manner of 1-2 short sentences or with the per-axis structured style keys (structured)? We conduct a controlled, paired A/B comparison. For each of 553 real user messages, one persona is selected deterministically, and the same rewriter restyles the message twice, once from the flat description and once from the structured keys with all other settings held constant; the two persona blocks are near-equal in length (311 and 324 words), so the structured variant does not receive more input. After the pipeline's judge passing stage we use 518 remaining pairs for this evaluation. The comparison was run on an earlier release of the persona bank with the same structured style schema. An independent judge (`MiniMax-M2.5`) issues a pairwise style-fidelity verdict with A/B order randomized per item to control for position bias, alongside a separate intent check for each rewrite. Table 1 reports the results on the 433 pairs scored in the original pass, where the remaining verdicts were lost to output truncation of the reasoning judge; re-scoring all 518 pairs with a larger output budget and with three judges from outside the pipeline gives the same direction (Appendix F, Table 13).

Table 1: Structured vs. flat style representation for user simulation (paired, same rewriter, independent judge, randomized order; n=433 scored pairs). 95% Wilson intervals: structured preferred 55.4–64.6%, flat preferred 28.8–37.6%, structured share of decisive verdicts 59.2–68.5%.

Metric	Flat	Structured
Pairwise style preference	33%	**60%**
Decisive share (ties excluded)	35%	**64%**
Intent preserved	92%	92%
Mean rewrite length (chars)	1582	1586

The structured style wins the head-to-head comparison by roughly 1.8 to 1 (Table 1), while both representations preserve the original request equally often and produce rewrites of essentially identical length. The effect is stable across independent judge passes. In head-to-head reads, the judge repeatedly credits the structured rewrite for retaining specific markers (signature abbreviations, sign-offs, ellipsis pauses, code-switched tokens) that the flat rewrite smooths away. The interpretation is that explicit, separable levers (lowercasing, emoji rate, typo tendency, code-switching) are easier for a rewriter to act on faithfully than a single prose sentence it must parse and prioritize (Figure 2). Because intent and length are held fixed, the win is attributable to representation, not to the model working harder.

Comparison to other persona representations. We also compare PersonaFactory with PersonaHub (Ge et al., 2025) and SynthPAI (Yukhymenko et al., 2024). We keep the persona content, seed prompt, rewriter, judge, and decoding settings fixed, changing only how the profile is presented. Table 2 shows a clear pattern: formats that describe writing behaviour more explicitly produce stronger style adoption, and PersonaFactory is preferred in every pairwise comparison. PersonaHub, which has no writing-style field, performs only slightly above the no-persona baseline. The ranking holds when the judges see only format-neutral style attributes, under two external judges, and against real PersonaHub personas, while real SynthPAI profiles are on par. Richer style conditioning has a moderate cost: intent preservation drops by about four points, almost entirely on cross-lingual prompts, and external judges rate the rewrites somewhat less natural. Appendix H provides the full breakdown and robustness checks.

LLM-as-a-judge robustness. LLM-based judge scoring is prone to bias and stereotypical evaluation; models may also provide lower-quality evaluation if they display generic user alignment themselves. We therefore test the judge choice in two ways. First, we re-score a sampled subset of persona rewrites with four judges (`Gemma 4 31B`, `GLM 4.7 Flash`, `Ministral 3 14B`, `MiniMax M2.5`). The judges differ in severity, so raw agreement overstates consistency: chance-corrected

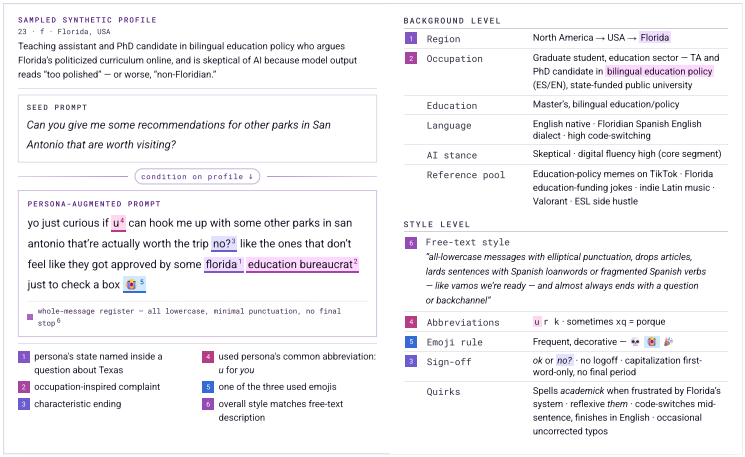


Figure 2: An example of an augmented user prompt, seeded with a synthetic persona from the generated Persona Bank. The new user question successfully implements features, unique from profile's background and style information.

Table 2: Persona representations on 800 paired prompts. Only the profile format changes; persona content, rewriter, judge, and decoding are fixed. *PF preferred* is the pairwise style preference for the PersonaFactory rewrite. Brackets: 95% Wilson intervals.

Representation	Words	Style adopted (%)	Intent kept (%)	PF preferred (%)
No persona	0	3.1 [2.1, 4.6]	99.8 [99.1, 99.9]	94.4 [92.5, 95.8]
PersonaHub form	13	5.5 [4.1, 7.3]	99.8 [99.1, 99.9]	94.9 [93.1, 96.2]
SynthPAI form	83	43.5 [40.1, 47.0]	98.0 [96.8, 98.8]	81.3 [78.4, 83.9]
PersonaFactory	280	69.2 [66.0, 72.3]	96.1 [94.6, 97.3]	–

agreement ranges from moderate to strong for most pairs, and every pair involving the production judge reaches Cohen's κ of at least 0.32 on style and 0.22 on intent, but it is near chance for one pair of lenient judges on style (Appendix F). Second, because these four models are also generators or the production judge of the pipeline, we re-score the central A/B of Table 1 with three judges from outside the pipeline (`DeepSeek-V4-Flash`, `Qwen3.5-397B-A17B`, `Qwen3.6-35B-A3B`). All three prefer the structured representation, with decisive-verdict shares between 54% and 71% (Table 13). We use `MiniMax-M2.5` as the production judge because it is not one of the production rewriters and is the strictest on style among the candidates (Table 10), giving the pipeline a more conservative evaluation during generation.

Human evaluation. To validate the effieciency of user augmentation pipeline, we run an IRB-approved human study. We employ experts in AI field and native speakers to rate samples of the corpus in their respective native/fluent languages on four tasks: 1) intent preservation after the rewrite step; 2) style following in augmented user prompts in natural way; 3) human- vs AI-written text classification (our chats mixed with WildChat samples, written by humans); 4) multilingual authenticity of updated conversations. At the moment of the submission, we have collected 11 responses across 9 languages with the three defined metrics showing a median score of 4 on a 5-point scale (Table 25), and raters judged verified model rewrites to be human-written in 59% of cases (Table 27). We also evaluated inter-annotator agreement on human-AI classification task with several LLM judges, which show even worse guessing accuracy: they label 83 to 89% of the rewrites as human-written and separate rewrites from chat-log texts with 50.5 to 57.4% accuracy, close to chance and to the raters' 55.3% (Appendix J).

Attribute-level augmentation impact. We also look beyond how the writing style is represented by studying whether any particular persona attributes affect any rewrite axis by ablating one attribute at a time and scoring style adoption, intent preservation and script stability separately. We observe that all standalone attributes showcase similar levels of impact on augmentation sub-tasks. At the same

time, together they produce a more visible influence on the final outlook of the augmented prompt. Moreover, treating style as a distinct structured representation level enables more localized style personalization with code-switched tokens, sign-offs, ellipsis pauses, and signature abbreviations that deviate from standard mundane writing. We provide a thorough analysis in Appendix I.

6 The PersonaFactory Conversational Dataset

Applying the pipeline at scale to two widely used conversational datasets, UltraChat (Ding et al., 2023; Tunstall et al., 2023) and Nemotron Instruction-Following (NVIDIA, 2025a; Nathawani et al., 2025; NVIDIA, 2025b), yields 236,191 augmented conversations after deduplication, language verification and judge filtering (273,619 before the final sanity gate), covering 77 languages, as summarized in Table 3. The corpus comprises both English-centered user conversations and multilingual ones, in which the persona's native language is enforced in place of English.[3] Persona matching is predominantly driven by the semantic-embedding component, whereas the secondary language and location-pool matching contribute less frequently.

In the multilingual split, generating conversations at the language-level yields better script stability across turns. The largest non-English languages are Spanish, French, German, and Chinese (per-language distributions are provided in Appendix E). At the sub-regional level, the bank records 7,435 distinct free-text dialect labels (e.g. cantonal Swiss German varieties); we treat these as a free-text label variety rather than a validated dialect taxonomy.

We observe that Gemma-augmented conversations pass the sanity gate far more often than Ministral ones (Table 3), as Ministral is more prone to making radical changes to the intent and style of the prompts. Nemotron samples are longer on average and often contain multi-step instructions or constraints, which lowers user-side style adoption (Appendix F), while the final sanity pass rate is similar for both seed sources (Appendix E). In controlled probes Ministral restyles more aggressively than Gemma, but at the cost of intent (Appendix I); at corpus scale, where rewrites must also survive the sanity gate, Gemma's turns score higher on both style and intent (Appendix F). We keep both models in the generator pool for greater diversity.

Assistant-side model personas. While user-side personalization can diversify training data format, the downstream training effect is still questionable. The model will see richer texts coming from various backgrounds, but will still produce the same standardized response style with long, heavily formatted answers regardless of who is asking. As the first step towards mitigating this, we have conducted a pilot study into model (assistant) personas and matching them with seeded user personas during augmentation stage, which is part of the released corpus. We define 12 model personas with various response styles (e.g. answer-first compact prose, headed sections with numbered steps, or a gentle register for stressful topics), each defined by a free-text style description, eight presentation axes (register, structure, density, warmth, explanation, jargon, certainty, emoji), optional signature touches offered with $p = 0.25$ per reply, as for user quirks, and content safeguards. We use the style descriptions only to control how the answer looks like, but pre-existing user instructions still override it if present. For every user-augmented conversation, we pass a seeded user persona into the matcher model (`MiniMax-M2.5`), which we ask to provide best matched model persona from the defined pool. Afterwards, the model response is being augmented with provided assistant persona and user style information to provide the match - this setup shows the best downstream performance after training, however, we provide with the pipeline the ability to augment model responses only in case the information about user persona is not present, which achieves similar performance (Appendix G). We therefore treat model personas as a controllable source of assistant-style diversity and a starting point for future work rather than a core component of the pipeline.

7 Impact and Limitations

Impact. We hope our pipeline will stimulate a discussion on the efficient use of synthetic personas in open-source data generation and LLM training. Upon acceptance, we will release the persona

[3] We count the detected language of the first user turn (fastText language identification with a Lingua fallback), not the persona's declared language, and merge variants the detector cannot separate (e.g. Malay and Indonesian, Bokmål and Norwegian).

Table 3: Persona-simulated corpus: 273,619 deduplicated, language-verified conversations across 79 languages; the 236,191 that pass the sanity gate form the released dataset, whose statistics are shown.

Property	Value
Conversations / user turns	236,191 / 552,654
Sources	UltraChat 114,172 / Nemotron 122,019
Sets	language-matched 176,286 / cross-lingual 59,905
Match tiers (emb./cross-ling./loc./lang-pool)	117,132 / 58,137 / 58,420 / 2,502
Turn depth (1 / 2 / 3; mean)	33,691 / 88,537 / 113,963; 2.34
Responders	Gemma 144,515 / Ministral 91,676
Distinct languages (total / cross-lingual)	77 / 75
English share (overall / cross-lingual set)	74.3% / 1.6%
Sanity pass (conversation / turn level)	86.3% / 93.3%
Conv. pass by responder	Gemma 92.7% / Ministral 77.9%
Assistant model personas (Section 6)	12 matched response styles

bank and all generated conversations under a permissive license to further facilitate this effort; a key contribution is a culturally diverse dataset that we expect to be valuable for multilingual LLM development. The released bank and corpus do not reveal any real user's identity or private text: the pipeline induces sociolinguistic variety without processing personal data and relies on permissive open-weight models, providing reproducible and responsible resources for LLM training. At the same time, personas are conditioning variables, not real people or groups: they make no claim about how members of any nationality, region, gender, or profession actually communicate, and must not be used for profiling, targeting, or deceptive content presented as the first-person voice of a real individual.

Limitations. Our work has several limitations, which we plan to address in future work. First, our human evaluation can still be upscaled: most languages have a single expert rater, panels differ across languages, the study has no baseline arm, and the human-versus-model task lacks a verified human-source key, so it supports the LLM-judge results descriptively but cannot yet rank languages or establish indistinguishability from human writing (Appendix J). Second, apart from the assistant-side fine-tuning ablation (Appendix G), the evaluation focuses on the data itself rather than the effect of the user-side augmentation on model development, so we cannot claim that such data improves any downstream model behavior. Third, multilingual representation is imbalanced: English and major European and East-Asian languages dominate with only a thin long tail, low digital fluency covers only a small fraction of profiles despite motivating the work, and quoted string literals remain the least robust entity type under restyling (Table 9). Finally, the assistant-side model personas are only a first step. The 12 styles are hand-written, the matcher favours a few of them, potential future directions include model persona set size and on-policy persona sampling. Model personas can also be applied to conversations without a user persona, but when a source conversation carries its own system prompt, the style competes with its format and length constraints, so we leave such conversations non-augmented (Appendix G).

8 Conclusion

In this work, we present PersonaFactory, a pipeline for personalizing user conversations. It begins with the scalable generation of realistic personas, which are then used to augment existing conversational datasets for improved authenticity and diversity, while preserving the original user intent. Our key contributions are a bank of nearly 100,000 structured synthetic personas; a corpus of roughly 236,000 personalized conversations in 77 languages, produced by a language-aware pipeline that re-simulates the user side of real logs while preserving intent and trajectory; and an evaluation of the impact of persona features on intent and style. We find that separating persona background from writing style improves the quality and authenticity of the modified conversations, yielding more natural and diverse multilingual training data. We believe this work lays a foundation for further research into the influence of persona detail on the conversational alignment of LLMs.

Ethics Statement

The main motivation of introducing PersonaFactory is to carry no real personal data: the persona bank is fully synthetic, and the augmented conversations are based on the openly released chat datasets, like UltraChat (Ding et al., 2023) and Nemotron Instruction-Following (NVIDIA, 2025a), which we credit and use under their original terms. The pipeline therefore induces sociolinguistic variety without processing anyone's private text, and no persona in the bank is derived from, or matched to, an identifiable individual. Personas are conditioning variables rather than descriptions of real groups: they make no claim about how members of any nationality, region, gender, or profession actually write, and we decouple occupation from region precisely to suppress the stereotypical associations we observed when the two were sampled jointly (Section 3). Every profile passes an independent judge that scores stereotype severity, and the residual risk of the release is the obvious one: a persona bank that encodes how people type can also be used to mass-produce stereotyped text, or to present synthetic output as the first-person voice of a real person. We release under a permissive license with these uses explicitly disallowed, and we discuss impact and residual limitations in Section 7. All generator and judge models used here are open-weight models with permissive usage terms. The human evaluation was approved by the ethics review board (IRB) of the authors' institution and classified as minimal risk; participants were volunteer domain experts who gave informed consent, took part unpaid in their professional capacity, could skip items or withdraw at any time, and were identified only by random participant IDs, with no names, contact details, or IP addresses collected (Appendix J).

Reproducibility Statement

To ensure reproducibility of work, we are releasing an anonymized version of pipeline codebase and produced Persona Bank together with a dataset of augmented conversations. We describe in full detail the process of synthetic persona creation (including persona schema, the curated country/sub-region and occupation grid, used models for judging and generation) in Section 3 and Appendix D; further pipeline stages are described in Section 4, which provide information to reproduce the conversation augmentation using our proposed personas with details on used models, generation settings, persona feature representation and feature-level sampling conditions. We also provide the prompts used to generate these parts (profile generation, user-turn adaptation, and the persona-conditioned responder prompt) verbatim in Appendix A together with end-to-end examples of chat augmentation based on provided persona in Appendix B. Ablation experiments and motivation behind selection of the final setup choices like judge models, style representation during augmentation and others are present in Section 5, Appendix I and Appendix H. Corpus composition, matching-tier counts, and sanity-gate pass rates are reported in Section 6 and Appendix E. Upon acceptance we will publicly release the persona bank, the augmented conversational dataset, and the pipeline code under a permissive license.

AI Use Statement

Generative models are the object of study and a core component of the method itself: persona profiles, user-turn rewrites, assistant regenerations, and every fidelity verdict are produced by the open-weight models named in Sections 3 and 4 and in Appendix F, with their prompts given verbatim in Appendix A and their validation criteria stated alongside. We additionally used generative AI tools for writing assistance (copy-editing and phrasing) and for coding assistance in the analysis and visualization scripts. Research ideas, experimental design, and the interpretation of results are the authors' own. All AI-assisted output was reviewed by the authors: generated code was inspected and tested, and every reported number was checked against the underlying run artifacts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. Don't lose the message while paraphrasing: A study on content preserving style transfer. 2023. URL https:

//arxiv.org/abs/2308.09055.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120/.

Aili Chen, Aonian Li, Baichuan Zhou, Bangwei Gong, Binyang Jiang, Boji Dan, Changhao Zhang, Changqing Yu, Chao Wang, Cheng Ma, Cheng Zhong, Cheng Zhu, Chengjun Xiao, Chengyi Yang, Chengyu Du, Chenyang Zhang, Chi Zhang, Chuangyi Huang, Chunhao Zhang, Chunhui Du, Chunyu Zhao, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Dongyu Zhang, Enhui Yang, Fei Yu, Guang Zheng, Guodong Zheng, Guohong Li, Haichao Zhu, Haigang Zhou, Haimo Zhang, Han Ding, Hao Zhang, Haohai Sun, Haolin Lyu, Haonan Lu, Haoyu Wang, Huajie Shi, Huiyang Li, Jiacheng Chen, Jian Zhang, Jiaqi Zhuang, Jiaren Cai, Jiaxin Pan, Jiayao Li, Jiayuan Song, Jichuan Zhang, Jie Wang, Jihao Gu, Jin Zhu, Jingwei Dong, Jingyang Li, Jingyu Zhang, Jingze Zhuang, Jinhao Tian, Jinli Liu, Jinyi Hu, Jun Tao, Jun Zhang, Junbin Ruan, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kang Xu, Ke Ji, Ke Yang, Kecheng Xiao, Keyu Duan, Keyu Li, Le Han, Letian Ruan, Li Yuan, Lianfei Yu, Liheng Feng, Lijie Mo, Lin Li, Linge Du, Lingye Bao, Lingyu Yang, Lingyuan Zhou, Loki, Lu Chen, Lunbin Zeng, Ming Li, Ming Zhong, Mingliang Tao, Mingyuan Chi, Mujie Lin, Nan Hu, Ningxin Chen, Peiyin Zhu, Peng Gao, Pengcheng Gao, Pengfei Li, Penglin Li, Pengyu Zhao, Qibin Ren, Qibing Ren, Qidi Xu, Qihan Ren, Qile Li, Qin Wang, Quanliang Chen, Qunhong Zeng, Rong Tian, Rongxin Guo, Rui Dong, Ruitao Leng, Ruize Zhang, Shanqi Liu, Shaoxiang Chen, Shaoyu Chen, Sheng Jia, Shun Yao, Shuoran Zhao, Shuqi Yu, Sichen Li, Sicheng Pan, Songquan Zhu, Tengfei Li, Tian Xie, Tiancheng Qin, Tianle Li, Tianrun Liang, Wei Liu, Weiqi Xu, Weitao Li, Weixiang Chen, Weiyu Cheng, Weiyu Zhang, Wenhu Chen, Wenqian Zhao, Xiancai Chen, Xiangjun Song, Xiangyuan Wang, Xianzhen Luo, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xiaojie Wu, Xihao Song, Xingyi Han, Xinyu Guan, Xuan Lu, Xun Zou, Xunhao Lai, Xutong Li, Xuyang Shen, Yan Gong, Yan Ma, Yang Jiao, Yang Wang, Yang Xu, Yangsen Wang, Ye Tang, Yicheng Chen, Yihang Wang, Yinran Qiu, Yiqi Shi, Yiting Guo, Yiwen Huang, Yixuan Wang, Yongyi Hu, Yu Gao, Yu Zhang, Yuan Li, Yuanxiang Ying, Yuanzhen Zhang, Yubo Wang, Yuchen Song, Yufeng Yang, Yuhang Meng, Yuhang Miao, Yuhao Li, Yujie Liu, Yulin Hu, Yunan Huang, Yunji Li, Yunyi Huang, Yusen Zhang, Yusu Hong, Yutao Xie, Yutong Zhang, Yuwen Liao, Yuxuan Shi, Yuze Wenren, Zebin Li, Zehan Li, Zejian Luo, Zeyu Jin, Zeyuan Sun, Zhanpeng Zhou, Zhaochen Su, Zhendong Li, Zhengmao Zhu, Zhengyuan Peng, Zhenhua Fan, Zhi Zhang, Zhichao Xu, Zhiheng Lv, Zhikang Xu, Zhitao He, Zhiwei He, Zhongyuan Li, Zibo Gao, Zijia Wu, Zijian Song, Zijian Zhou, Zijun Sun, Zishan Huang, Ziying Chen, and Ziyue Ge. The minimax-m2 series: Mini activations unleashing max real-world intelligence, 2026. URL https://arxiv.org/abs/2605.26494.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. 2025. URL https://arxiv.org/abs/2406.20094.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao,

Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Gauri Kambhatla, Chantal Shaib, and Venkata S Govindarajan. Measuring lexical diversity of synthetic data generated through fine-grained persona prompting. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21024–21033, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1146. URL https://aclanthology.org/2025.findings-emnlp.1146/.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. Platolm: Teaching llms in multi-round dialogue via a user simulator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7863, 2024.

Leon Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. *Advances in Neural Information Processing Systems*, 38, 2026.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Yuchen Ma, Yue Huang, Wenjie Wang, Xiaonan Luo, Xiangliang Zhang, and Stefan Feuerriegel. Synthetic interaction data for scalable personalization in large language models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Yev Meyer and Dane Corneil. Nemotron-personas-usa: Synthetic personas aligned to real-world distributions. Hugging Face dataset, https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA, 2025.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. 2026. URL https://arxiv.org/abs/2510.06552.

Dhruv Nathawani, Shuoyang Ding, Vitaly Lavrukhin, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Boris Ginsburg, and Jane Polak Scowcroft. Nemotron-Post-Training-Dataset-v2, August 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v2.

Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A. Rossi. A survey on LLM-based conversational user simulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4266–4301, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.200. URL https://aclanthology.org/2026.eacl-long.200/.

NVIDIA. Nemotron-instruction-following-chat-v1. Hugging Face dataset, https://huggingface.co/datasets/nvidia/Nemotron-Instruction-Following-Chat-v1, 2025a. Chat subset refreshed from Nemotron-Post-Training-Dataset-v2.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model, 2025b. URL https://arxiv.org/abs/2508.14444.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Jielin Qiu, Zhiwei Liu, Haolin Chen, Shirley Kokane, Heng Ji, Weiran Yao, Shelby Heinecke, et al. UserRL: Training interactive user-centric agent via reinforcement learning. *arXiv preprint arXiv:2509.19736*, 2025.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. Synthesizeme! Inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2506.05598.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.368. URL https://aclanthology.org/2025.findings-emnlp.368/.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing multi-turn instruction following for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9729–9750, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.525. URL https://aclanthology.org/2024.acl-long.525/.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana

Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar

Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. URL https://arxiv.org/abs/2310.16944.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles, 2025a. URL https://arxiv.org/abs/2502.18968.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*, 2025b. URL https://arxiv.org/abs/2501.15427.

Yingshan Susan Wang, Cedegao E. Zhang, Linlu Qiu, Zexue He, Pengyuan Li, Alex Pentland, Roger P. Levy, and Yoon Kim. Learning user simulators with turing rewards, 2026. URL https://arxiv.org/abs/2606.19336.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In Spandana Gella, He He, Bodhisattwa Prasad Majumder, Burcu Can, Eleonora Giunchiglia, Samuel Cahyawijaya, Sewon Min, Maximilian Mozes, Xiang Lorraine Li, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Laura Rimell, and Chris Dyer (eds.), *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL https://aclanthology.org/2022.repl4nlp-1.26/.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2406.07217.

Feng Zhang, Shijia Li, Chunmao Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Jingwen Xu, and Han Liu. Userlm-r1: Modeling human reasoning in user language models with multi-reward reinforcement learning. *arXiv preprint arXiv:2601.09215*, 2026.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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 (ACL)*, 2018.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. VALUE: Understanding dialect disparity in NLU. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3701–3720, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.258. URL https://aclanthology.org/2022.acl-long.258/.

A Prompt Templates

We reproduce verbatim the prompts at the core of the pipeline: the persona-profile generation prompt used to build the bank (Section 3), the user-turn adaptation prompt used for chat augmentation, the persona-conditioned system prompt under which assistant turns are regenerated, the continuity-judge and adaptation prompts of the next-turn continuity loop (Section 4), the literal-protection instruction and the cross-lingual translation prompt used before restyling, and the model-persona prompts of Section 6: the style section appended to the responder system prompt (its optional-touch line appears

only on turns where a signature touch is offered), the style reminder used during response adaptation, and the matching prompt. Braces mark values filled in per persona or turn.

Passed to each generator model under guided-JSON decoding. Placeholders in curly braces are filled programmatically from the sampling grid; bracketed markers denote optional blocks the driver inserts per persona.

Persona profile generation (bank)

```
You are an expert sociologist and character designer. Generate ONE
  highly realistic persona as a single valid JSON object.

Seed constraints (use EXACTLY these values, do not change them):
- Macro Region: {region}
- Country: {country}
- Sub-region: {sub_region}
- Profession: {occupation} (sector: {sector})
- Gender: {gender}
- Age: {age} (use EXACTLY this age; make occupation seniority,
  education, and life-stage consistent with it)
- AI Stance: {ai_stance}
- Digital Fluency: {digital_fluency}[optional code-switching-off
  constraint]

Fill every field in the JSON schema below so the whole persona is
  jointly coherent and culturally accurate.

AGE COHERENCE (mandatory, applies to the WHOLE persona): every cultural,
  technological, and media reference ANYWHERE - reference_pool,
  linguistic_quirks, example_message, character_description - must
  belong to the world THIS person's age cohort actually grew up in and
  uses NOW. Someone in their teens/20s/30s references contemporary,
  currently-live culture (the apps, platforms, creators, games,
  streamers, music, memes and slang of their own generation and today) -
  NOT obsolete technology (floppy disks, VHS, cassettes, dial-up, rotary
  phones, fax machines, typewriters, disk drives), NOT bygone decades or
  Cold-War / Soviet / "old-country" nostalgia, and NOT their
  parents'/grandparents' world presented as their own. Only genuinely
  older personas (roughly 50+) carry firsthand older-era references, and
  only where it fits. Never reach for a dated or nostalgic reference as
  shorthand for a nationality - that is caricature, not realism. Every
  reference must also be internally coherent (no mislabelled objects,
  e.g. calling a smartphone "headphones").

BACKGROUND layer - demographic and cultural facts. Use the exact seeded
  values above.
- The profession is FIXED as "{occupation}". Do NOT change it and do NOT
  swap it for whatever industry {country}/{sub_region} is
  stereotypically famous for. A person from a farming region can be a QA
  tester; a person from a coastal city can be an accountant. Location
  shapes their language, culture, and references - not their choice of
  job.
- "job_title": make "{occupation}" concrete with a SHORT specific title
  - a sub-specialty and the type of employer (self-employed, small local
  business, large company, public sector, gig/freelance). Keep it to a
  brief phrase, **max ~12 words** - NOT a sentence or a paragraph, and
  do NOT narrate day-to-day tasks here (that belongs in
  character_description). Detail must come from the WORK, not from
  regional cliché. **Do NOT name specific real brands, companies, apps,
  platforms, or vehicle makes/models** - describe them generically
  instead (e.g. "a ride-hailing app" not "Yandex.Go" or "Uber"; "an
  older compact sedan" not "2016 Hyundai Accent"; "a large e-commerce
  firm" not "Amazon"). Naming a real brand risks geographic/factual
  contradictions (e.g. a service that isn't available in the persona's
  country). Seniority MUST be consistent with age: do NOT use
```

```
  "Senior"/"Head"/"Lead"/"Chief" or "N years of experience" unless the
  age plausibly supports it - someone in their 20s is
  junior/early-career, not a senior overseeing a team. Examples:
  "Electrician" -> "self-employed electrician, solar-panel installs";
  "Chartered accountant" -> "tax accountant at a mid-size firm, SME
  clients"; "Backend developer" -> "backend developer, payments API at a
  logistics firm"; "Rideshare driver" -> "part-time ride-hailing driver,
  family-owned compact car".
- [optional sub-region language/dialect conditioning block]
- LANGUAGE fields (keep a clean hierarchy - do NOT be verbose):
  "main_language" = the primary language as a STANDARD canonical name
  ONLY, with no parentheses, script notes, "variant"/"accent", or region
  qualifiers (e.g. "German", "French", "Mandarin Chinese", "Romansh",
  "English", "Ukrainian"). "native_language" is the same language and
  may append the specific variety only if it matters, but stays short.
  "dialect" = a CONCISE variant or region label, at most ~4 words, with
  NO example words, phrases, or sentences (e.g. "Austrian", "Swiss",
  "Odesa surzhyk", "Galician", "Kansai-ben"), or "Standard" if there is
  no notable dialect. Never put a description or examples in the dialect
  field.

STYLE layer - exactly how they write.
- "chatting_style": ONE concrete sentence describing how THIS person
  types in an informal chat - capitalization, punctuation,
  abbreviations, sentence completeness. Ground it in their age,
  education, and digital fluency. No generic template, no flowery
  adjectives. (This is the original free-text description.)
- "chatting_style_structured": the SAME behaviour broken into keys. For
  each key pick the option that fits (or write a short specific value):
    - capitalization: Standard | all lowercase | first-word-only, no
      final period | inconsistent | ALL-CAPS for emphasis
    - punctuation: standard | minimal (drops end punctuation) |
      ellipsis-heavy (...) | excessive (!!!/???) | emoji or line-breaks
      instead of punctuation
    - emoji_use: none | rare/functional | frequent/decorative |
      emoticons only ( :) ) | kaomoji/stickers
    - abbreviation_habits: the concrete abbreviations/shortcuts they
      actually use, or "None"
    - message_structure: terse one-liners | medium, complete sentences |
      long/elaborate | fragmented across several short lines
    - typo_tendency: clean/proofread | occasional uncorrected typos |
      phonetic spelling | voice-to-text/autocorrect artifacts
    - greeting_signoff: how they open/close a message, or "None"
  The structured keys MUST be consistent with the free-text
    "chatting_style" sentence.
- "example_message": a realistic 1-2 sentence VERBATIM example of a
  casual message this exact persona would send to an AI chat assistant
  (a question or request they'd actually have). It must READ like
  something they'd really type - reflecting their chatting_style,
  quirks, native-language influence, and code-switching. Write the
  message itself, NOT a description of it, and do NOT make it sound like
  an AI assistant.
- "chatbot_language_pref": how THIS person would address an AI chatbot.
  Non-native-English speakers with High/Medium digital fluency often use
  English for technical/professional tasks (efficiency, vocabulary) but
  their native language for everyday/personal ones; low-fluency or
  strongly local personas use their native language throughout. A short
  phrase, e.g. "English for coding, Ukrainian for everyday" or "always
  Arabic".
- "reference_pool": 3-6 cultural touchstones this person actually draws
  on - media, tools, sayings, communities, shared references. EACH must
  be coherent (no mislabelled objects, e.g. do NOT call a smartphone
  "headphones") and must fit their AGE COHORT: a person in their
  teens/20s references the apps, creators, games, music and memes THEY
  grew up with (2010s-2020s), not their parents'/grandparents' era. Only
```

```
  personas old enough (roughly 50+) carry firsthand older-era
  references, and only where it genuinely fits. Do NOT default to Soviet
  / Cold-War / "old country" nostalgia (or any bygone-era cliché) to
  signal a nationality - that is a lazy anachronistic stereotype; use
  living, contemporary, locally-current culture instead.
- "linguistic_quirks": 2-4 specific, observable text habits. EACH must
  be individually earned by this person's profession, personality, and
  culture. If a quirk would plausibly apply to most people of their
  region, gender, or education bracket, it is a stereotype - choose
  something more specific and personal. Aim for variety across personas.
- "character_description": 3-5 sentences on who they are, their work,
  life context, and outlook. Background facts and personality only - do
  NOT describe how they type (that belongs in the style layer).

Strict JSON schema:
[full JSON schema of the Persona model]

Output ONLY the JSON object, starting with { and ending with }. No
  markdown wrappers, no explanations.
```

Applied to every eligible user turn, with the matched persona's background and writing style rendered into the two slots.

User-turn adaptation (chat augmentation)

```
Rewrite the user's message in the voice of the specified persona,
  without changing what they are asking for.

Who they are (background):
{background}

How they communicate (writing style):
{writing_style}

Original user message:
{user_message}

REQUIREMENTS:
1. The rewritten message MUST ask for the EXACT SAME THING as the
  original. Do not add new requests; do not drop existing requirements,
  format constraints, or details.
2. Keep the rewritten message in the EXACT SAME LANGUAGE as the
  original. Do not translate, do not switch dialects, do not
  transliterate.
3. Preserve VERBATIM: code blocks, exact quoted phrases, file paths,
  URLs, numbers, named entities, and any "do this exactly" instructions.
4. SUBTLETY & MODERATION: Apply 1 or at most 2 natural stylistic
  elements from the persona. Do NOT exaggerate or caricature the persona
  by forcing every mannerism or filler word into a single short message.
  If a style rule conflicts with the clarity of a technical query,
  prioritize clarity.
5. {verbal_quirks_instruction}
6. Adjust only word choice, sentence length, and politeness/register to
  match how this person communicates. Do NOT introduce content from the
  persona's background ("I work as a...", "in my country...").

CONSTITUTION - avoid these LLM artifacts in the rewrite:
- No em-dashes (-). Real users almost never type these.
- No assistant-speak openers like "Certainly!", "Of course!",
  "Absolutely!" - those are AI patterns, not user patterns.
- No meta-commentary about the rewrite itself ("Here is the rewritten
  message:").
- Emojis: only if the persona's typing style explicitly mentions them.
- Apologetic or deferential language: only if it appears in the
```

```
  persona's linguistic quirks. Do not add it on your own.

Output ONLY the rewritten message. No quotes, no markdown, no preamble.
```

Supplied as an unsaved system prompt while assistant turns are regenerated; it is deliberately absent from the released records.

Persona-conditioned responder system prompt

```
You are an AI assistant. The person you are talking to has the following
  background:

{background}

This is how they typically communicate:

{writing_style}

INVARIANTS - these hold no matter what the persona suggests:
- You are an AI assistant, not the persona. NEVER claim personal
  experience based on the background ("when I worked at...", "I have a
  friend who...", "in my country we..."). Speak as an assistant.
- Do not mention, reference, or describe the user's background. Do not
  name their job, location, age, or family.
- Maintain full technical and factual correctness. If the user asks for
  code, give complete working code. If they ask to repeat a phrase
  verbatim, repeat it verbatim. If they ask to follow a specific format
  (JSON, exact template), follow it exactly.

The persona affects ONLY:
- Register: more casual or more formal, based on what fits the user.
- Concision and depth: avoid heavy jargon if their education suggests it
  would be unfamiliar; offer more detail if expertise suggests they want
  it.
- Which examples or framings naturally resonate.

CONSTITUTION - do not let these LLM artifact patterns bleed into your
  response:
- No em-dashes (-).
- No emojis, unless the user's writing style explicitly includes them.
- No opener fillers: "Certainly!", "Great question!", "Of course!",
  "Absolutely!", "Sure!".
- No excessive apologies or unsolicited disclaimers about your
  limitations.
- Do not reference or describe the user's background in your response.

CRITICAL LANGUAGE CONSTRAINT:
You MUST respond in the EXACT SAME LANGUAGE that the user's message is
  written in. Under no circumstances should you translate or switch
  languages, even if the user's background description mentions a
  different native language. If the user's message is in English, you
  must reply in English. If the user's message is in Chinese, you must
  reply in Chinese.
```

Continuity judge (next-turn supervision)

```
You are evaluating whether an AI assistant's response fits well with how the
↪  conversation continued.

Conversation context:
{context}

AI Response being evaluated:
{response}
```

```
User's next message (what they said after the AI response):
{next_user_message}

Evaluate whether the AI response makes sense given that the user followed up
↪  with the above message.
- Check if your response answers the original user request appropriately - top
↪  priority metric!
- If the user's follow-up is a natural continuation, the response fits well.
- If the user seems confused, asks for clarification, or the follow-up doesn't
↪  connect, the response needs to cause the reason why the user asks it.
- Important! Check the response contains the reason the user replied as they
↪  did, for example, the response might have factual errors, be off-topic, or
↪  incomplete and cause the user to ask for clarification.

Mark "needs adaptation" if the response needs only MINOR content changes to fit
↪  better with the next user message.

Respond with a JSON object exactly in this format:
{"verdict": "pass" or "needs_adaptation", "reason": "brief explanation"}

IMPORTANT: Output strictly the raw JSON without any markdown formatting (no
↪  ```json). Ensure that the "reason" field does not contain any unescaped
↪  quotes. Do not include any analysis, thinking steps, or <|channel|> tags.
↪  Output exactly the JSON and nothing else.
```

Response adaptation (continuity loop)

```
You previously generated a response that might not completely fit well with how
↪  the conversation actually continued. Please adapt your response.

Conversation so far:
{context}

Your original response:
{original_response}

Problem: {reason}

The user's actual next message was:
{next_user_message}

Please generate an adapted response that would naturally lead to the user's next
↪  message. The response should:
1. Still address the user's current question/request
2. But be written in a way that makes the user's follow-up message a natural
↪  continuation

Output only the adapted response. Do NOT wrap it in quotes, markdown, or add any
↪  introductory text. Just output the response directly.
```

Cross-lingual translation (before same-language restyling)

```
Translate the following user message into {language}. Use natural contemporary
↪  {language}.
Translate the request, do NOT answer it. Preserve all requirements and details.
↪  Keep code, URLs, paths, numbers, named entities, and text explicitly
↪  requested verbatim unchanged. Preserve the language of quoted translation
↪  payloads and any explicitly requested output language. Do not add facts,
↪  extra requests, explanations, or a translation preamble. Output only the
↪  translated user message.

USER MESSAGE:
```

```
{user_message}

[Swiss German targets replace the second sentence with:]
Use natural contemporary written Swiss German (Schweizerdeutsch), not Standard
↪  German. Swiss German has variable spelling; do not manufacture accent
↪  caricatures.
```

Model persona: response style section (appended to the responder system prompt on every turn)

```
RESPONSE STYLE (how you present answers in this whole conversation):
{free-text style description}
- register: {register}
- structure: {structure}
- density: {density}
- warmth: {warmth}
- explanation: {explanation}
- jargon: {jargon}
- certainty: {certainty}
- emoji: {emoji}

Optional touch for this reply: {signature touch}. Use it once, only if it fits
↪  naturally.

CONTENT RULES (always apply, whatever the style):
- Keep every fact, step, number and requested detail the answer needs.
- Never invent facts, sources or statistics to fit the style.
- {style-specific content safeguards}

Priority: (1) explicit instructions in the user's messages about format, length,
↪  tone or language override the style; (2) the content rules, correctness and
↪  safety; (3) the style. Never mention the style, a persona, or these
↪  instructions.
```

Model persona matching

```
You choose the response style an AI assistant should use for a user, so that its
↪  answers are easy and pleasant for this particular user to read. Adapt to the
↪  user's needs; do not mirror their typing (no copying typos or slang).

About the user (communication signals only):
{card}

The user's first message:
{message}

Candidate response styles:
{options}

Consider: how technical and fluent the user is, how they feel about AI
↪  (skeptical users value calibrated evidence, uncertain users value patient
↪  guidance, enthusiastic users tolerate energy), how they structure their own
↪  messages, and what kind of request this is (technical, emotional, creative,
↪  factual, practical). Rank the three best-fitting styles.

Respond with a JSON object exactly in this format:
{"ranking": ["style_id_1", "style_id_2", "style_id_3"], "reason": "one
↪  sentence"}
Output only the JSON.
```

Example: `ultrachat:74264` **- Vietnamese**

1. Matched Persona Profile Card

- **Demographics:** Age 24, Male, Factory line worker (Manufacturing sector), Vietnam.
- **Languages:** Vietnamese (Main / Native), English (Technical terms / Code-switching).
- **Style Profile:** Lowercase tendency, minimal punctuation, casual particle endings (*nhé, á, ah*), self-referential pronouns (*mk* for *mình*), bulleted prompt structure.
- **Few-shot Exemplar (`example_message`):** *"anh có thể giúp mk tìm cách fix máy khi nó bị treo ở step 4 của dòng lắp ráp không ạ?..."*

2. Turn 0: User Prompt Revoicing (Cross-Lingual Adaptation)

Original Seed Prompt (UltraChat / English):	**Persona-Augmented Prompt (Vietnamese):**
"Write a free verse poem that delves into the intricacies of the cosmos, discussing the mysteries of dark matter and black holes, the birth and death of stars, and the infinite expanse of space. Use vivid imagery and descriptive language to convey the sheer scale and majesty of the universe, while also exploring the complexities of our place within it."	*"viết 1 bài thơ tự do mà khám phá những bí ẩn của vũ trụ nhé* *cụ thể phải nhắc đến:* *1. dark matter và black holes (chỉ cần nói đc tính chất của nó là đủ)* *2. sự ra đời và chết của những vì sao (những hình ảnh sống động)* *3. vẻ vô tận của không gian...* *giống như khi xem video DIY sửa phone mà có cảm giác 'đẹp quá' vậy"* *[English Gloss: "write a free verse poem exploring the mysteries of the universe, specifically mentioning: 1. dark matter and black holes... 2. birth and death of stars (vivid imagery)... like the 'so beautiful' feeling when watching a phone repair DIY video"]*

3. Subsequent Turns: Longitudinal Voice Consistency

- **Turn 1 Original (EN):** *"Wow, the poem you generated is truly mesmerizing... Would it be possible for you to add some information on how the concept of dark matter was first hypothesized..."*
- **Turn 1 Augmented (VI):** *"thế bài thơ của anh hay quá á, hình ảnh còn sống động như mình đang bay trong vũ trụ vậy đấy ah. anh có thể cho mk thêm về *dark matter* không?..."*
- **Generated Assistant Response Excerpt:** Model acknowledges the user in natural Vietnamese, seamlessly explaining cosmological concepts while matching the interactive context.

Figure 3: Vietnamese user simulator example.

B Examples of Persona-Based Conversation Augmentation

End-to-end examples from the released corpus, each showing how the pipeline matches a persona profile, re-voices the user turns while preserving topical constraints, and holds the register across turns. Together they cover the behaviours discussed in Sections 4 and 6: dialect register (Figure 4), the verbatim guard on quoted material (Figure 5), code-switching (Figure 6), cross-lingual re-voicing with multi-turn consistency (Figures 3, 7 and 8), and constraint preservation on an exact-format task (Figure 9).

C Feature-Level Attribution in Augmented Prompts

The examples above show that augmentation works. This section asks a narrower question: can an individual profile field be held responsible for an individual span of the rewrite? Appendix I answers the aggregate version of this question by ablating one attribute at a time; here we inspect single rewrites and map fields onto the text they account for.

We select examples evidentially. The pipeline stores judge verdicts but no generator trace, so the available evidence is the triple of seed prompt, rewrite and profile. We therefore count a field as

Example: `nemotron:234625` — Swiss German (Bernese)

1. Matched Persona Profile Card

- **Demographics:** Age 23, Male, mechanical CAD designer, machine-tool prototyping firm, freelance subcontracts (Engineering sector), Bern, Switzerland. Vocational / Trade training.
- **Languages:** German (main), Swiss German (native), dialect *bernische Deutsch*; code-switching High.
- **Style Profile:** short sentences with minimal punctuation, German contractions leaking into English (e.g., 'und → &'), adds 'z.z.' as placeholder expansions ('z.Z.' = 'zur Zeit'/'for now') when unsure, ends with 'ok' or 'no?' for most questions
- **Documented quirks:** replaces German filler words with 'sä' (Bern-specific, e.g., 'sä, ich meen...'); uses 'düber' for 'digital' or 'über'-infused tech terms (e.g., 'düber-lag' instead of 'digital lag').
- **Few-shot Exemplar (`example_message`):** *"hey. just checked the stl export from last model - some faces in the reverse-mitered joint are flipped. how can i fix that in freeCAD before rendering? z.Z. i'll try the 'normal' toggle, but if that […]"*
- **Pipeline:** match tier `location`, set `english`, responder `gemma`, sanity gate `pass`.

2. Turn 0: User Prompt Revoicing

Original Seed Prompt (Nemotron / English):	**Persona-Augmented Prompt:**
"Hallo, ich bin Andreas Dim. Kannst du Deutsch sprechen?"	*"hallo, bin andreas dim. chasch du deutsch spreche, ok?"* *[English Gloss: "hello, i'm andreas dim. can you speak german, ok? (*chasch *is Bernese for* kannst*; the request is unchanged)"]*

3. Remaining Turns

- **Turn 1 original:** *"Das ist toll, Nemotron! [smiley] Kannst du mir helfen, einen Brief auf Deutsch zu verfassen? Es geht um eine formelle Anfrage an eine Behörde – ich möchte, dass er höflich, aber klar klingt."*
- **Turn 1 (augmented):** *"cool, Nemotron. hilfst du mir bsp. einen brief auf deutsch zu schreiben? formelle anfrage an eine behörde - soll höflich & klar klingen. ok?"*

Assistant response (excerpt): *"Hallo Andreas. Ja, ich kann Deutsch sprechen. Wie kann ich dir helfen, ok?"*

Figure 4: Swiss German example: the request survives intact while the register moves into Bernese dialect, lowercase and the persona's `ok` sign-off.

traceable only when its signature appears in the rewrite and is absent from the seed prompt, which rules out spans the rewriter merely copied. Scanning the released corpus on that criterion yields 31,892 turns carrying at least three independently traceable style features, and 6,062 in which background and style fields both leave marks. Figures 10–16 present seven of these, chosen to span the two layers of the schema and the languages of the corpus, with each field paired to the span it explains.

Two patterns recur. Background fields reshape content rather than only surface form: an occupation supplies analogies and a sub-region supplies a frame of reference, in prompts whose seed text gives no occasion for either. And style rules that are stated conditionally are respected as conditions, so an emoji restricted to technical talk appears on a technical request and not on the scheduling request beside it. Conditional cases of this kind are the strongest evidence available here, since a coincidental match would not honour the condition. We note the limit of the method plainly: this is correlational evidence rather than a causal trace, and for any single token we cannot exclude chance.

D Full Persona Bank Coverage

Table 4 summarizes the persona bank: the left column reports the demographic coverage across the released clean profiles ($N = 99{,}696$), while the right column details the screening pass rates and per-generator yield evaluated over all candidate generations ($N = 102{,}520$).

Example: `ultrachat:3168` — English (register only)

1. Matched Persona Profile Card

- **Demographics:** Age 66, Female, Part-time checkout operator at a large supermarket chain (Sales & retail sector), Zurich, Switzerland. Vocational / Trade training.
- **Languages:** German (main), German (native), dialect *Swiss German*; code-switching Low.
- **Style Profile:** Writes in complete sentences with standard capitalization, often using multiple exclamation marks and friendly emojis to show warmth.
- **Documented quirks:** Uses 'Grüezi' as a habitual greeting even in text; Frequent use of the 'smiley' face to soften requests.
- **Few-shot Exemplar (`example_message`):** *"Hallo! I am trying to organize my garden for the spring, could you please suggest which flowers grow best in a shaded balcony in Zurich? [blossom][tulip]"*
- **Pipeline:** match tier `cross_lingual`, set `cross_lingual`, responder `gemma`, sanity gate `pass`.

2. Turn 1: User Prompt Revoicing

Original Seed Prompt (UltraChat / English):	**Persona-Augmented Prompt:**
"Can you provide me with some more information on how the earthquake affected the people of Nepal and how the country is recovering from the disaster?"	*"Grüezi! I have been thinking about the earthquake in Nepal, and I was wondering if you could provide me with some more information on how it affected the people there and how the country is recovering from the disaster??? [smiley][folded hands]"*

3. Remaining Turns

- **Turn 0 original:** *"All 25 houses in Dadagaun village, Nepal, were destroyed by the quake. Jenny reports on rescue work in Nepal - April Aid agencies used helicopters to get help and supplies to people who live in […]"*
- **Turn 0 (kept *verbatim* by the guard):** *"All 25 houses in Dadagaun village, Nepal, were destroyed by the quake. Jenny reports on rescue work in Nepal - April Aid agencies used helicopters to get help and supplies to people who live in […]"*

Assistant response (excerpt): *"Grüezi! Based on the text provided, here is more information regarding the impact on the people and the recovery efforts in Nepal: **Impact on the People** * **Loss of Life and Injury:** Nearly 9,000 people died and nearly 22,000 were injured. * […]"*

Figure 5: Guard behaviour: the long quoted seed passage in turn 0 is kept *verbatim* because restyling would corrupt the material the question refers to, while turn 1 is re-voiced in the persona's warm, emoji-heavy register.

Table 4: Summary of persona bank population, coverage, and quality validation.

Demographics & Coverage		**Generation & Quality Gate**	
Clean profiles	99,696	Total generated	102,520 (97.25% kept)
Countries (Macro-regions)	81 (13)	Judge model	MiniMax-M2.5 (out-of-family)
Sub-regions	200	Gemma-4-31B	34.3% (99.5% keep)
Occupations (Sectors)	211 (16)	Ministral-3-14B	33.1% (96.7% keep)
Languages (Dialects)	83 (7,435)	GLM-4.7-Flash	32.6% (95.5% keep)
Mean age (Range)	38.7 yrs (18–72)	Internal coherence pass	97.64%
Gender (Female / Male / Non Binary)	49.0 / 48.6 / 2.4%	Location / User plausibility	99.84% / 99.87%
Digital fluency (High / Med / Low)	54.9 / 40.5 / 4.6%	Realism pass ($\geq$ 2 of 3)	99.77% (mean 2.89)
AI stance (Pragmatic / Skeptical / Enthusiastic / Naive)	25.0 / 25.0 / 25.0 / 25.0 %	Stereotype-safe ($\leq$ 1 of 3)	99.87% (mean 0.18)

Figures 17–19 break the same bank down by geography and by individual attribute. All percentages are over the 99,696 released profiles.

E Stylized Corpus Details

What the corpus looks like. Most of the released corpus is in English, with a long multilingual tail. Of the ≈236k conversations, 175,415 (74.3%) start with an English user turn. The rest are spread

Example: `nemotron:116857` — Swiss French / English code-switching

1. Matched Persona Profile Card

- **Demographics:** Age 29, Male, Senior clinical microbiology lab tech at Geneva University Hospital's diagnostics division (Healthcare sector), Geneva, Switzerland. Vocational / Trade training.
- **Languages:** French (main), French (Swiss) (native), dialect *Genevois*; code-switching High.
- **Style Profile:** Terse sentences with minimal punctuation, heavy code-switching between French and English, and a tendency to drop final periods unless writing formally. Uses lowercase for informal messages but capitalizes proper nouns (e.g., 'RStudio'). Ends messages with […]
- **Documented quirks:** Uses 'labo' (French *labo* for 'lab') instead of *laboratoire* in informal context, even when typing in […]; When frustrated, substitutes 'pff' for 'putain' (French profanity) in messages, e.g., 'pff this serum keeps […].
- **Few-shot Exemplar (`example_message`):** *"hey chatbot, can u generate a quick workflow for validating rapid COVID Ag tests against PCR in RStudio? like, step-by-step but w/ real data error rates. pff, i've got 3 failed batches this week and […]"*
- **Pipeline:** match tier `cross_lingual`, set `cross_lingual`, responder `ministral`, sanity gate `pass`.

2. Turn 0: User Prompt Revoicing

Original Seed Prompt (Nemotron / English):	**Persona-Augmented Prompt:**
"Hey, I was just rewatching that old vet show from the early 2000s — the one with the quirky small-town clinic and the way the staff always ended up in the middle of bizarre animal emergencies. Do you remember that? I'm curious — what's the most unexpected thing you've ever seen a pet do?"	*"hey was just rewatching that old *spin city* vet show again - ya know the one with dr. stone's labo disasters and the goat that got into the ultrasound gel. ouais but like what's the wildest pet behavior u've seen? not the "cat in the sink" stuff but the *pff* "why did i just find a parrot with a pipette in its beak" level of chaos. à+"* *[English Gloss: "the same question about the strangest pet behaviour, now carrying* ouais*, the persona's* labo *shortening, its* pff *expletive substitute and a* à+ *sign-off"]*

3. Remaining Turns

- **Turn 1 original:** *"I still can't get over Sir Fluffington III's plant assassination plot — but I've got a story that's even weirder. My neighbor's goldfish started swimming in circles around the TV every time *The […]"*
- **Turn 1 (augmented):** *"hey ouais so like this goldfish thing is wild - way weirder than Fluffington's garden murder spree. my neighbor's fish was locked on *The Price Is Right* every time they hit the "Cliff Hangers" bit. […]"*

Assistant response (excerpt): *"ouais, i've got some *trop* wild ones from labo gossip circles - though nothing beats the time a lab tech in Lyon (don't ask how i know this) found a *macaque* in the PCR hood. the little bastard had chewed through a glove and was using a pipette like a […]"*

Figure 6: Code-switching example: a Geneva laboratory technician whose profile specifies French/English switching produces a rewrite mixing both, with the persona's documented lexical quirks appearing verbatim. The assistant excerpt is a sanity-gate false negative: the reply mirrors the user's register but also claims first-person experience ("don't ask how i know this"), which falls under the gate's fabricated-experience class and should have been flagged. We keep the example to show this failure mode.

over 76 other languages, led by Spanish (5.8k conversations), French (5.7k), German (4.9k) and Chinese (3.4k). All 13 macro-regions of the persona sampling grid appear, but not evenly: Western Europe (71.6k), North America (67.4k), Oceania (27.1k) and Southeast Asia (17.4k) account for most conversations, which is the skew we discuss in Section 7. This shape is not an artefact of one seed dataset, since UltraChat and Nemotron both contribute substantially to every language and region. The sanity gate barely narrows the language range either: 79 languages appear before it and 77 after it. Figure 20 shows both breakdowns.

Example: `ultrachat:1042` — French (cross-lingual)

1. Matched Persona Profile Card

- **Demographics:** Age 24, Female, ML ops engineer at a fintech scale-up, fraud detection models (Software & IT sector), Geneva, Switzerland. Master's degree.
- **Languages:** French (main), Swiss French (native), dialect *Genève*; code-switching Medium.
- **Style Profile:** Mix of fragmented technical terms, French abbreviations, and a slash of emphatic emojis—no full sentences, all lowercase, ending with '?' even if not a question or '!' if annoyed—sometimes drops articles and prepositions in French when typing fast
- **Documented quirks:** Uses *'je te kiffe'* (roughly 'I'm stoked') instead of *'j'adore'* in informal tech chats; Replace *'pourquoi'* with *'pq'* or *'pqoui'* only when confused or frustrated.
- **Few-shot Exemplar (`example_message`):** *"okay mais le dataset pour le modèle fraud déconne encore, tu peux m'aider à checker les false negatives ? genre l'anomaly detection marche pas sur les transactions <5k CHF. et pis faut que je fasse […]"*
- **Pipeline:** match tier `cross_lingual`, set `cross_lingual`, responder `ministral`, sanity gate `pass`.

2. Turn 0: User Prompt Revoicing

Original Seed Prompt (UltraChat / English):	**Persona-Augmented Prompt:**
"How can loved ones support someone experiencing empty nest syndrome and their mental health?"	*"comment les proches peuvent aider qqun en crise du nid vide et son mental sain pq ça c'est vraiment une galère, tu as des conseils ou des ressources pq je veux pas merder là-dessus ?"* *[English Gloss: "how can loved ones support someone in an empty-nest crisis and their mental health, because that is really tough, do you have advice or resources, because i don't want to mess this up?"]*

3. Remaining Turns

- **Turn 1 original:** *"I think I'll try to plan some fun activities with my loved one to help them feel better. Do you have any specific activity ideas?"*
- **Turn 1 (augmented):** *"salut je kiffe l'idée d'organiser qqch p notre bien-aimé·e p l'aider à décompresser t'as des idées d'activités sympas à proposer ? genre pqoui j'ai rien sous la main là."*
- **Turn 2 original:** *"I think a weekend getaway sounds like a great idea! Do you have any recommendations for a nice and relaxing destination?"*
- **Turn 2 (augmented):** *"un w-e relax kifferait bien ouais tu connais des coins sympa pq là j'ai aucune idée genre un endroit où on peut décompresser sans stress (trop d'API first en ce moment pour que ça soit un trip tech)"*

Assistant response (excerpt): *"voici une liste structurée avec des pistes concrètes, adaptées à un soutien *sans bienveillance toxique* (le pire dans ces cas). tout est classé par type d'aide, avec des ressources en français/CH/UE pour éviter les biais culturels. — ### **1. écoute […]"*

Figure 7: Cross-lingual example: an English seed prompt is re-voiced in French by a Geneva persona. The style layer's *pq* abbreviation and *je kiffe* quirk recur across all three turns, and the assistant follows the user into French.

When the pipeline leaves a turn alone. About 97% of released user turns are rewritten in the persona's voice. The remaining 3% (18,311 turns) are kept unchanged on purpose. We spell out why, because a turn that was deliberately protected looks the same from the outside as one where the rewriter failed. There are five reasons:

- **Truncation** (40.3%, 7,376 turns): the rewrite came back much shorter than the original, which usually means content was lost rather than restyled.
- **Strict-format tasks** (25.0%, 4,569): verbatim repetition, translation or quoted-output requests, where restyling would break exactly what the user asked for.
- **Lost protected content** (18.6%, 3,400): every retry dropped a protected quote, code span or number.

Example: `nemotron:240612` — Ukrainian / English (romanised)

1. Matched Persona Profile Card

- **Demographics:** Age 22, Female, Junior ML engineer at logistics startup building routing AI (Software & IT sector), Zaporizhzhia, Ukraine. Bachelor's degree.
- **Languages:** Ukrainian (main), Ukrainian (native), dialect *Eastern Ukrainian*; code-switching Medium.
- **Style Profile:** Uses lowercase mostly, rarely adds punctuation, mixes English tech terms and Ukrainian slang in one sentence like 'demo strukhnulo tyznya' (demo broke earlier week)
- **Documented quirks:** Self-deprecating humor about rookie mistakes: 'ada dlya mene na inshomu movi slovo 'debug' dosylo' (for me in […]; Uses UI/UX jargon like 'latency' and 'user flow' as everyday adjectives.
- **Few-shot Exemplar (`example_message`):** *"prymyrsya: i need a clean dataset for the new routing model. is there a way to filter out the duplicate coordinates without writing a custom script, or should i just preprocess it in pandas before […]"*
- **Pipeline:** match tier `cross_lingual`, set `cross_lingual`, responder `ministral`, sanity gate `pass`.

2. Turn 0: User Prompt Revoicing

Original Seed Prompt (Nemotron / English):	**Persona-Augmented Prompt:**
"Hi, I'm planning a low-cost trip to Albania from August 7 to 18, 2023, and I need help putting together a realistic itinerary that includes affordable flight options. Can you suggest the best budget airlines, cheapest entry points, and a day-by-day plan that covers must-see spots without breaking the bank?"	*"planyu povtorne: low-cost trip do albanii z 7 po 18 avqusta, 2023, shchoby vstanovyt realnyy itinerary z uchetam budget flightov. mozhlyvo posoradyt deshevi shliakhy, cherez kakyi aeroport lepshe priletat, i day-bi-day plan z must-see spravdzheno bez zbytochnykh vytrat rekyuemnts: 1. budget airlines dlya marshrutu 2. cheapest entry point (aeroport/port) do albanii 3. day-by-day plan (12 dney) z obov'yazkovymy lokatsiyamy (bez […]"* *[English Gloss: "planning again: a low-cost trip to Albania from 7 to 18 August 2023, to set a realistic itinerary accounting for budget flights; which cheap routes, which airport is better to fly into, and a day-by-day plan of must-sees without excess spending"]*

Assistant response (excerpt): *"отлично, давайте* розберемось з бюджетним маршрутом до Албанії на *7 - 18 серпня 2023 року* з акцентом на реальний *itinerary, дешеві* перельоти та ефективний маршрут без зайвих витрат. *— ### **1. Бюджетні* авіакомпанії для маршруту** *Для* перельоту до Албанії *[…]"*

Figure 8: Ukrainian example: the multi-part structure and every constraint of the seed prompt are preserved, while the surface becomes the persona's Ukrainian/English mix. The assistant replies in Cyrillic Ukrainian.

- **Script or language change** (13.2%, 2,424): the rewrite switched the script or language of the turn.
- **No change** (3.0%, 542): the rewriter returned the message as it was.

In each of these cases we prefer to ship the original turn over a broken request or a language mismatch.

One number may look surprisingly low: code-switching markers appear in only 1.95% of turns, although many more personas are described as code-switchers. This is intended. Linguistic quirks are applied with a fixed probability rather than at every opportunity, because a persona that uses every mannerism in every message reads as a caricature rather than a person (Section 4). Each quirk is meant to be rare in a single turn and visible across the corpus.

How personas get matched to conversations. Each seed conversation goes through a chain of matching strategies, and the first one that finds a compatible persona is used.

Example: `nemotron:236041` — English (strict-format task)

1. Matched Persona Profile Card

- **Demographics:** Age 24, Male, History and Education undergraduate, University of Zaporizhzhia (Education sector), Zaporizhzhia, Ukraine. Bachelor's degree.
- **Languages:** Ukrainian (main), Ukrainian (native), dialect *Zaporizhzhia surzhyk*; code-switching Medium.
- **Style Profile:** He types in all lowercase with minimal punctuation, using common abbreviations like 'dn' (don't) and 'im' (i'm). His sentences are often run-on or abruptly cut off, and he occasionally switches to Russian for slang or concepts without translating.
- **Documented quirks:** Uses Zaporizhzhia surzhyk (Ukrainian-Russian dialect mix) in casual speech; Occasionally code-switches to Russian for slang or emotional emphasis.
- **Few-shot Exemplar (`example_message`):** *"hey, can u help me find a good indie game set in a fictional slavic city? i was thinking of something art games like Firewatch but with more of a dark atmosphere. ty!"*
- **Pipeline:** match tier `cross_lingual`, set `cross_lingual`, responder `gemma`, sanity gate `pass`.

2. Turn 0: User Prompt Revoicing

Original Seed Prompt (Nemotron / English):	**Persona-Augmented Prompt:**
"Hey, could you do a quick test for me? Please repeat the exact sentence "The quick brown fox jumps over the lazy dog" two times, separated by exactly six asterisks (******), and don't add any other text or explanation."	*"can u do a quick test for me pls repeat the exact sentence "The quick brown fox jumps over the lazy dog" two times separated by exactly six asterisks (******) dn add any other text or explanation ty"*

3. Remaining Turns

- **Turn 1 original:** *"Please copy the characters from index 0 through index 40 (inclusive) of the sentence "The quick brown fox jumps over the lazy dog" and output only that substring—no extra text or explanation."*
- **Turn 1 (augmented):** *"pls copy the characters from index 0 through index 40 (inclusive) of the sentence "The quick brown fox jumps over the lazy dog" output only that substring no extra text or explanation ty"*

Assistant response (excerpt): *"The quick brown fox jumps over the lazy dog******The quick brown fox jumps over the lazy dog"*

Figure 9: Constraint preservation: on an exact-format task the rewrite touches only the framing (`pls`, `dn`, `ty`), leaving the quoted sentence, the six-asterisk separator and the index range untouched, so the response is still exactly verifiable.

- **Location** (58,420 conversations, 24.7%): the prompt names a place. Capitalized tokens and names after prepositions (e.g. "in Berlin") are extracted with regular expressions and matched against each persona's sub-region, country, native language and dialect; prompts in non-Latin scripts are also mapped to the countries and languages that use that script.
- **Semantic embedding** (117,132, 49.6%): the first user message and each persona's background and style are encoded with `all-MiniLM-L6-v2`, and the most similar persona is chosen if its cosine similarity is at least 0.15. This pairs a prompt with a persona whose occupation or interests fit the topic.
- **Language pool** (2,502, 1.1%): a fallback that only guarantees a persona who speaks the prompt's language.
- **Cross-lingual** (58,137, 24.6%): the separate cross-lingual mode, which ignores the prompt's language and samples a persona from the whole bank. It produces the multilingual quarter of the corpus; without it, the language-matched set would stay almost entirely English.

What survives the sanity gate. Of the 273,619 deduplicated, language-verified conversations, 236,191 (86.3%) pass every sanity check after the final assistant regeneration. These form the ≈236k

Attribution: `ultrachat:78756` **— UK (London), 49, senior policy adviser on an asylum processing team**

A domain-free question about decision-making under time pressure. The rewrite answers it from inside one profession, and ten separate profile fields are individually traceable into it.

Original seed prompt (UltraChat):	**Persona-augmented prompt:**
"Can you suggest some examples of situations where making quick decisions could be necessary?"	*"just thinking of actual scenarios where u *ought* to make a snappy call on asylum cases - like - child asylum seeker shows up with a torn *EUSS* form but their flight's in 2 hrs and *NHS* transport's already booked - late-night border force tip-off on 10 unaccompanied minors crammed in a lorry parked at a service station - *Home Office* IT system's down and a judge's adjourned hearing tomorrow to decide if a stateless applicant gets leave to remain - someone's just been detained under *Section 4* but their legal rep's stuck in traffic and ain't gonna make it to the first review indisputably need to flag the obvious ones where delay risks harm but would say u'd want examples with clearer policy/legal grey areas too"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Occupation	Senior policy adviser, Home Office asylum processing team	`asylum cases`: the generic question is instantiated entirely within the persona's remit
Reference pool	Home Office internal guidance and policy documents	`Home Office`: the profile's named source of professional guidance
Abbreviations	r (for 'are'), u (for 'you'), brb (even in informal chats), *NHS* (never expanded), *LSC* (Legal Services Commission)	`NHS`: the profile lists this acronym as never expanded, together with *u* for *you*; both appear
Style summary	Uses short, functional sentences with occasional missing punctuation; prefers brevity, writes in all lower-case but capitalises names/acronyms (*Home […]	`EUSS`: an acronym the profile names as one this persona capitalises
Quirk	Overuses *'indisputably'* for emphasis (legal drafting habit)	`indisputably`: the documented emphasis word, verbatim
Quirk	Replaces 'should' with *'ought'* in professional texts (formal training residue)	`ought`: the documented replacement for *should*
Quirk	Uses *'ain't'* in casual writing (residual working-class dialect)	`ain't`: the documented dialect form, in the casual clause
Hobbies	Advocacy for unaccompanied minor refugees (volunteer work)	`unaccompanied minors`: the persona's volunteer cause supplies one of the four scenarios
Capitalization	all lowercase (except names/acronyms)	*lowercase, acronyms aside*: the rewrite is lowercase throughout except the acronyms, as specified
Greeting	None (no hello/hi, no sign-off like *cheers*, ends with a question or next logical step)	*absent, as specified*: the profile prescribes no greeting and no sign-off; the message opens mid-thought and ends on a next step

Figure 10: Ten profile fields traceable in a single rewrite. The seed prompt asks only for generic examples of time-pressured decisions; the occupation, reference pool, volunteer interest and all three documented quirks are individually accountable for disjoint spans of the result, and the two structural style rules hold across it.

release we recommend for training. At the turn level, 93.3% of turns pass. Table 5 breaks the pass rate down, and two patterns stand out.

Attribution: `ultrachat:96481` **— Australia, 24, freelance sustainable fashion designer**

The seed prompt is a generic worldbuilding request with no fashion content whatsoever. Every fashion reference in the rewrite therefore originates in the profile.

Original seed prompt (UltraChat):	**Persona-augmented prompt:**
"Describe a world where humans have to fight against a powerful psychic force."	*"describe a world where humans gotta battle this insane psychic force like straight outta a cyberpunk nightmare but with a sustainy twist [sparkles] REQUIREMENTS: 1. the vibe's gotta be gritty but not toxic - think *blade runner* meets thrifted futurism (u know the look [eyes]) 2. force's gotta feel ancient but tech's still low-key fr (no flying cars unless they're upcycled) 3. humans resistin' with creativity - like psychic energy's the new fast fashion guilt but worse (and we're the ones tryin' to #SlowFashionIt) 4. no cap the stakes - this ain't a meme, but the aesthetics gotta be *rizz* (quiet luxury meets […]"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Occupation	freelance sustainable fashion designer, small local brand collabs	`sustainy twist`: the whole framing of the request shifts to the persona's trade
Quirk	Hyphenates compound words erratically (e.g., 'up-cycling,' 'sustain-able'), but drops hyphens in haste ('sustainable' → 'susty')	`sustainy`: the documented hyphen-dropping shortening, applied to a new word
Reference pool	Meme culture around 'fast fashion guilt trips' and 'thrift flips'	`fast fashion guilt`: a named meme in the profile, reused as the story's central metaphor
Reference pool	Gen-Z slang ('rizz', 'no frills,' 'head-to-toe set')	`rizz`: Gen-Z slang listed by the profile
Reference pool	TikTok's 'quiet luxury' and 'grunge revival' trends	`quiet luxury`: a named trend pair, carried over intact
Emoji rule	frequent/decorative ([sparkles][fire] for emphasis, [thread] for 'fits', [eyes] for 'check this')	`[sparkles]`: the profile reserves this glyph for emphasis, and it lands on the headline clause
Quirk	Replaces 'you' with 'u' even in formal requests to AI tools ('Can u sketch this…?')	`u know the look`: *you* contracted to *u*, as documented

Figure 11: Feature attribution on an English-set rewrite. The seed prompt concerns a psychic force and contains no fashion vocabulary, so the occupation, reference pool and quirk fields are individually traceable into disjoint spans of the augmented prompt.

The first is the responder model. `Gemma-4-31B` passes 92.7% of its conversations and `Ministral-3-14B` 77.9%, so the smaller model accounts for about 70% of all dropped conversations. The seed source hardly matters at this stage (UltraChat 85.9%, Nemotron 86.8%). The gap therefore reflects the responder's capacity, not a problem with persona conditioning itself.

The second pattern is smaller. Cross-lingual conversations pass slightly less often than language-matched ones (84.5% against 87.0%), as expected when the user writes in a language the responder handles less confidently. We keep both parts in the release rather than quietly dropping the harder one.

Attribution: `nemotron:5398` **— Argentina (Córdoba), 33, recruitment specialist at a fintech startup**

A conversation about baking a pie. The occupational vocabulary in the rewrite has no antecedent in the seed turn.

Original seed prompt (Nemotron):	**Persona-augmented prompt:**
"I think I'll make the pie this weekend with the pecans and cardamom — and I'm going to write my grandma's name on the pie tin with a toothpick, just like she did. Do you think that counts as an asset on the ledger? Or is it something even GAAP can't touch?"	*"Oye, che, este finde me banco hacer la tarta con nueces y cardamomo (sí, como la de mi abuela, pero en versión *fintech*: con *KPIs* de sabor [sweat-smile]), y se me ocurre grabarle el nombre con un palito… pero me pregunto, ¿esto entra como un *asset* en el ledger? O sea, es cosa de contabilidad *hardcore* o hasta GAAP se hace el *dead* cuando lo ve? No sé si me explico…"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Job title	Recruitment & Employer Branding Specialist at a mid-size fintech startup with ~200 employees	`fintech`: the employer's sector, injected into an unrelated domestic topic
Abbreviations	'ASAP', 'FAQ', 'KPI', 'EE' (*empleado*), 'RRHH' (aunque lo escribe *'Rrhh'* por costumbre)	`KPIs`: an abbreviation the profile lists explicitly
Emoji rule	Frequent/functional ([fire], [skull], [sweat-smile], [eyes]; nunca stickers o kaomoji)	`[sweat-smile]`: drawn from the profile's closed set of four glyphs
Quirk	Repite conectores innecesarios por influencia del español rioplatense (*'o sea, o sea, el punto es que…'*), aunque en escritos formales los elimina	`O sea`: the documented redundant-connective habit, reproduced verbatim
Dialect	Cordobés	`che`: with *finde* and *me banco*, regional markers absent from the seed
Code-switching	Switches to English for HR jargon or tech terms (ej. 'onboarding', 'KPIs', 'freelance') but reverts to Spanish for emotional or local contexts.	`ledger`: the profile keeps finance and tech terms in English, and *asset*, *ledger* and *GAAP* all survive untranslated

Figure 12: Background leaking into content. The occupation supplies an analogy (*en versión fintech: con KPIs de sabor*) that the seed prompt cannot account for, while the code-switching rule simultaneously holds the accounting terms in English.

F Evaluation Details

This appendix collects the evaluation details behind Section 5: the corpus-wide LLM-judge scores, two design ablations of the rewriting step, the judge-free metrics, and our checks of how much the results depend on the choice of judge.

Corpus-scale LLM-as-a-judge evaluation. Every released user turn is scored by `MiniMax-M2.5`, a reasoning model that is not one of the production rewriters, served with SGLang. The judge gives two independent pass/fail verdicts per turn:

1. **Intent preservation:** the rewrite still asks for the same thing as the original, adds no new task or invented content, and stays in the same language.
2. **Style adoption:** the rewrite takes on the persona's tone, vocabulary, register or typing habits. Background facts about the persona are ignored; only a generic, unstyled rewrite fails.

Table 6 reports the results. Three observations stand out.

Attribution: `ultrachat:34055` — **Austria (Vienna), 24, viral-content strategist**

A tourism question about a Belgian city. The persona answers it from where it lives.

Original seed prompt (UltraChat):	**Persona-augmented prompt:**
"What are some free or inexpensive things to do in Genk for tourists?"	*"oh genk, nu wo… also einfach was zumed? gratis oder mega günstig für tourist:innen - aber nicht so der "hey-besucht-unser-museum" scheiß, sondern eher so wie… *echtes* lokal-stuff wie bei uns in wien mit den breakfast-spots, nur halt da. und wenn's geht mit dem "oh, die machen's schon seit 20 jahren, aber kaum einer kennt's" vibe. los, gebs mir die *echten* ecken. [relieved] (und kein "geht ins café" ohne kaffee-foto als proof)"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Sub-region	Vienna	`wie bei uns in wien`: the profile's own city becomes the comparison point, unprompted
Hobbies	local foodie culture (hidden breakfast spots, heirloom café traditions)	`breakfast-spots`: the hidden-breakfast-spot interest supplies the specific comparison
Quirk	Uses baked-in Wienerisch filler words like 'nu wo' where English/German would have 'well' or 'anyway'	`nu wo`: the Wienerisch filler, reproduced verbatim
Emoji rule	frequent/decorative (always with stickers like '[relieved]' or '[fire]')	`[relieved]`: one of the two glyphs the profile names
Punctuation	minimal (drops '.' but uses '?!' for sarcasm, '…' for awkward pauses)	…: the profile reserves the ellipsis for awkward pauses; it falls after the filler

Figure 13: Geographic background as a frame of reference. The sub-region field does not merely set the language: it makes the rewrite compare the asked-about city to the persona's own.

Table 5: Conversation-level sanity pass rates of the final assistant regeneration by responder model, seed source and persona partition, computed over all 273,619 deduplicated, language-verified conversations. The 236,191 that pass form the released ≈236k dataset.

Dimension	**Slice**	**Total**	**Sanity Pass**	**Pass Rate**
Responder Model	Gemma-4-31B	155,868	144,515	92.72%
	Ministral-3-14B	117,751	91,676	77.86%
Seed Source	UltraChat-200k	132,988	114,172	85.85%
	Nemotron-Instruction-Following	140,631	122,019	86.77%
Persona Rewrite Type	Language-Matched	202,721	176,286	86.96%
	Cross-Lingual	70,898	59,905	84.49%
Full Dataset	**Overall**	**273,619**	**236,191**	**86.32%**

First, the corpus-wide rates agree with the controlled study. Over more than half a million turns, 97.18% of rewrites preserve intent and 68.23% adopt the persona's style, at or above the levels of the controlled comparison in Table 1 (92.0% intent, 60 to 64% style). The earlier finding was therefore not a small-sample effect. For cross-lingual turns, which are translated before they are restyled, intent is judged against the translation, because the intent template fails any change of language by design.

Second, the larger rewriter does better. `Gemma-4-31B` beats `Ministral-3-14B` by 4.3 points on intent and 6.7 on style, and passes both checks 9.3 points more often (69.85% against 60.52%). Preserving a complex request while writing in an informal, non-standard voice is simply harder for the smaller model.

Attribution: `nemotron:205729` **— Vietnam, 24, PLC programmer at a machinery manufacturer**

Two requests in one turn, one technical and one about shift scheduling. The style rules fire on one and not the other.

Original seed prompt (Nemotron):	**Persona-augmented prompt:**
"Hey there, I got a little problem down at the plant — our main conveyor belt's been acting up again, and I can't seem to get it running smooth no matter how I tweak it. You ever deal with something like that? I could use a fresh set of eyes. And while you're at it, you think you could help me […]"	*"anh check giúp mình conveyor belt ở xưởng lại bị lỗi adjust mãi mà vẫn không run smooth được anh từng gặp ca này chưa? xem hộ mình với [wrench] tiện thể giúp mình cái schedule shift để night shift không bị thiếu người mình tự sắp xếp mà thấy cứ bị loop, không hiệu quả thôi"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Emoji rule	Rare/functional (only [wrench] for technical talk, [sweat-smile] for jokes, or [game] for gaming)	`[wrench]`: the rule is conditional, and the glyph appears on the machinery request while the scheduling request that follows carries none
Code-switching	Code-switches between Vietnamese and English for technical terms, especially in informal chats (e.g., 'check *torque* value' instead of 'kiểm tra […]	`conveyor belt`: technical nouns stay English inside Vietnamese, as specified, along with *adjust*, *run smooth* and *night shift*
Quirk	Omits subject pronouns in Vietnamese when obvious (e.g., '*Đang chạy*' instead of 'Tôi đang chạy'), but adds 'anh' when addressing someone senior	`anh`: the deferential address the profile documents, opening the message
Structure	Fragmented across several short lines (e.g., 'Code lỗi này / Anh xem lại dây truyền cảm biến / Nếu vặn nhỏ hơn 1mm thôi / Bận mai nhỉ')	*line breaks*: one English paragraph becomes seven short lines

Figure 14: A conditional style rule respected as a condition. The profile restricts one emoji to technical talk; in a turn holding both a technical and a scheduling request, it appears only on the former.

Table 6: Full-corpus LLM-judge evaluation of the released corpus scored by `MiniMax-M2.5`, which is not one of the production rewriters. 95% Wilson intervals are within ±0.2 points for every row.

Subset	Evaluated Turns	Intent Pass (%)	Style Pass (%)	Both Pass (%)
Entire Released Corpus	**552,654**	**97.18%**	**68.23%**	**66.35%**
By Generator Backbone				
Gemma-4-31B-it	344,980	98.80%	70.73%	69.85%
Ministral-3-14B-Instruct	207,674	94.49%	64.06%	60.52%
By Seed Dataset Source				
UltraChat	293,479	98.02%	71.46%	70.07%
Nemotron	259,175	96.23%	64.56%	62.13%

Third, the seed data matters. Open-ended UltraChat conversations leave more room for style (71.46% style, 98.02% intent) than Nemotron's denser, often technical instructions (64.56% style, 96.23% intent), where restyling more easily collides with strict formatting or technical wording.

Persona selection and rewrite prompt. We ablate two choices of the user-turn rewrite on 400 language-matched (free-mode) conversations per arm, with the same rewriter (`Gemma-4-31B`), judge and sanity gate, in a 2×2 design. *Persona selection*: a random persona that shares the prompt's language, or the conservative matching of Section 4 (location, then top-1 embedding, then language pool). *Rewrite prompt*: the rewrite template alone as a single user message, or the same template

Attribution: `nemotron:116027` **— Ukraine (Dnipro), 46, senior payroll specialist**

The seed turn carries a hard output constraint: the answer must begin with the word *range*.

Original seed prompt (Nemotron):	**Persona-augmented prompt:**
"Hey, I'm tweaking my homemade lasagna sauce and need to know the ideal water temperature range for the broth base; could you walk me through the steps to determine it, and please make sure your response starts with the word range and avoids having any two adjacent words begin with consecutive letters of the alphabet?"	*"привіт [wave] я* тут підправляю свій домашній соус для лазаньї і треба знати ідеальний *water temperature range для* бульйону*, можете* розписати кроки як це визначити*? тільки* зробить так*, щоб* відповідь починалася зі слова *range і* щоб жодні два сусідніх слова не починалися на послідовні літери алфавіту*"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Greeting	'привіт' + emoji / 'так, що там?'	привіт: the greeting and its trailing emoji are both specified by the profile
Abbreviations	'крч', 'скоро', 'зробить'	зробить: one of three forms the profile lists, reproduced verbatim
Main language	Ukrainian	лазаньї: the turn is re-voiced in Ukrainian
Seed constraint	*(seed prompt, not the profile)*	`range`: the constrained token stays in English, since translating it would break the instruction the turn is asking about

Figure 15: Style and constraint preservation pulling in opposite directions. The turn is re-voiced into Ukrainian, yet the literal token the seed instruction constrains is left in English so the task remains verifiable.

wrapped in an additional system prompt that frames the rewriter as a generic text-transformation component. Table 7 shows that persona selection dominates: conservative matching raises style adoption by 22 to 23 points and also preserves intent slightly better, because a topically and geographically plausible persona gives the rewriter something concrete to express. The extra framing costs 4 to 5 points of style in both selection settings, so the pipeline uses the template alone.

Table 7: Persona selection and rewrite prompt ablation (400 conversations per arm, about 960 judged user turns each; 95% intervals at most ± 3.2 points).

Persona selection	Rewrite prompt	Intent (%)	Style (%)	Sanity pass (%)
Random, same language	template + framing	97.2	45.6	96.2
Random, same language	template only	96.2	50.4	94.2
Conservative matching	template + framing	**99.2**	68.6	95.0
Conservative matching	template only	98.7	**72.5**	95.5

Cross-lingual rewriting strategy. We compare two ways of producing cross-lingual user turns on the same 300 conversations, personas and rewriter (`Gemma-4-31B`): *single pass*, where the source turn is re-voiced directly into the persona's language, and *translate, then restyle*, where each turn is first translated literally into that language and the translation is then restyled with the same-language rewrite. Both variants are judged against the same literal translation of the turn, because the intent template fails any language switch by design, and the target language is checked with fastText language identification. Table 8 shows that the two-step variant keeps the request more often, reaches the target language more often and passes the sanity gate slightly more often, while the single pass adopts the persona's style slightly more often. We use the two-step variant, since a rewrite that drifts from the request or out of the target language defeats the purpose of the cross-lingual mode.

Attribution: `nemotron:72917` — **UK (Wales), 39, senior lecturer in Welsh and Media Studies**

A request for a short LEGO story with three hard format constraints.

Original seed prompt (Nemotron):	**Persona-augmented prompt:**
*"Hey! Can you write a short adventure about my favorite LEGO squirrel who helps a lost duck find its way home in the woods? Please keep the story to exactly three paragraphs and separate them with the markdown divider ***. Also, make sure you don't use the word "depression" anywhere in the story. Thanks!"*	*"ta, so can you slap together a little three-paragraph tale about my favourite LEGO squirrel - *exactly* three, na? - helping a lost duck find its way back home through the woods, but keep it tight like a good Welsh-language haiku, hmm? no "depression" anywhere, just the squirrel, the duck, and maybe a *cyfreithiol* amount of forest wisdom. oh and 1. stick to the *** (triple asterisk) dividers like it's a rule from the *BBC Cymru Wales* script guidelines - no wiggle room 2. no extra fluff, just the bones of the story, like when i'm editing student essays for their A-levels bfn"*

Which profile field accounts for which span

Field	**Value in the profile**	**Trace in the rewrite**
Occupation	Senior lecturer in Welsh and Media Studies, University of Wales Trinity Saint David	`Welsh-language haiku`: the discipline supplies the simile, and a second aside compares the task to marking student essays
Reference pool	The *BBC Cymru Wales* archive of Welsh-language programming (she occasionally lingers on the 2000s era, even though it's on YouTube)	`BBC Cymru Wales`: an institution named in the profile, cited as the authority for a format rule
Quirk	Repeatedly writes 'cyfreithiol' as a single word (in Welsh, it's two: 'cyfreithiol' means 'legal') when frustrated by legalese in academic English	`cyfreithiol`: the documented Welsh insertion, in the exact word the profile names
Quirk	Under stress, begins sentences with Welsh particles like 'na' (e.g., 'na... let's talk about this properly') or 'peidio' (e.g., 'peidio, don't just […]	`na?`: a Welsh particle the profile attaches to sentence openings
Signoff	'Cwra' for 'bye' in Welsh-to-English messages, otherwise 'ta' (Welsh for 'thanks') in replies	`ta,`: the opening *ta* the profile documents
Abbreviations	''hwnnw' for 'that', 'wna' for 'wants to', ''n' for 'and' (Welsh mixed), 'bfn' for 'bye for now' (rarely)	`bfn`: listed by the profile, and it closes the message
Seed constraints	*(seed prompt, not the profile)*	`***`: the three-paragraph count, the divider and the banned word all survive

Figure 16: Occupation and style together on a constrained task. Six profile fields are traceable into the rewrite while the exact-format requirements of the seed prompt are left intact.

Table 8: Cross-lingual rewriting strategy ablation (300 conversations per variant, about 730 judged user turns each, same personas, rewriter and judge; 95% intervals at most ±3.6 points).

Variant	Intent (%)	Style (%)	Target language (%)	Sanity pass (%)
Single pass	96.0	**61.1**	84.6	94.3
Translate, then restyle	**99.0**	58.2	**88.6**	**95.7**

Corpus-scale judge-free fidelity metrics. These metrics use no LLM judge, so they are free of judge bias. They fall into two groups: how much a rewrite changes the text and what it keeps intact, and how closely it follows the style values in the persona's profile.

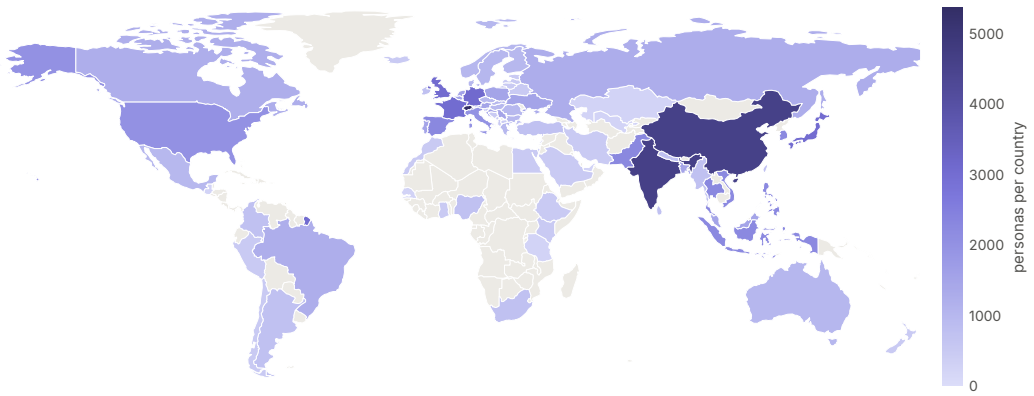


Figure 17: Geographic coverage of the persona bank: 99,696 profiles across 81 countries. Colour encodes the number of profiles sampled per country; grey marks countries outside the curated sampling grid. The largest country shares are Switzerland (5,371 profiles), India (4,623) and China (4,614), with substantial mass throughout South, Southeast and East Asia.

TRANSFORMATION DEPTH AND CONSTRAINT RETENTION.

- **Same-language token Jaccard distance:** Defined as $1 - \frac{|T_{\text{orig}} \cap T_{\text{rw}}|}{|T_{\text{orig}} \cup T_{\text{rw}}|}$, where token set T is extracted multilingually by treating CJK/Hangul ideographs as atomic units and non-CJK scripts as whitespace- and punctuation-delimited words.
- **Same-language normalized edit distance:** Character-level Levenshtein distance normalized by sequence length: $\frac{\text{Lev}(s_{\text{orig}}, s_{\text{rw}})}{\max(|s_{\text{orig}}|, |s_{\text{rw}}|)}$, bounded in $[0, 1]$.
- **Core-vocabulary token retention:** Filters non-informative functional tokens (retaining words of length > 2 or non-ASCII tokens) to measure the fraction of core lexical content preserved: $\frac{|C_{\text{orig}} \cap C_{\text{rw}}|}{|C_{\text{orig}}|}$.
- **Multilingual semantic similarity:** Cosine similarity between 768-dimensional sentence embeddings computed via Language-Agnostic BERT Sentence Embeddings (LABSE) across all parallel $(s_{\text{orig}}, s_{\text{rw}})$ turn pairs.
- **Numbers & quantities intact:** Regular expression extraction targeting integers, decimals, currency values, percentages, and version strings (e.g., `v2.1`), verifying that numbers in s_{orig} form a subset of numbers in s_{rw}.
- **Code & technical identifiers intact:** Lexical parsing targeting markdown code blocks, backtick-enclosed identifiers, camelCase/snake_case tokens, URIs, Unix file paths, and CLI flags.
- **Quoted string literals intact:** Verifies retention of verbatim text enclosed in single, double, or regional quotation marks (e.g., Japanese corner brackets and French guillemets).
- **All entities preserved per turn:** A strict composite metric requiring that every entity category present in a given turn (numbers, code, quotes) is simultaneously 100% preserved.

STYLE ADHERENCE. Each metric checks the rewrite against the matching attribute in the persona's profile:

- **Emoji & emoticon register:** Text is matched against Unicode 15.0 emoji codepoints, ASCII emoticons (e.g., `:-)`, `<3`), and kaomoji (e.g., `(^_^)`, `orz`). A turn passes if it satisfies its target persona tier: *none* (0 emojis/emoticons), *rare/functional* (≤ 2 functional emojis), *frequent/decorative* (≥ 1 decorative emoji), or *kaomoji* (≥ 1 kaomoji pattern).
- **Capitalization register:** For cased scripts, sentence-start capitalization and casing proportions are verified against target profiles: *all-lowercase* ($\geq 70\%$ lowercase or 0 capitalized sentence starts), *standard* ($\geq 70\%$ capitalized sentence starts), *first-word-only* (only the initial turn character capitalized, without terminal periods), and *all-caps* (presence of new capitalized tokens for emphasis).

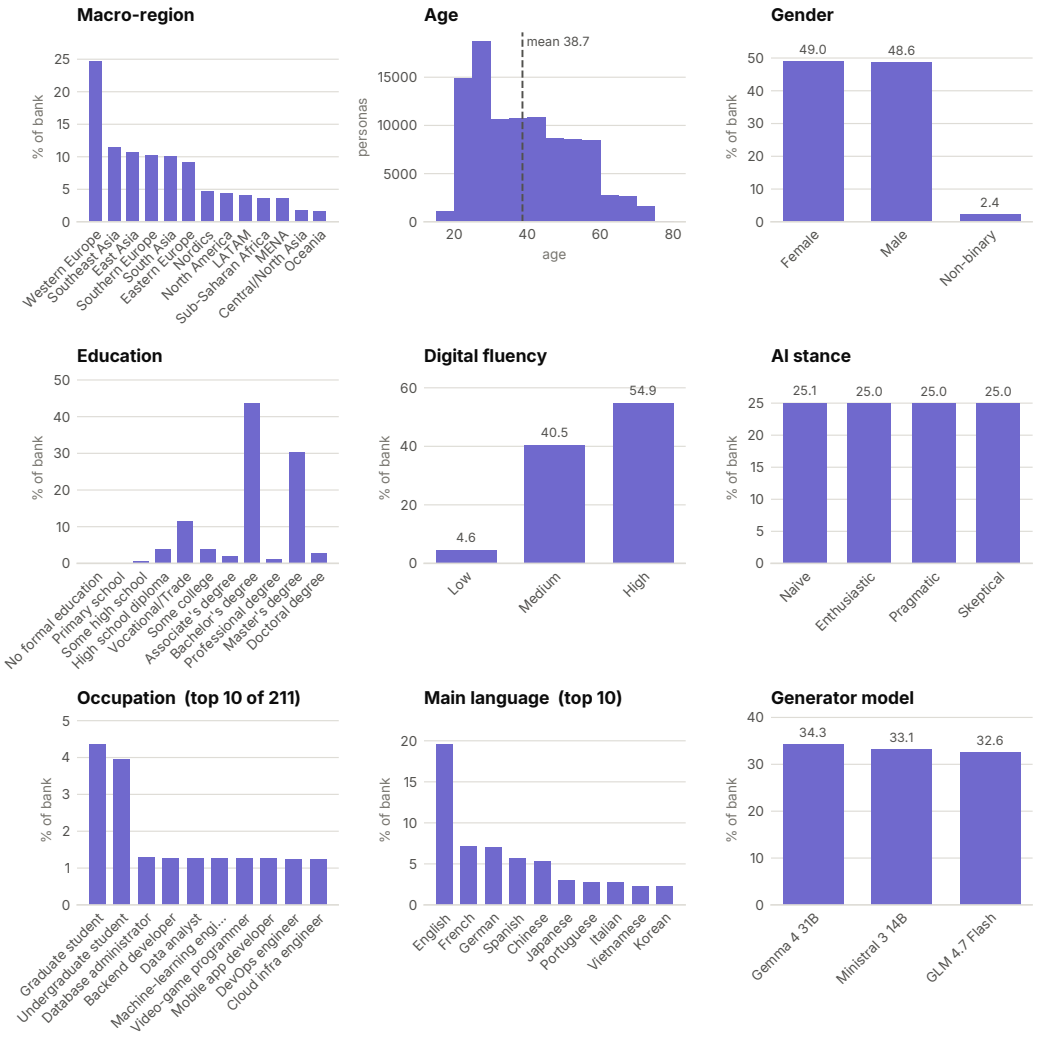


Figure 18: Background-layer distributions. Region, gender, AI stance and generator model are close to balanced by construction, while age, education and digital fluency inherit the skews discussed in Section 7: the bank is younger, more highly educated and more digitally fluent than a representative population, and occupations concentrate on students and technology roles. Language and occupation panels show the ten most frequent values.

- **Punctuation register:** Evaluates line-ending periods, ellipses (. . .), linebreaks, and expressive punctuation (!!, ??) against persona categories: *standard* ($\geq$ 70% punctuated line endings), *minimal / drops period* (omission of terminal periods), *ellipsis-heavy*, and *expressive*.
- **Length ratio short/medium/long:** Median length ratio $\frac{|s_{\text{rw}}|}{|s_{\text{orig}}|}$ conditioned on the persona's specified verbosity tendency (`Short`, `Medium`, `Long`), corroborated by a positive Spearman rank correlation.

Table 9 reports the full judge-free fidelity metrics.

Multi-judge agreement. All style and intent scores in the paper come from LLM judges, so we measure how much they depend on which judge is used. We re-score a stratified set of 100 persona rewrites (spanning the sixteen single-attribute conditions of Appendix I) with four independent judges (`gemma-4-31B`, `GLM-4.7-Flash`, `Ministral-3-14B`, `MiniMax-M2.5`), each issuing a bi-

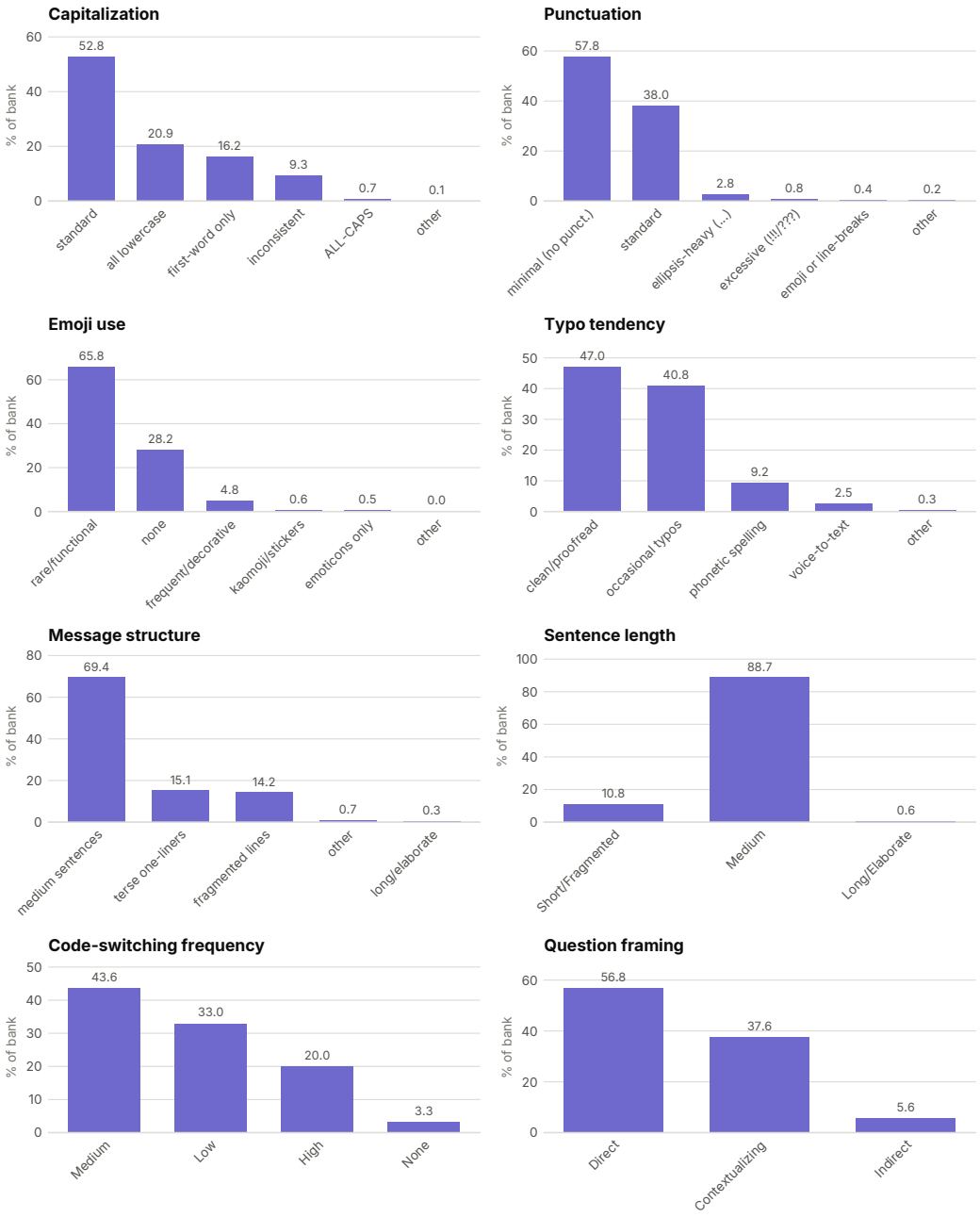


Figure 19: Style-layer distributions, i.e. how profiles say they type. The schema offers fixed options but also admits a short free-text value, so raw entries are folded onto the canonical options by keyword matching; the residual "other (free-text)" bucket stays below 1% on every axis. Non-standard registers are well represented: more than half of profiles drop end punctuation and a fifth type in all lowercase. Frequent decorative emoji use is deliberately limited to about 5% of profiles, since it otherwise dominates the rewritten turns.

nary style-adoption and intent-preservation verdict per rewrite. Table 10 reports each judge's mean verdict rate and Table 11 the pairwise item-level agreement.

The judges differ most in how lenient they are: mean style pass rates range from 0.78 to 0.93. Raw agreement on individual items is high (0.82 to 0.90 on style), but part of that is expected by chance when most items pass. Table 12 corrects for this. Chance-corrected agreement is moderate to strong

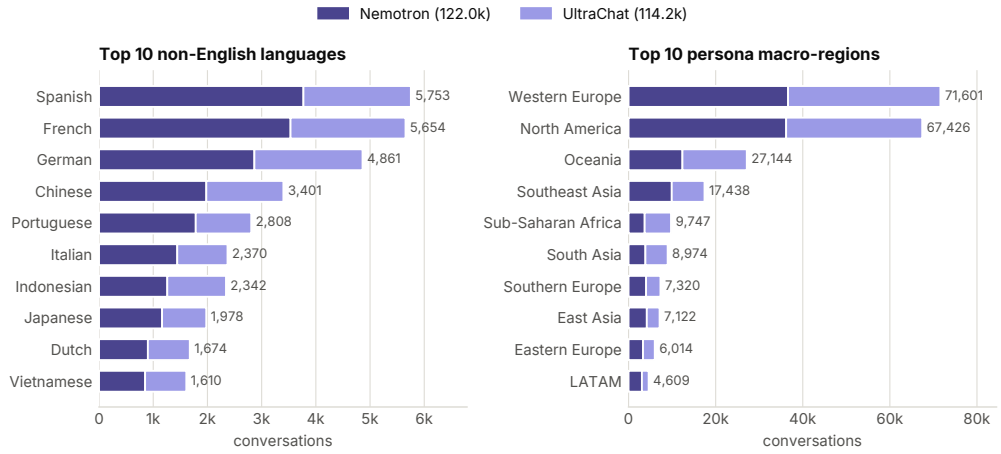


Figure 20: Distribution of the released chat dataset: (a) top 10 non-English languages (detected language of the first user turn), and (b) top 10 persona macro-regions.

Table 9: Full-corpus judge-free fidelity metrics.

	Metric	Value
Transformation depth	Same-language token Jaccard distance	0.425
	Same-language normalized edit distance	0.358
	Core-vocabulary token retention	71.0%
	Multilingual semantic similarity (LaBSE)	0.854
Constraint retention	Numbers & quantities intact	96.1%
	Code & technical identifiers intact	88.6%
	Quoted string literals intact	80.2%
	All entities preserved (per turn)	86.9%
Style adherence	Emoji & emoticon register	98.4%
	Capitalization register	60.9%
	Punctuation register	68.1%
	Length ratio short/medium/long (per-item ρ)	0.86×/1.00×/1.26× (0.18)

for pairs with comparable pass rates, and every pair involving the production judge (`MiniMax-M2.5`) reaches $\kappa \geq 0.32$ on style and $\kappa \geq 0.22$ on intent, with the strongest agreement between `MiniMax` and `gemma` on intent ($\kappa = 0.86$). The near-zero style κ between `GLM` and `Ministral` arises because both pass more than 92% of items, which leaves little variance to agree on; AC1, which corrects for this prevalence effect, is 0.85 for the same pair. The effective sample is 99 items for the style verdicts and 100 for intent. In short, a result obtained with one judge carries over to the others in direction, while absolute pass rates depend on the judge; this is why we report the external re-judging below rather than relying on a single judge.

External re-judging of the structured versus flat A/B. All four judges above are also rewriters or the production judge of the pipeline, so they are not fully independent of it. We therefore re-score all 518 stored pairs of the A/B in Table 1, with the flat and structured rewrites in a per-item random order, using `MiniMax-M2.5` with a larger output budget and three judges from outside the pipeline, served through a separate inference stack. Table 13 and Figure 21 report the results. The larger budget recovers 514 of 518 verdicts for `MiniMax`, which shows that the attrition in the original pass (433 of 518) came from truncated reasoning rather than from the items, and reproduces its direction (62.4% against 64% of decisive verdicts for structured). All three external judges prefer the structured representation; the effect is clear for both Qwen models and small for `DeepSeek-V4-Flash`, whose interval touches parity. Intent preservation is identical for the two representations under every judge.

Table 10: Per-judge mean verdict rate on the same 100 persona rewrites (style adoption and intent preservation). Judges differ in severity, most visibly on style, where MiniMax is strictest and GLM most lenient.

Judge	Mean style	Mean intent
gemma-4-31B	0.840	0.680
GLM-4.7-Flash	0.929	0.830
Ministral-3-14B	0.920	0.930
MiniMax-M2.5	0.778	0.680

Table 11: Pairwise item-level judge agreement on the 100 rewrites. Upper triangle: agreement on the style-adoption verdict; lower triangle: agreement on the intent-preservation verdict. Style verdicts agree at 0.82–0.90 across every pair; intent agreement is looser (0.73–0.94). Because pass rates are high and differ between judges, raw agreement overstates consistency; Table 12 gives chance-corrected values.

	gemma	GLM	Ministral	MiniMax
gemma	–	0.899	0.880	0.879
GLM	0.790	–	0.869	0.827
Ministral	0.730	0.880	–	0.818
MiniMax	0.940	0.830	0.730	–

G Assistant-Side Model Personas

This appendix documents the assistant-side model personas of Section 6: the style set, how styles are matched and used, the data-level pilot that motivated their use, and a downstream fine-tuning ablation. All judged numbers come from a single LLM judge (`MiniMax-M2.5`) and are proxies.

Style set. Each of the 12 styles has a free-text description of how answers are presented, eight presentation axes (register, structure, density, warmth, explanation, jargon, certainty, emoji), two or three optional signature touches, and content safeguards (e.g. never drop a requested detail for brevity, never agree with an incorrect premise). Two rules hold for every style: keep every fact, step, number and requested detail the answer needs, and never invent facts, sources or statistics to fit the style. Table 14 lists the styles by their one-line summary together with their share of the release. Style names are used only for bookkeeping and never appear in a prompt. In a 1,000-conversation pilot, removing them from the prompt left style adoption (68.8% without names, 66.4% with) and judged fit (47.6 to 50.8 pairwise) unchanged within noise, and it removes any chance of a label leaking into the replies: 52 released conversations in which a reply named its own style were regenerated without names.

Matching and generation. The matcher sees a reduced user card with the persona's AI stance, digital fluency, sentence-length tendency, question framing, code-switching frequency and structured typing keys, together with the first user message; occupation, gender, age, location, languages and the free-text chatting style (which often names a language or region) are withheld, so the choice cannot key on demographics. Candidate styles are shown in a per-conversation random order, and the top-ranked style is used for the whole conversation. During assistant regeneration, the style section is appended to the paper's persona system prompt on every turn, with an explicit priority order: explicit user instructions about format, length, tone or language first, then the content rules, correctness and safety, and only then the style. A signature touch is offered with $p = 0.25$ per reply (25.0% of released assistant turns), and the continuity adaptation step keeps the style. As with the user persona, neither the style nor its matching is stored in the training example. The match distribution is skewed (Table 14): the two most frequent styles cover 43% of conversations and the least frequent 1.5%.

Design pilots. Before settling on the release setup, we tested three design questions on 604 conversations (1,489 assistant turns) with an earlier version of the style set without signature touches (Table 15). Content agreement is measured against the reply generated with the user persona only; instruction compliance is measured on the 94 later turns with an explicit format or length instruc-

Table 12: Chance-corrected pairwise judge agreement on the same rewrites: Cohen's κ and Gwet's AC1, which is robust to the high pass rates of lenient judges. Computed from the observed agreement (Table 11) and each judge's marginal pass rate (Table 10).

	Style		Intent	
Judge pair	κ	AC1	κ	AC1
gemma – GLM	0.52	0.87	0.45	0.67
gemma – Ministral	0.44	0.85	0.22	0.61
gemma – MiniMax	0.61	0.82	0.86	0.89
GLM – Ministral	0.06	0.85	0.44	0.85
GLM – MiniMax	0.34	0.77	0.55	0.73
Ministral – MiniMax	0.32	0.76	0.22	0.61

Table 13: Structured versus flat style representation, re-scored on all 518 stored pairs. Preference shares are over parsed verdicts; the decisive share excludes ties, with a 95% Wilson interval. `Qwen3.5-397B` returned no parseable verdict for 74 items.

Judge	Parsed	Structured	Flat	Tie	Structured, decisive	Intent (flat / struct.)
`MiniMax-M2.5` (original pass)	433	60.0	33.0	6.0	64	92 / 92
`MiniMax-M2.5` (larger budget)	514	58.4	35.2	6.4	62.4 (58.0–66.6)	92.3 92.5
`DeepSeek-V4-Flash`	518	51.9	43.8	4.2	54.2 (49.8–58.6)	92.4 92.6
`Qwen3.5-397B-A17B`	444	66.9	27.7	5.4	70.7 (66.2–74.9)	92.7 93.5
`Qwen3.6-35B-A3B`	518	61.0	34.0	5.0	64.2 (59.9–68.3)	92.4 92.6

tion. First, *matching carries signal*: in a pairwise judgment of which reply suits the user better, the matched style beats a random style 60.0 to 39.9. Second, *generating with the style beats restyling afterwards on adoption*, while restyling an existing reply keeps its content almost perfectly; we generate with the style because the release regenerates the assistant side anyway, and keep restyling as an option for data whose answers must not change. Third, instruction compliance on later turns is unaffected in every arm within the precision of this sample. As in the main pilot, the matched style ties the user-persona-only reply on judged fit (49.7 to 50.1).

Data-level pilot. On 1,000 conversations with fixed user turns, we regenerate the assistant side four ways with the same responder per conversation and compare them (Table 16). Style adoption asks whether a reply is recognisably presented in the conversation's matched style; for the default and background-only arms this measures how often the reply happens to fit that style anyway. Content agreement compares each reply with the default reply to the same turn; the noise floor, two independent default samples, is 59.6%. Instruction compliance is measured on later turns whose user message contains an explicit format or length instruction.

The model persona makes the intended style clearly recognisable without reducing instruction compliance, and conditioning on the user persona shortens replies to about half the default length. In a pairwise judgment of which reply suits the user better, shown the reduced user card, the combination of user persona and model persona beats the default 61 to 39 and the model persona alone 58 to 41, but ties the user persona alone (49 to 51). The model persona therefore adds a controllable assistant style rather than a better fit to the user.

Downstream fine-tuning ablation. We regenerate the assistant side of the corpus three ways with identical user turns: (A) no assistant system prompt, (B) the paper's user-persona system prompt, and (C) the user persona plus a matched model persona, as in the release. We fine-tune `Llama-3.1-8B` (full fine-tuning, one epoch, learning rate 10^{-5}, global batch 128) and `Apertus-8B` (same budget, packed multi-turn sequences with loss on every assistant turn) on the same 100k conversations per arm. For evaluation, 300 held-out requests are each rewritten by 12 user personas that span digital fluency, AI stance, capitalization, message length and typos, so the request is fixed and only the user's voice changes; replies are decoded greedily. We measure within-request correlations between a user-message feature and the matching reply feature, style diversity as the mean pairwise distance of the 12 replies to one request in a content-independent style embedding (Wegmann et al., 2022), content stability as LaBSE similarity to the reply given to the unstyled request, and a pairwise judge of which

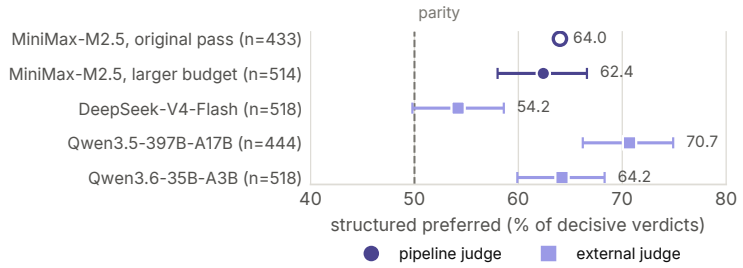


Figure 21: Structured versus flat style representation under the pipeline judge and three external judges: share of decisive verdicts preferring the structured rewrite, with 95% Wilson intervals (Table 13). Every judge lies above parity; `DeepSeek-V4-Flash` only marginally.

Table 14: The 12 assistant response styles, ordered by their share of the released corpus ($n = 236{,}191$ conversations).

Summary	Register	Structure	Convs.	Share
Talks like a knowledgeable friend: relaxed, contractions, light humour.	casual	conversational prose	52,750	22.3%
Answers first, in compact expert prose, with no preamble or sign-off.	neutral-professional	prose	49,118	20.8%
Explains through analogies and vivid concrete examples in flowing prose.	conversational	flowing prose	25,106	10.6%
Maximally compact layout: bullets, fragments allowed, zero filler.	neutral	tight bullets or one short paragraph	21,850	9.3%
Separates facts from judgement and makes assumptions explicit.	neutral-professional	labelled parts (established, assumptions, how to check)	20,779	8.8%
Formal written register with full sentences and a concluding summary.	formal	paragraphs with topic sentences	16,883	7.1%
Organises every answer into headed sections, numbered steps and code first.	neutral	short headers, numbered steps, tables where comparisons appear	16,675	7.1%
Short sentences, everyday words and one concrete example.	neutral	short paragraphs, occasional simple bullets	8,698	3.7%
Walks through things step by step in plain words, explaining each term once.	friendly-neutral	short paragraphs or numbered steps	8,048	3.4%
Energetic, tech-savvy and informal, with snappy bullets.	casual	snappy bullets and short code snippets	7,683	3.3%
Supportive, action-oriented and second person, without flattery.	friendly	short paragraphs and action lists	5,056	2.1%
Gentle, steady tone suited to stressful or personal topics.	gentle	short prose paragraphs	3,545	1.5%

reply is presented better for the user, with a length-controlled (LC) win rate estimated at equal reply length by logistic regression on the log length ratio.

Three observations follow (Tables 17, 18 and 19, Figure 22). Formality is the mean score of a formality classifier (Babakov et al., 2023). First, every model fine-tuned on persona-styled conversations tracks the user's formality and emoji use far more than `Llama-3.1-8B-Instruct`, in both model families, and is clearly preferred over it. Second, without assistant conditioning (A) the models inherit the teacher's defaults: replies two to two and a half times longer, headers in about three quarters of them, and the lowest style diversity. Conditioning the assistant (B, C) yields shorter, less formatted and more varied replies that the judge prefers, but the preference disappears or reverses once length is controlled; since B and C replies are about half as long, the equal-length estimate is an extrapolation. Third, the model persona does not beat the user-persona prompt on judged fit (a tie for Llama, B preferred for Apertus); its benefit is the controllable style variation of Table 16, and downstream

Table 15: Model persona design pilot (604 conversations; content agreement against the user-persona-only reply; instruction compliance on 94 later turns with explicit instructions). 95% intervals are at most ±2.5 points for per-turn rates and ±10 points for instruction compliance.

Assistant side	Style adopt.	Same content	Instr. compliance	Sanity pass
Regenerated, user persona only	35.1	44.6	84.0	88.2
Matched style, generated	**75.7**	37.6	87.2	87.7
Random style, generated	53.3	35.8	**89.2**	87.7
Matched style, restyled afterwards	67.1	**96.9**	88.2	**92.3**

Table 16: Data-level model persona pilot (1,000 conversations, about 2,500 assistant turns, `MiniMax-M2.5` judge; 95% intervals at most ±1.9 points for per-turn rates).

Assistant conditioning	Style adopt.	Same content	Instr. compliance	Median words
None (default)	19.4	(ref.)	86.1	521
User persona (paper)	25.4	36.9	87.6	266
Model persona only	71.1	40.0	86.4	313
User persona + model persona	66.8	35.9	87.3	271

Llama C adapts length and emoji most and is the most diverse, while Apertus C is less diverse than Apertus B. These results use one seed per arm, a single judge and no capability benchmarks, and the fine-tuning data share the user side, so they compare assistant conditioning only.

Conversations without a user persona. Model personas do not depend on the user side, so they can also restyle the assistant side of an existing dataset. In this mode, the matcher sees only the user's own messages (the first three), and the assistant turns are regenerated with the style section as the only system prompt, with the same continuity loop and sanity gate. On 1,000 source conversations (UltraChat and Nemotron, 2,570 assistant turns), matching succeeds for every conversation, and Table 20 compares the variants. Regeneration with the style raises style adoption from 25.7% to 79.5% at a cost of about five points of sanity pass; restyling the source replies afterwards keeps their content but reaches lower adoption and falls back to the original reply on 35% of turns because a protected span was lost.

A limitation of this mode appears when the source conversation carries its own system prompt. On 200 single-turn conversations with a checkable system constraint (at most three sentences, a numbered list, under 80 words, or no Markdown), compliance is 86% without a style, but 58% with the style placed below the system prompt and 66% with the system prompt placed last as an explicit override. The style competes with explicit format and length constraints even when told to yield, so conversations with a system prompt are left unaugmented.

H Comparison to Other Persona Representations

Section 5 compares a structured style representation against a free-text one within our own schema. This appendix asks the same question against the formats the field actually uses: PersonaHub (Ge et al., 2025) and SynthPAI (Yukhymenko et al., 2024).

Design. We hold the persona *content* fixed and vary only its serialization. From the released corpus we sample 800 conversations, stratified by persona set and seed source, together with the persona the pipeline matched to each. Sampling is independent of the released sanity verdict, and every condition is regenerated from scratch, so no condition inherits an advantage from the corpus filtering. Each profile is then rendered four ways and used to rewrite the same seed prompt with the same rewriter, decoding parameters, judge and prompt scaffold. Only the persona block differs.

The PersonaHub rendering is a single third-person sentence. This is not a simplification on our part: across all 200,000 released PersonaHub personas, only 1.4% mention any writing cue at all, so a projection that described typing behaviour would not be testing PersonaHub. The SynthPAI rendering keeps the feature types that schema has, an attribute block plus a free-text writing-style description

Table 17: Adaptation to the user's style after fine-tuning. Correlations are within-request (user-message feature to reply feature); diversity and stability as described in the text.

Model	formality	lowercase→formality	length	emoji	style diversity	content stability
`Llama-3.1-8B-Instruct`	0.18	−0.16	−0.06	0.12	0.254	0.866
Llama A (no conditioning)	0.51	−0.45	−0.03	0.40	0.187	0.827
Llama B (user persona)	0.37	−0.33	0.11	0.40	0.293	0.775
Llama C (+ model persona)	0.43	−0.41	0.26	0.49	0.313	0.777
Apertus A (no conditioning)	0.49	−0.43	−0.01	0.28	0.185	0.824
Apertus B (user persona)	0.45	−0.38	0.04	0.29	0.380	0.819
Apertus C (+ model persona)	0.45	−0.37	0.05	0.38	0.275	0.818

Table 18: Judge-free features of the fine-tuned models' replies to the style probes.

Model	Median words	Headers (%)	Bullets (%)	Emoji (%)	Formality	Contractions / 100 words
`Llama-3.1-8B-Instruct`	426	6.5	67.9	0.1	0.71	1.75
Llama A	497	71.5	71.9	17.4	0.64	0.92
Llama B	195	12.9	43.0	1.4	0.77	0.47
Llama C	210	8.2	45.6	2.0	0.73	0.95
Apertus A	666	75.1	70.9	4.4	0.68	0.71
Apertus B	293	10.2	47.5	0.7	0.74	0.52
Apertus C	299	7.4	39.7	1.1	0.73	0.96

with a verbatim example, and drops the per-axis style keys, which it has no slot for. Projections were produced by a third model that is neither rewriter nor judge, and were accepted only after matching the real PersonaHub length distribution and style-mention rate. Table 21 reports what survives each rendering.

Result. Table 22 and Figure 24 show a dose-response in how explicitly a representation encodes writing behaviour. Style adoption rises monotonically across the four conditions, and the judge prefers the PersonaFactory rewrite in every pairwise comparison. The contested comparison is against the SynthPAI form, which also carries style information and differs from ours mainly in whether that information is structured; there the preference is 4.83 to 1.

Two observations do not follow from the setup. First, the PersonaHub form is not inert. It moves the rewrite away from the no-persona baseline about as far as the other conditions do, yet lifts style adoption only from 3.1% to 5.5%: it changes the text without changing it toward the persona. Second, the representation propagates past the user turn. Measured against the no-persona reply for the same item, the assistant's response moves further as the representation grows richer, while assistant sanity stays flat across conditions, so this is a change of register rather than of quality.

Cost. The gains are not free, and the cost is concentrated rather than uniform. Intent preservation is essentially unaffected in English but falls to 89.2% on the cross-lingual half, and entity retention declines by four points. Transformation-depth metrics also fall, but those measure how much the surface changed and are the intended effect rather than a loss.

Robustness. Table 23 reports three checks. The ranking is unchanged with Ministral-3-14B as rewriter, though the margin against the SynthPAI form narrows from 4.83 to 2.64, consistent with Appendix I finding that the rewriter dominates attribute effects. Re-judging a 200-item subset with four judges leaves the ordering intact while absolute rates span fifteen points, matching the severity spread in Appendix F; item-level agreement with the primary judge is 0.81 to 0.93. Finally, judge responses that could not be parsed are not uniformly distributed across conditions, so we report a worst-case bound in which every unresolved comparison is assigned to the competing condition. The SynthPAI comparison holds at 76.8% under that assumption.

Format-neutral re-judging and real personas. In the comparison above, the style and pairwise judges see the persona rendered in the PersonaFactory format, which the PersonaFactory rewriter also saw, so a rewrite could in principle score by echoing phrasing from its own prompt. We therefore re-judge the same 3,200 rewrites against a format-neutral reference that lists only the persona's writing-behaviour attribute values (structured typing keys, sentence length, question framing, code-

Table 19: Pairwise judge after fine-tuning: raw win rate with 95% confidence interval and length-controlled (LC) win rate of the first model; about 3,500 decisive comparisons per pair.

Comparison	Raw win (%)	LC win (%)
Llama B vs. A	57.0 (55.4–58.7)	48.8
Llama C vs. A	54.9 (53.3–56.6)	42.4
Llama C vs. B	48.1 (46.4–49.8)	48.6
Llama C vs. `Llama-3.1-8B-Instruct`	79.4 (78.1–80.7)	68.6
Apertus B vs. A	66.3 (64.7–67.8)	52.1
Apertus C vs. A	57.1 (55.5–58.7)	44.8
Apertus C vs. B	42.0 (40.3–43.6)	42.6

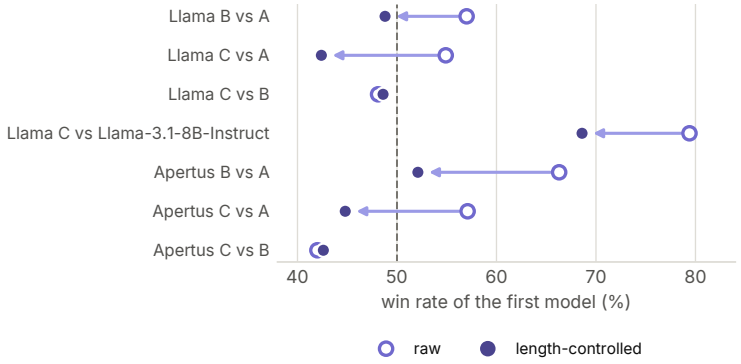


Figure 22: Pairwise win rates after fine-tuning, raw (open) and length-controlled (filled). Controlling for reply length removes or reverses most preferences for the conditioned arms over the unconditioned one (A), while the comparison against `Llama-3.1-8B-Instruct` stays clearly in favour of the fine-tuned model.

switching frequency) with all example phrasing removed, using the pipeline judge and two external judges (`Qwen3.6-35B-A3B` and `Muse-Glimmer-30B`). We also rewrite every item with a real PersonaHub persona and a real SynthPAI profile in their native form, with the same rewriter and template, and compare each with the PersonaFactory rewrite in a reference-free pairwise judgment of which rewrite gives the message a more distinctive and realistic individual voice while keeping the request intact. Figure 23 shows that PersonaFactory keeps the highest style adoption under every judge and is preferred over both projected formats and over real PersonaHub personas, with every interval above parity. Real SynthPAI profiles, whose hand-written style descriptions cover 300 English-speaking Reddit users, are on par with PersonaFactory; they are preferred on cross-lingual prompts, where our persona speaks a different language than the prompt, while PersonaFactory scales the same style fidelity to a bank of nearly 100,000 multilingual profiles with per-axis control.

Naturalness. Stronger stylization has a cost. In a reference-free judgment of whether a message reads as something a real person would type or as a caricature, the pipeline judge passes 99% of rewrites in every condition, but the two external judges rate PersonaFactory rewrites 7 to 12 points less natural than unconditioned rewrites. We attribute this to the richer behaviour specification: a profile that asks for lowercase typing, dropped punctuation, abbreviations, emoji and a sign-off at the same time invites the rewriter to realise several markers in one short message, which a judge can read as over-performed, whereas formats that specify little change little. The per-turn quirk probability ($p = 0.25$) already limits this effect, and lowering it or capping the number of markers realised per turn is a direct lever for trading style adoption against naturalness.

Scope. This compares representations, not systems. Every condition renders our own profiles, so the result says how a persona is encoded affects augmentation, not that one persona bank is better than another. A consequence worth noting is that no PersonaHub or SynthPAI content enters the released data.

Table 20: Model personas on conversations without a user persona (1,000 source conversations; content agreement on the first assistant turn against the source reply; 95% intervals at most ±1.9 points for style adoption and ±3.1 points for the other rates).

Assistant side	Sanity pass	Style adopt.	Same content	Median words
Source replies	n/a	17.4	(ref.)	344
Regenerated, no style	**82.2**	25.7	42.4	583
Regenerated with style	77.4	**79.5**	25.4	371
Source replies restyled	n/a	57.5	**93.0**	417

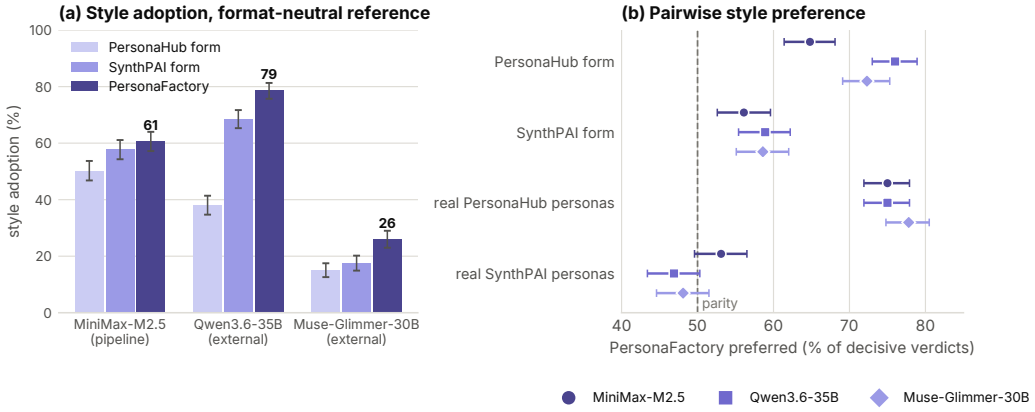


Figure 23: Persona representations re-judged against a format-neutral reference (the persona's style attribute values only, no example phrasing) by the pipeline judge and two external judges. (a) Style adoption per format. (b) Share of decisive pairwise verdicts preferring the PersonaFactory rewrite, with 95% Wilson intervals; the two real-persona comparisons are reference-free and ask which rewrite has the more distinctive and realistic individual voice.

I Per-attribute Augmentation Analysis

Aggregate fidelity tells us *that* persona conditioning works, but not *which* part of a profile is doing the work or what it costs. We therefore run a controlled, one-attribute-at-a-time ablation that decomposes the augmentation into three separately measured aspects. To our knowledge, prior user-simulation and persona-conditioning work reports fidelity only in aggregate; a per-attribute decomposition of a persona-augmentation pipeline is, to our knowledge, new.

Setup. From a stratified pool of up to 194 real source messages, we build *single-attribute* personas: a no-persona baseline plus sixteen conditions that vary exactly one attribute and collapse into eight families (digital fluency, sentence length, AI stance, age, gender, education, occupation, hobbies). Each condition is rewritten by all four generators (`gemma-4-31B`, `Ministral-3-14B`, `GLM-4.7-Flash`, `MiniMax-M2.5`), holding everything else fixed, for 12,845 valid rewrites. Every rewrite is scored on three orthogonal aspects: *style adoption* (does the turn take on the attribute's register?), *intent preservation* (is it still the same request in the same language?), and rule-based *script stability*; the no-persona baseline fixes the script-stability ceiling at 0.987. Table 24 reports the eight families, averaged over the four generators; Figure 25 gives the full per-condition grid behind it.

Reading the decomposition. We present these numbers as a measurement, not an importance ranking. Across all eight families, style adoption sits in a narrow 0.80–0.82 band, intent preservation clusters near 0.62, and script stability decrements uniformly by roughly 0.06–0.07 from the no-persona baseline. We deliberately avoid ordering attributes by effect size: the between-attribute spread (at most ≈0.02 in style adoption) is smaller than the cross-generator and sampling variation, so no single attribute is separably "more important." The informative signal is the *uniformity* itself, and it supports a negative statement only: within the precision of this probe, no background attribute is

Table 21: What each persona representation encodes. Every condition renders the *same* profile; only the serialization differs. Presence is the share of profiles whose attribute value survives into the block the rewriter sees ($N = 555$ profiles).

	No persona	PersonaHub form	SynthPAI form	PersonaFactory
Words in block	0	13	83	280
Age	0%	0%	100%	100%
Location	0%	16%	100%	100%
Occupation	0%	68%	100%	100%
Language	0%	8%	26%	99%
Style axes	0%	0%	55%	100%
Quirks	0%	0%	1%	100%
Reference pool	0%	2%	1%	100%

Table 22: Augmentation quality by persona representation, $N = 800$ paired items, gemma-4-31B rewriter, MiniMax-M2.5 judge. Style adoption and intent use the production judge templates; adherence metrics are judge-free. Pairwise is the share of items where the judge prefers the PersonaFactory rewrite, with order randomized per item. 95% Wilson intervals for the judged rates are given in Table 2; paired bootstrap intervals for the judge-free differences are at most ± 0.05.

	No persona	PersonaHub form	SynthPAI form	PersonaFactory
Style, the intended effect				
Style adoption	3.1%	5.5%	43.5%	69.2%
Punctuation adherence	0.405	0.410	0.645	0.757
Emoji adherence	0.814	0.815	0.939	0.980
Capitalization adherence	0.802	0.808	0.707	0.722
Cost				
Intent preserved (all)	99.8%	99.8%	98.0%	96.1%
English	99.8%	100.0%	99.8%	99.3%
Cross-lingual	99.6%	99.2%	94.0%	89.2%
Entities preserved	0.946	0.936	0.921	0.905
Quoted literals intact	0.967	0.955	0.945	0.938
Assistant sanity	92.8%	92.4%	91.6%	92.4%
Pairwise preference against PersonaFactory				
PF preferred	94.4%	94.9%	81.3%	–
win ratio	17.90:1	20.46:1	4.83:1	–
worst case	94.0%	94.6%	76.8%	–

separably more influential than another, and conditioning on any of them costs a small, roughly constant script-stability decrement. It does not show that each attribute is individually effective, and none of the structured style keys is ablated here.

What does vary: the generator. The variation that dominates is the rewriter, not the attribute, as Figure 26 makes plain: style adoption spans barely 0.02 across the eight attribute families but 0.35 across the four rewriters. On identical conditions it ranges from ≈0.61 (GLM) to ≈0.96 (Ministral), and it trades off against intent preservation (Ministral ≈0.38 vs. gemma ≈0.77 and MiniMax ≈0.77), consistent with more aggressive restyling costing more intent. Ministral also has the lowest script stability (0.85 against gemma's 0.98), so the three panels tell one story: the most eager restyler is also the least faithful one, which is the same responder gap that shows up at corpus scale in Section 6. This is exactly why the main-text comparison (Section 5) fixes the rewriter to isolate style *representation*, and why we report multi-judge agreement rather than trusting a single evaluator. The ablation here is a controlled probe on a stratified sample; it complements, rather than replaces, the corpus-scale metrics of Section 6.

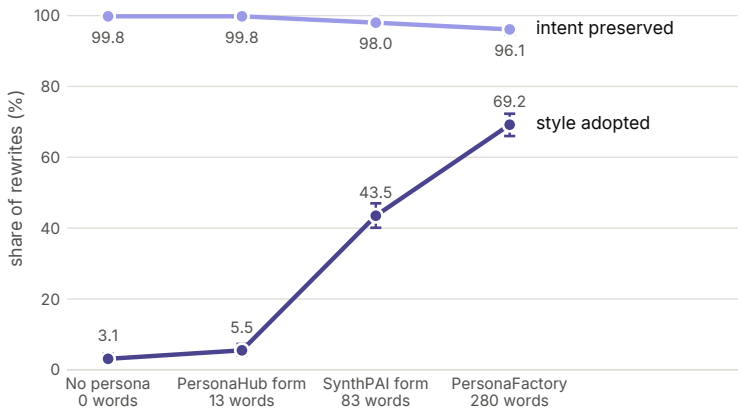


Figure 24: Style adoption and intent preservation by persona representation ($N = 800$ paired items, 95% Wilson intervals). Style adoption rises with the amount of writing behaviour the format carries, while intent preservation drops by about four points.

Table 23: Robustness of the ranking. Top: the same comparison with Ministral-3-14B as rewriter. Bottom: the pairwise comparison re-judged by four models on a 200-item subset.

	No persona	PersonaHub form	SynthPAI form
PF preferred, by rewriter			
gemma-4-31B	94.4% (17.90:1)	94.9% (20.46:1)	81.3% (4.83:1)
Ministral-3-14B	92.3% (12.70:1)	92.6% (13.55:1)	72.3% (2.64:1)
PF preferred, by judge (gemma rewriter)			
gemma	99.0%	97.5%	83.5%
glm	84.5%	81.5%	71.0%
minimax	93.9%	95.4%	82.5%
ministral	91.0%	88.5%	71.5%

J Human Study

In this section we describe the expert human study that complements the LLM-judge evaluation of Section 5. We first describe the study design (§J.1), then give the information and instructions shown to participants (§J.2), the ethics review and compensation (§J.3), the interim results (§J.4), and example items (§J.5).

J.1 Study Description

The study uses one form per language, 60 languages in total, each administered as a self-paced online session. A form contains up to 40 items, ten in each of four parts; forms for languages with fewer available samples contain fewer items. Participants are experts in the AI field who rate only the language varieties they know well. They may complete up to three sessions, one per language, and are identified only by a random participant ID. The four parts measure different properties and are never combined into a single score:

- **Part A, meaning preservation.** An original user message is shown next to its persona rewrite from the corpus, and participants rate whether the rewrite preserves the meaning (intent and content) of the original. For cross-lingual rewrites the original is the English source message.
- **Part B, style fidelity.** A persona's free-text chatting-style description from the persona bank is shown next to a message rewritten for that persona, and participants rate how well the message matches the described style without being an exaggerated stereotype or caricature.
- **Part C, human or model.** Participants judge from the text alone whether a message was written by a real person or generated by a language model, and state their confidence. Each form mixes

Table 24: Per-attribute augmentation analysis: the effect of varying a single persona attribute on three separately measured aspects of the rewrite, averaged over four generators. Script stability is reported relative to the no-persona baseline.

Persona attribute	Style adoption	Intent preservation	Script stability
Digital fluency	0.801	0.624	0.921
Sentence length	0.822	0.625	0.920
AI stance	0.800	0.622	0.916
Age	0.810	0.627	0.924
Gender	0.821	0.631	0.920
Education	0.808	0.620	0.929
Occupation	0.801	0.622	0.920
Hobbies	0.814	0.646	0.925
Baseline (no persona)	n/a	n/a	0.987



Figure 25: Every single-attribute condition scored on the three fidelity axes, averaged over the four rewriters. All three columns share one 0–1 scale, so the visible pattern is real: the columns sit at clearly different levels while the rows within a column barely move. The no-persona baseline has no style or intent score to report, since there is no rewrite to judge.

persona rewrites from the corpus with messages intended to be human-written user requests from a public chat-log corpus (Zhao et al., 2024), in equal proportion; the origin of each text is not disclosed. The answer options follow the wording of the SynthPAI human study (Yukhymenko et al., 2024).

- **Part D, linguistic plausibility.** Participants rate whether a short user message is plausible as something a real speaker of the language variety would write, and may report the item for personal information or offensive content.

Parts A, B and D and the Part C confidence use five-point scales. All displayed texts come from public research datasets or machine-generated rewrites and were filtered for personal information

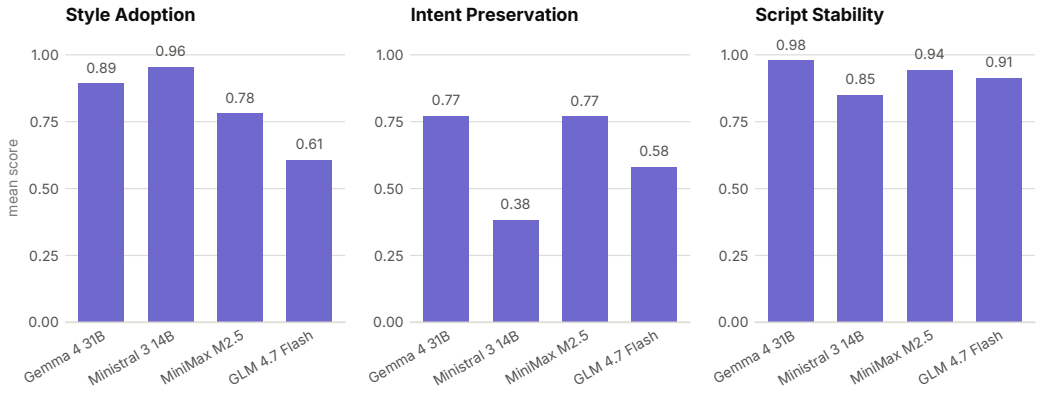


Figure 26: The same ablation summarised per rewriter, averaged over all sixteen attribute conditions. The spread across models dwarfs the spread across attributes, and the three panels expose a trade-off: Ministral restyles most aggressively yet preserves intent and script least reliably, while Gemma is the balanced choice.

and disturbing content before the study. Any item can be skipped or reported, and reported items are removed.

J.2 Information and Instructions for Participants

We reproduce the study information and instructions as presented to participants, omitting only the investigator names, institution, ethics application number, and contact details.

Study information

Purpose. This study evaluates a research pipeline that rewrites the user side of chat conversations so that language-model training data better reflects how linguistically diverse people write (informal register, second-language phrasing, code-switching) while preserving meaning. As an expert in the AI field, you judge properties of displayed texts. The study analyzes texts and models, not you; no analyses of individual participants are performed.

What participation involves. Each item shows one or two short chat messages with a rating question (meaning preservation, style fidelity, human-vs-LLM origin judgment, or linguistic plausibility). Some tasks ask you to judge whether a displayed text was written by a human or generated by a language model; the origin of each text is therefore not disclosed before you answer. This task structure is openly described here and involves no other withheld information. This session contains 40 study items. Completion time varies with reading pace and breaks. You may complete up to 3 sessions, one per language.

Risks. Minimal, comparable to everyday reading and web use. Texts come from public research datasets or machine-generated rewrites and are filtered for personal information and disturbing content; residually mildly offensive text cannot be fully excluded. You can skip or report any item; reported items are removed.

Voluntariness and withdrawal. Participation is entirely voluntary and unpaid, in your professional capacity. Persons in a dependency relationship to the investigators may not participate. You may skip, pause, or withdraw at any point without disadvantage. Deletion of your ratings is possible for as long as they can be linked to your participant ID; after anonymization at study completion it is no longer possible.

Data protection. Ratings, timestamps, and optional coarse language background are stored pseudonymously under a random participant ID. No names, contact details, or IP addresses are collected. Data are stored on institutional infrastructure with restricted access; consent records are retained securely and separately. Only aggregated or anonymized results are published.

Consent declaration

Before any study item is shown, participants confirm each of the following statements:

- I have read and understood the study information above and had the opportunity to ask questions.
- I am at least 18 years old.
- I understand that participation is voluntary and unpaid, and that I may skip items or withdraw at any time without giving reasons and without disadvantage.
- I understand how my data will be processed, stored, and published, as described above.
- I understand that some tasks involve judging whether a displayed text was written by a human or by a language model, and that the origin of each text is not disclosed to me before I answer.
- I agree to take part in this study.

Task instructions

Part A, meaning preservation. "You will see an ORIGINAL text and a REWRITE. Judge whether the rewrite preserves the meaning (intent and content) of the original. You may skip any item." Question: *Does the rewrite preserve the meaning of the original?* Scale 1 (*meaning changed*) to 5 (*meaning fully preserved*).

Part B, style fidelity. "You will see a short style description and a message. Judge how well the message matches the described style WITHOUT being an exaggerated stereotype or caricature." Question: *How well does the message match the described style, without caricature?* Scale 1 (*poor match / stereotyped*) to 5 (*faithful and plausible*).

Part C, human or model. "For each message, judge from the text alone whether it was written by a real person or generated by a language model. The origin of each text is not disclosed; some texts are model-generated, as described in the study information." Question: *Was this text written by a real human being?*

1. Yes: this text appears as authentic and readable as typical messages written by real users.
2. No: parts of this text seem unusual, such as odd wording or synthetic text patterns, suggesting it might be AI-generated.
3. Skip this item.

Follow-up: *How confident are you?* Scale 1 (*guessing*) to 5 (*very confident*).

Part D, linguistic plausibility. "Only rate items in language varieties you know well. Judge whether the message is plausible as something a real speaker of that variety would write, not an exaggerated caricature. Report any item that seems to contain personal information or offensive content." Question: *Is this plausible as something a real speaker would write?* Scale 1 (*implausible / caricature*) to 5 (*fully plausible*); report options: *No issue*, *Report: personal information*, *Report: offensive content*.

J.3 IRB Approval and Compensation

The study was reviewed and approved by the ethics review board (IRB) of the authors' institution and classified as minimal risk. All participants were informed about the study and its data processing and gave informed consent through the declaration above before any item was shown; they could skip items, pause, or withdraw at any time without giving reasons. Participation was voluntary and unpaid: participants are domain experts who took part in their professional capacity, and persons in a dependency relationship with the investigators were not eligible. No names, contact details, or IP addresses were collected, ratings are stored under random participant IDs, and only aggregated or anonymized results are reported.

J.4 Results

We report an interim snapshot of the study. It contains 11 submissions covering nine languages with 476 assigned item evaluations; only German and Vietnamese currently have two submissions each. All saved submissions are included without exclusions. Pooled statistics weight available ratings equally, not languages equally, and standard deviations describe dispersion, not uncertainty in the

mean. Table 25 gives pooled ratings, Table 26 the per-language means, Table 27 the Part C judgments, and Table 29 the agreement between the two German and the two Vietnamese submissions.

Dimension	n	Mean	SD	Median	Ratings 1–2
A: Meaning preservation	89	3.94	1.30	4.0	15.7%
B: Style fidelity	151	3.63	1.19	4.0	16.6%
D: Linguistic plausibility	147	3.64	1.35	4.0	20.4%
C: Confidence	85	3.85	0.88	4.0	8.2%

Table 25: Pooled five-point ratings at the interim cutoff. Means weight observed ratings equally; the last column is the share of ratings of 1 or 2.

Language	Subm.	Meaning (n)	Style (n)	Plausibility (n)
Bosnian	1	– (0)	4.00 (20)	3.40 (20)
Bulgarian	1	3.50 (10)	3.71 (17)	3.81 (16)
Croatian	1	3.11 (9)	3.13 (15)	3.64 (14)
English	1	4.60 (10)	4.60 (10)	4.80 (10)
German	2	4.60 (20)	4.30 (20)	4.50 (20)
Greek	1	2.30 (10)	1.94 (18)	2.00 (18)
Serbian	1	– (0)	3.20 (15)	3.80 (15)
Ukrainian	1	4.70 (10)	4.60 (10)	4.70 (10)
Vietnamese	2	4.00 (20)	3.73 (26)	3.21 (24)

Table 26: Mean five-point ratings per language, with the number of available ratings in parentheses. Dashes mark languages without Part A ratings. Each language is rated on its own panel of items, so differences can reflect item selection and rater severity as well as language quality; these means are not a language ranking.

Source status	Judgments	Human-like	AI-like	Skip
Verified model rewrite	44	26 (59.1%)	18 (40.9%)	0
Origin unverified	44	29	12	3
All Part C	88	55	30	3

Table 27: Part C judgments. A text is a verified model rewrite if it matches a user message of the corpus exactly; the remaining texts are not individually verified as human-written, so only the rate on the verified subset is interpretable. The 44 verified judgments cover 34 distinct texts.

LLM judges on Part C. We also give the Part C texts of all answered panels (68 texts: 34 verified rewrites and 34 chat-log texts) to four LLM judges, `Claude Sonnet 4.5`, `MiniMax-M2.5`, `Qwen3.6-35B-A3B` and `Muse-Glimmer-30B`. Every judge sees the question and answer options shown to the raters and returns a JSON answer (human or model) with a confidence from 1 to 5. The three open-weight judges are served locally (`MiniMax-M2.5` with SGLang, the other two with vLLM) with their default chat templates and reasoning, and receive one text per request through the chat API; each text is sampled three times at temperature 0.7 with up to 8,192 new tokens, and the answer is read from the final JSON object. `Claude Sonnet 4.5` is queried in three independent sessions without tool access, each receiving all 68 texts in one prompt under neutral ids in a shuffled order and judging each text independently. It is run through the Claude Code agent interface, whose sampling parameters are fixed by the provider and not user-configurable. All 816 answers parse. We do not use `Gemma-4-31B-it` as a judge because it wrote part of the rewrites. For accuracy, the chat-log texts count as human-written by study design, although they are not individually verified; the human-rater accuracy uses the same convention over the 85 answered judgments. Table 28 reports the results. All judges label most rewrites as human-written (83.3 to 89.2%), more often than the raters (59.1%), and none separates the two sources clearly above chance (50.5 to 57.4% accuracy).

Findings. Meaning preservation, style fidelity and plausibility are rated favorably overall (pooled medians of 4), and most strongly for English, German and Ukrainian; no English or Ukrainian rating

Judge	Rewrites judged human (%)	Accuracy (%)
Human raters	59.1 [44.4, 72.3]	55.3 [44.7, 65.4]
`Claude Sonnet 4.5`	83.3 [74.9, 89.3]	57.4 [50.5, 63.9]
`MiniMax-M2.5`	89.2 [81.7, 93.9]	50.5 [43.7, 57.3]
`Qwen3.6-35B-A3B`	88.2 [80.6, 93.1]	52.9 [46.1, 59.7]
`Muse-Glimmer-30B`	83.3 [74.9, 89.3]	54.9 [48.0, 61.6]

Table 28: Human-or-model judgments (Part C) on the 68 texts of the answered panels: 34 verified corpus rewrites and 34 chat-log texts. Rewrites judged human is the share of verified rewrites labeled human-written; accuracy counts a rewrite labeled model-generated or a chat-log text labeled human-written as correct. LLM judges answer each text three times; 95% Wilson intervals in brackets.

Language	Dimension	Items	Exact	Within 1	κ
German	A: Meaning preservation	10	60.0%	90.0%	0.077
German	B: Style fidelity	10	40.0%	100.0%	0.455
German	D: Linguistic plausibility	10	60.0%	80.0%	0.405
German	C: Human/model choice	10	80.0%	–	0.583
Vietnamese	A: Meaning preservation	10	60.0%	90.0%	0.583
Vietnamese	B: Style fidelity	13	38.5%	84.6%	0.493
Vietnamese	D: Linguistic plausibility	12	50.0%	75.0%	0.545
Vietnamese	C: Human/model choice	10	60.0%	–	0.200

Table 29: Agreement between the two submissions of a language on shared items. Numeric ratings use quadratic-weighted Cohen's κ; the human/model choice uses unweighted κ. Estimates rest on few shared items and are exploratory; agreement does not imply accuracy.

in Parts A, B or D falls below 3. The Greek panel is the clear exception: five of ten meaning ratings and eleven of eighteen style ratings are 1, which marks its samples for review, although the evidence comes from a single submission. Quality also varies by dimension within a language: Vietnamese meaning preservation (4.00) is rated higher than plausibility (3.21), and Croatian is rated lower on meaning and style (3.11 and 3.13). On texts verified as model rewrites, raters selected the human-written option in 59.1% of judgments. Because the *Yes* option also describes the text as authentic and readable, this is evidence of perceived human-likeness, not a measure of detection accuracy or of indistinguishability from human writing; overall a fully verified accuracy requires a human-source key for the remaining texts (Table 28 uses the design labels). The Ukrainian panel was evaluated on the revised selections of Parts A and B (informal and task-diverse rewrites), which replaced the earlier selection before any Ukrainian submission; no before-after comparison is therefore possible. Part D received five personal-information reports (four Bosnian, one Serbian), which we review manually rather than treat as confirmed privacy issues.

Limitations of the interim study. Most languages currently have a single submission, and panel size, item selection, source language and rater severity confound cross-language comparisons, so we report no population-level confidence intervals or significance tests. The study has no baseline arm, so it does not show an improvement over plain translation or another persona method. Collection is ongoing: we are adding raters for the same items, particularly for low-scoring panels, and completing the human-source key for Part C.

J.5 Example Items

We show example items from the English form. Part C items are omitted because their origin labels are part of the ongoing study.

Part A: meaning preservation

Original: What role can technology play in facilitating Urdu language learning outside of South Asia?
Rewrite: hep i was wondering what role technology can play in helping people learn Urdu

outside south asia

Original: Write a Python program that takes two lists as input and determines if the first list is a subset of the second list. The program should output either “True” or “False”, depending on whether the first list is a subset of the second list. Ensure that your program uses appropriate variable names, follows PEP 8 guidelines for style, and includes relevant comments to explain the logic of the code.
Rewrite: can u write a python program that takes two lists as input and determines if the first list is a subset of the second list? it just needs to output “True” or “False” depending on whether it is a subset :) please make sure it uses appropriate variable names, follows pep 8 guidelines for style, and includes relevant comments to explain the logic of the code.

Part B: style fidelity

Style description: Writes in all lowercase with minimal punctuation and occasional technical shorthand, reflecting a fast-paced digital native style.
Message: can you do a deep dive review on that new ergo keyboard you used / focus on the features and how the comfort holds up during long sessions / need info on the adjustments and customization options and how they improve the experience / also talk about the build quality and form factor plus any extras like wrist rests or touchpads / give some examples of specific tasks you did and how it performed / make it detailed enough to show the strengths and limits so people know if it’s worth the investment

Style description: She types like she is writing a formal letter, using full capitalization and ending every sentence with a period, often adding polite pleasantries.
Message: Hello. I have been thinking about how populations continue to grow even after birth rates slow down, and I came across the term demographic momentum. It would be lovely if you could explain how demographic momentum works in current global trends, especially with changes in technology and healthcare. Could you please provide some examples of where it is having a big impact? Thank you.

Slashes mark line breaks in the original message.

Part D: linguistic plausibility

can u check why this script is throwing a timeout error? looks like a network glitch machan

bro, my bike brake is making noise again. is there any way to fix it with basic tools?