Can Large Language Models Personalize Dialogues to Generational Styles?

Pier Felice Balestrucci[1], Ondřej Dušek[2], Luca Anselma[1], Alessandro Mazzei[1]

[1]University of Turin, Computer Science Department, Turin, Italy
[2] Charles University, Faculty of Mathematics and Physics, Prague, Czechia
{pierfelice.balestrucci, luca.anselma, alessandro.mazzei}@unito.it
odusek@ufal.mff.cuni.cz

Abstract

We investigate how large language models (LLMs) can produce personalized dialogue responses, specifically focusing on whether they reflect linguistic styles pertaining to different generations: Baby Boomers, Generation X, Generation Y, and Generation Z. We create P-MultiWoZ, a personalized, generation-specific version of MultiWOZ 2.2, by prompting LLMs, and validate its alignment with the original dataset through automatic and human evaluations. To validate the appropriateness of generational linguistic traits, we introduce GeMoSC, a corpus of generation-annotated movie dialogues. Linguistic analysis and perplexity test suggest that P-MultiWoZ reflects patterns consistent with GeMoSC. Finally, a human evaluation reveals that annotators were able to mostly correctly identify the generation behind P-MultiWoZ dialogues, based only on a single query-reply pair.[1]

1 Introduction

Communication style is a complex interplay of linguistic, cultural, and contextual elements that shape interactions. Understanding and adapting to different styles can significantly enhance the effectiveness and satisfaction of communication (Miehle et al., 2022, 2020). Various factors influence communication styles, including age (Jansen et al., 2022), self-concept (Hansford and Hattie, 1987), education (Mackenzie, 2000), all of which shape how individuals express themselves and engage with others. Several studies compare how individuals of different generations communicate (Subramaniam and Razak, 2014; Raslie, 2021), taking into account factors such as exposure to younger generations' language through social media (Conny et al., 2024). These distinctions are typically based on categorizations such as Baby Boomers (*Boomers*, born between 1946-64), Generation X (*Gen X*, 1965-80), Generation Y/Millennials (*Gen Y*, 1981-96), and Generation Z (*Gen Z*, 1997-2012). Understanding communication style differences is crucial for designing adaptable communication strategies.

With the rapid advancement of large language models (LLMs), there is growing interest in evaluating their ability to emulate human language (Jumelet et al., 2024), particularly in adapting to generational styles and preferences (Liu et al., 2024a). By adapting to different styles, personalities, and user preferences, LLMs can facilitate more natural and engaging interactions (Su et al., 2021; Xu et al., 2023; Zheng et al., 2021).

This paper examines generational communication styles as a case study of personalization in LLMs. We pose the crucial question: *"Can user profiles be integrated into LLM prompts to personalize dialogue based on generational communication styles?"* to understand whether LLMs contain implicit knowledge of how individuals from different generations express themselves and whether this knowledge allows for the personalization of dialogues.

We choose the dialogue domain to address this question since user-specific linguistic traits, such as generational styles, become more apparent in less formal situations (Mairesse et al., 2007). Furthermore, task-oriented dialogues are particularly well-suited for this purpose, as they emphasize the importance of aligning system outputs with user preferences. Therefore, we use the MultiWOZ 2.2 dataset (Budzianowski et al., 2018; Zang et al., 2020) as a testbed. We create generation-specific speaker profiles and paraphrase MultiWOZ dialogues through LLMs, resulting in the personalized corpus P-MultiWOZ. The final P-MultiWOZ consists of 240 dialogues (60 per generation variant: P-MultiWOZ B, X, Y, Z), with a total of 14,720 user-system turns, where both user and system utterances are style-adapted. Additionally, we build

[1]All the data and code from this paper are released under CC-4.0-BY license at: https://github.com/PierBale/P-MultiWoZ

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 64–77
November 4-9, 2025

a custom corpus, GeMoSC (Generational Movie Script Corpus), containing utterances of movie characters belonging to different generations, to conduct detailed linguistic analyses and comparisons.

We assess content consistency of P-MultiWOZ with the original MultiWOZ dataset and evaluate whether the generation-specific paraphrases effectively reflect generational styles, through automatic analyses and human annotations. The results show that P-MultiWOZ retains the same content as the original dataset and shares linguistic patterns corresponding to the generations represented in the movie corpus. Furthermore, human annotators were able to recognize whether a conversation extracted from P-MultiWOZ belonged to a specific generation in most cases. This demonstrates that LLMs can personalize conversations based on generational communication styles, reflecting linguistic patterns that align with those of real individuals.

2 Related Work

Numerous studies have investigated the factors contributing to effective conversations (See et al., 2019; Serban et al., 2016; Mazzei et al., 2022), highlighting the importance of personalizing dialogue systems to enhance their capabilities. Notably, conversational dynamics are significantly influenced by factors such as an individual's knowledge (Janarthanam and Lemon, 2014), physical abilities (Nuovo et al., 2024) and age (Pennebaker and Stone, 2003). Moreover, successful interactions often depend on the ability of interlocutors to adapt to their dialogue counterpart (Friedberg et al., 2012), an issue mostly overlooked in current dialogue systems (Kumar and Dusek, 2024).

Several approaches to conversational models increasingly aim to integrate personal characteristics—such as age, gender, geographical location, and physical abilities—through both explicit mechanisms (Jansen et al., 2022; Qian et al., 2017) and implicit modeling (Kottur et al., 2017; Gur et al., 2018; Balestrucci et al., 2024). LLMs have emerged as powerful tools for generating and personalizing text based on diverse styles, by prompting strategies and fine-tuning. Recent studies have demonstrated their ability to adapt linguistic output to specific stylistic preferences, making them valuable for applications ranging from automated writing assistance to dialogue systems (Reif et al., 2022; Liu et al., 2024b).

While corpora designed to embed personal information into conversations are emerging, creating opportunities to enhance dialogue personalization (Chen et al., 2024; Kim et al., 2023), current research often overlooks user-specific traits such as generational conversational patterns. Our work highlights the importance of adapting dialogue systems to better align with users' characteristics.

3 P-MultiWOZ

To evaluate LLMs' ability to produce generation-specific language, we create P-MultiWOZ, a personalized version of the MultiWOZ 2.2 (Zang et al., 2020) task-oriented dialogue dataset. We first prompt an LLM to define user profiles that reflect the distinctive communication styles of Boomers, Gen X, Gen Y, and Gen Z. These profiles guide further prompting of LLMs to simulate generational communication styles and generate paraphrased versions of a subset of MultiWOZ 2.2 dialogues.

To ensure content consistency, we evaluate the paraphrased dialogues through an automatic analysis using LLMs, complemented by a manual annotation of a random small subset of these dialogues. Based on these evaluations, we identify the most effective LLM for generating the final paraphrased dialogues. We then analyze how closely these dialogues align with the expected styles of their corresponding generations (see Sections 4 and 5).

MultiWOZ 2.2 (Zang et al., 2020) is a large-scale, multi-domain, human-human conversational corpus of over 10,000 dialogues covering multiple domains such as booking hotels, restaurants, and transportation services, making it one of the most comprehensive and challenging datasets for building robust dialogue systems capable of handling complex, multi-turn interactions across different contexts (Peng et al., 2021).

User Profile Generation To define profiles reflecting generational communication styles, we adopted the initial step of the pipeline proposed by Li et al. (2024), i.e., prompting GPT-4o[2] with an emphasis on two key attributes: the birth year and gender of the users.[3] We generated a total of 20 user profiles, evenly distributed with 5 profiles per generation and 10 profiles per gender.

Trial Generation Based on the user profiles, we prompted two of the most prominent mid-sized

[2] https://openai.com/index/hello-gpt-4o/

[3] All prompts used here are provided in Appendix A.

Paraphrase LLM		FLAN-T5 large EM
Mistral 7B v0.3	DA	68.77%
	Slot-Value	90.29%
Llama 3.1 8B	DA	67.20%
	Slot-Value	89.17%

Table 1: Average exact match (EM) rates for dialogue act (DA) and slot-value classifications, comparing LLM-produced paraphrases to original MultiWOZ utterances.

Paraphrase LLM	#I	#NC	#M	#Others
Llama 3.1 8B	3	4	10	0
Mistral 7B v0.3	1	3	6	0

Table 2: Error analysis across 104 utterances for each model, with numbers of errors found. #I: Incorrect — paraphrased content contradicts the original; #NC: Not Checkable — content cannot be verified against the original; #M: Misleading — paraphrased content misleads relative to the original; #Others: Other Issues — errors due to grammar, style, irrelevance, or redundancy.

open-source models, LLaMa 3.1 8B (Dubey et al., 2024) and Mistral 7B Instruct v0.3,[4] to generate paraphrased versions of a subset of dialogues from MultiWOZ 2.2.[3] We generated 385 paraphrased turns per model.

Automatic Quality Verification To assess that the content of the paraphrases has not changed, inspired by previous works (Zhu et al., 2023; Tavares et al., 2023), we prompt FLAN-T5 large (Chung et al., 2024) to obtain classifications of dialogue acts and slot-value pairs for the paraphrases.[5] We expect the paraphrases to align with the gold MultiWOZ annotations. We designed prompt templates for the dialogue act and slot value classification, explicitly providing the LLMs with three response options[6] (i.e., the LLM is asked to pick the intent for the utterance or the value of a single slot).[3]

Table 1 presents the results of this analysis. The results are positive compared to Tavares et al. (2023), who employ similar strategies. The lower scores for dialogue act classification are likely attributable to the higher complexity arising from the multiple definitions of intents and domains, whereas questions focusing on finding a single slot value (e.g., "north" for the slot "area") are more straightforward to resolve.

Human Review To verify the automatic results, we recruited two volunteer annotators with good English proficiency, who reviewed 208 randomly selected utterances (evenly split between the two models and generations) and annotated errors using the taxonomy from Kasner and Dusek (2024). Both models performed well, exhibiting few errors across all categories (see Table 2). The annotators achieved an inter-annotator agreement measured by Cohen's κ (Cohen, 1960) of 0.84, indicating near-perfect agreement (See Appendix C for further details).

Final P-MultiWOZ Generation We picked Mistral 7B Instruct v0.3 as the primary model for additional turn generation due to its slightly better performance in preserving both content and intent. The final P-MultiWOZ consists of 60 dialogues, encompassing 1,840 user-system turns for each generation (dubbed P-MultiWOZ B, P-MultiWOZ X, P-MultiWoz Y, P-MultiWOZ Z). For example, P-MultiWOZ Z includes the sentence "Hey there! Got any info on trains heading to Cambridge, departing on Saturday? 😄" as a paraphrase of the original sentence "Can you help me find a train going to Cambridge leaving on Saturday?"

4 Evaluating Generational Differences

To check how well the different P-MultiWOZ corpora reflect generational differences, we built a new dataset based on movie scripts that contain dialogues from characters belonging to different generations. This dataset, called Generational Movie Script Corpus (GeMoSC) is used to (1) linguistically compare and identify common features with the P-MultiWOZ corpora, (2) fine-tune generation-specific LLMs and calculate their perplexity on the various P-MultiWOZ corpora. This is to verify that, e.g., the model finetuned on the Boomer GeMoSC corpus shows a lower perplexity when evaluated on the Boomer corpus from P-MultiWOZ.

GeMoSC We started with the publicly available Movie Scripts Corpus dataset,[7] which includes a comprehensive list of films (1909–2021) and their scripts, organized by character. To assign characters to a generation, one of the authors manually annotated selected scripts, classifying characters' gender and year of birth using the online IMDb

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[5] Slot-value pairs represent structured attributes extracted from user utterances, such as [pricerange: cheap] in a hotel booking request.

[6] The three-option choice simplifies LLM output parsing.

[7] https://www.kaggle.com/datasets/gufukuro/movie-scripts-corpus

Corpus	MSTTR	Tokens	Intj.	Dep.L.
GeMoSC-B	0.67	21.15	0.23	44.85
GeMoSC-X	0.67	19.55	0.23	40.76
GeMoSC-Y	0.68	19.15	0.25	40.11
GeMoSC-Z	0.67	19.96	0.26	41.56
P-MW B (U)	0.65	17.93	0.17	44.20
P-MW B (S)	0.67	25.37	0.12	59.08
P-MW X (U)	0.66	15.02	0.23	36.10
P-MW X (S)	0.67	23.46	0.10	53.79
P-MW Y (U)	0.67	16.99	0.35	40.26
P-MW Y (S)	0.68	24.06	0.10	54.42
P-MW Z (U)	0.66	17.99	0.42	52.93
P-MW Z (S)	0.69	23.18	0.10	51.62
MW 2.2 (U)	0.63	13.65	0.41	28.88
MW 2.2 (S)	0.66	17.89	0.20	37.55

Table 3: Statistics across GeMoSC, P-MultiWOZ (P-MW) and MultiWOZ 2.2 (MW): Mean Segmental Type-Token Ratio (MSTTR), average number of tokens (Tokens), of interjections (Intj.), of named entities (NE), and average dependency length (Dep.L.). (U) and (S) refer to user vs. system utterances.

database,[8] following selection guidelines to ensure realistic dialogues (see Appendix E). The process assumes that screenwriters create characters who speak in a manner consistent with their supposed year of birth, which is in line with basic screenwriting conventions (McKee, 1997; Dancyger and Rush, 2012). We annotated a total of 75 films (2018-2021) and refined the scripts by removing text describing character actions or scene directions, applying simple pattern-matching rules.

Linguistic Analysis The linguistic analysis in Table 3 highlights several features across GeMoSC, P-MultiWOZ, and MultiWOZ 2.2[9] corpora, grouped by generations and turns (user and system).

The mean segmental type-token ratio remains consistent across all datasets, indicating similar lexical diversity. System turns are notably longer and syntactically more complex, as shown by higher token counts and dependency lengths reflecting their task-oriented design. Notably, MultiWOZ 2.2 exhibits lower dependency lengths and token counts in user turns, suggesting more concise and less complex interactions compared to P-MultiWOZ. Generational differences are particularly evident in expressiveness. Younger simulated users, such as those in GeMoSC-Z and P-MultiWOZ Z, use interjections more frequently than older generations. Conversely, the number of interjections is lower in system responses, which remain more neutral. For instance, the texts in P-MultiWOZ Z are more direct, incorporating slang, abbreviations, and even emojis, as seen in the following examples: "*Alright, cool! Let's set up Ashley for 8 peeps, from Saturday for 3 nights. Got it?*" or "*Absolutely, that works for me!*". Similarly, this immediacy is reflected in some excerpts from GeMoSC-Z, such as "*Olivia's cool, yeah?*" or "*I'm going to see if it works*". P-MultiWOZ, thus, shares key features with human communication styles, exhibiting consistent lexical diversity alongside variations in syntactic complexity and generational expressiveness.

Perplexity Test Perplexity (Jelinek et al., 1977) evaluates a language model's ability to predict word sequences, with lower values indicating greater confidence and accuracy. We use it to check if a model fine-tuned on a generation-specific GeMoSC corpus exhibits lower perplexity when tested on the corresponding corpus from P-MultiWOZ. Table 4 presents the perplexity scores of LLaMa 3.1 8B evaluated on the P-MultiWOZ generational datasets, both in its base version and after fine-tuning on the GeMoSC generational corpora: B (Boomers), X (Gen X), Y (Gen Y), and Z (Gen Z). Details on the experimental settings are provided in Appendix F. Fine-tuning on GeMoSC generational corpora consistently improves the model's performance on corresponding P-MultiWOZ datasets. In particular, LLaMa-B, -Y, and -Z achieve the best alignment with Boomer, Gen Y, and Gen Z datasets, respectively, demonstrating the effectiveness of targeted fine-tuning in capturing generational linguistic patterns. The lower performance P-MultiWOZ X can be probably attributed to the base model's weaker performance on GeMoSC-X. To further investigate the impact of fine-tuning, we performed a Mann-Whitney U test (Mann and Whitney, 1947) to evaluate whether the differences in loss values between the fine-tuned models and the baseline were statistically significant. The results indicate highly significant differences in all comparisons, with $p \ll 0.005$ in every case ($p \approx 1.48 \times 10^{-16}$). However, no statistically significant differences were found among the fine-tuned generational models.

5 Human Evaluation

To further assess whether P-MultiWOZ adapts to generational communication styles, we conducted

[8] https://www.imdb.com

[9] For this analysis, only the dialogues shared between MultiWOZ 2.2 and P-MultiWOZ are considered.

Model	P-MW B	P-MW X	P-MW Y	P-MW Z
LLaMa-B	**7.29**	7.52	7.36	8.26
LLaMa-X	7.53	7.53	7.36	8.26
LLaMa-Y	7.42	7.42	**7.26**	8.13
LLaMa-Z	7.38	**7.35**	7.38	**8.12**
LLaMa Base	233.60	277.18	242.18	243.16

Table 4: Perplexity scores for LLaMa 3.1 8B on P-MW (P-MultiWOZ), both in its base versions and after fine-tuning on GeMoSC's generational corpora: B (Boomers), X (Gen X), Y (Gen Y), and Z (Gen Z).

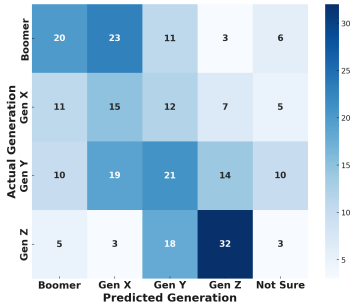

Figure 1: Confusion Matrix illustrating human classification performance across generational categories.

a human evaluation. Participants were tasked with identifying the generation of P-MultiWOZ conversations and providing feedback. We recruited 26 English-proficient volunteers via mailing lists and supplied them with a Qualtrics[10] questionnaire, estimating a 15-minute completion time. Participants provided informed consent and answered profiling questions on gender and birth year. Of the 26 participants, 19 were male, 7 female; 11 were from Gen Z, 11 Gen Y, and 4 Gen X. Each participant completed 5 questions, each featuring 5 pairs of conversation excerpts contrasting user-system turns from two generations. Participants identified the perceived generation or selected "Not Sure." Figure 1 shows a confusion matrix summarizing the results. A pattern emerges in which participants were able to identify the correct generation more consistently than the others. This was particularly true for Gen Z, where participants more easily recognized these dialogues. The highest confusion occurred between consecutive generations, e.g., Boomer conversations were often attributed to Gen X. Participants' comments supported these findings, highlighting subtle differences between consecutive generations. The predominance of Gen Z and Gen Y participants likely facilitated recognition of their own generations.

[10] https://www.qualtrics.com

%	Boomer	Gen X	Gen Y	Gen Z
Boomer	35.09	40.35	19.3	5.26
Gen X	24.44	33.33	26.67	15.56
Gen Y	15.62	29.69	32.81	21.88
Gen Z	8.62	5.17	31.03	55.17

Table 5: Confusion matrix (%) of the human evaluation across generations.

%	Precision	Recall	F1-Score
Boomer	43	35	39
Gen X	25	33	29
Gen Y	34	33	33
Gen Z	57	55	56

Table 6: Precision, Recall, and F1-score (%) for identifying each generation in the human evaluation.

It is important to note, however, that the classification accuracy observed in Figure 1 does not directly reflect the overall quality of the P-MultiWOZ dataset. In fact, distant generations exhibit greater discrimination power, while most misclassifications occur between adjacent generations, suggesting that certain stylistic overlaps exist across generational boundaries, as shown in Table 5. This phenomenon may reflect the inherent difficulty of the task, rather than a limitation of the dataset itself.

Indeed, Table 6 shows that Gen Z dialogues are recognized with the highest reliability, while Gen X and Gen Y show lower and more comparable performance. Boomer dialogues lie in between, suggesting that style cues for older and younger generations are easier to distinguish, whereas consecutive generations overlap more strongly.

From this analysis, we can conclude that the LLM-generated P-MultiWOZ data capture generational communication styles consistent with those of real individuals.

6 Limitations

This study represents a significant first step toward the generational stylistic analysis of language models, although it has several limitations. The models employed are small due to computational constraints, yet they yield interesting and promising results. Mid-sized and larger LLMs may offer higher performance but also incur substantially greater computational costs. The P-MultiWOZ dataset, though limited in size, has shown potential in emulating certain generational linguistic features, and we plan to expand it in the future to make it a 1:1 replica of MultiWOZ, ensuring greater representa-

tiveness. The GeMoSC dataset, despite being small as well, represents what we consider an original and important contribution, born from an extensive manual annotation process. We chose to create this dataset instead of using pre-existing conversational data, such as that extracted from social media, which could compromise user privacy, as part of an ethical choice. Our annotation and selection guidelines (see Appendix E) ensure the texts conform to the appropriate generational styles.

Paraphrasing existing dialogues instead of producing free-form new dialogues may constrain the capture of generational patterns beyond style, such as openness. While this mitigates hallucinations and reduces stereotype reinforcement, it may also limit the representation of deeper generational aspects in dialogue acts.

Lastly, the human experimentation involved an unbalanced number of participants representing different generations, with the Boomer generation being absent. Despite these limitations, the results still confirm the potential of LLMs in emulating distinctive generational styles, allowing for further research in this direction.

7 Ethical Considerations

The human evaluation campaign, crucial for manually verifying the personalization of P-MultiWOZ, involved several anonymous volunteer annotators with the sole requirement of being proficient in English. The annotation task took approximately 15 minutes to complete. Furthermore, before participating in the experiment, they signed an informed consent form explicitly stating, among other things: *i*) to be aware of the objectives of this research; *ii*) to participate on a voluntary basis; *iii*) to be of legal age; *iv*) to be aware that the study complies with current data processing and protection regulations, both at the national and EU level; *v*) to be aware of the possibility of withdrawing from the study at any time, without explanation, without any penalty, and with the assurance that their data would not be used.

Moreover, we acknowledge that personalizing LLMs based on generational traits risks reinforcing stereotypes, leading to biased assumptions and unfair treatment. To mitigate these risks, transparency measures, bias audits, and fairness evaluations are essential. Our future work will focus on balancing personalization with ethical safeguards to ensure inclusivity and prevent unintended harm.

Acknowledgements This work was partially funded by the 'Multilingual personalization through perspective-aware Language Modeling' project in partnership with Amazon Alexa. This work was co-funded by the European Union (ERC, NG-NLG, 101039303). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Pier Felice Balestrucci, Silvia Casola, Soda Marem Lo, Valerio Basile, and Alessandro Mazzei. 2024. I'm sure you're a real scholar yourself: Exploring Ironic Content Generation by Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494, Miami, Florida, USA.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large-scale Multi-domain Wizard-of-oz Dataset for Task-oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024*, pages 13650–13665, Torino, Italy.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-finetuned Language Models. *J. Mach. Learn. Res.*, 25:70:1–70:53.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Conny Conny, Nudia Yultisa, Rakhmat Wahyudin, and Tri Indah Rezeki. 2024. Linguistic Shift Among Gen Z in Computer-mediated Communication. *English Review: Journal of English Education*, 12(3):959–970.

Ken Dancyger and Jeff Rush. 2012. *Alternative Scriptwriting: Successfully Breaking the Rules*. Routledge.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Heather Friedberg, Diane J. Litman, and Susannah B. F. Paletz. 2012. Lexical Entrainment and Success in Student Engineering Groups. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL*, pages 404–409, USA.

Izzeddin Gur, Dilek Hakkani-Tür, Gökhan Tür, and Pararth Shah. 2018. User Modeling for Task Oriented Dialogues. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018*, pages 900–906, Athens, Greece.

B. Hansford and J. Hattie. 1987. Perceptions of Communicator Style and Self-concept. *Communication Research*, 14:189 – 203.

Srinivasan Janarthanam and Oliver Lemon. 2014. Adaptive Generation in Dialogue Systems Using Dynamic User Modeling. *Comput. Linguistics*, 40(4):883–920.

Lennert Jansen, Štěpán Lars Laichter, Arabella Sinclair, Margot van der Goot, Raquel Fernandez, and Sandro Pezzelle. 2022. Controllable Text Generation for All Ages: Evaluating a Plug-and-play Approach to Age-adapted Dialogue. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 172–188, Abu Dhabi, United Arab Emirates (Hybrid).

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Jaap Jumelet, Willem H. Zuidema, and Arabella Sinclair. 2024. Do Language Models Exhibit Human-like Structural Priming Effects? In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 14727–14742, Bangkok, Thailand.

Zdeněk Kasner and Ondrej Dusek. 2024. Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12045–12072, Bangkok, Thailand.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona Expansion with Commonsense Knowledge for Diverse and Consistent Response Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik*, pages 1131–1141, Croatia.

Satwik Kottur, Xiaoyu Wang, and Vítor Carvalho. 2017. Exploring Personalized Neural Conversational Models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3728–3734, Melbourne, Australia.

Nalin Kumar and Ondrej Dusek. 2024. LEEETs-Dial: Linguistic Entrainment in End-to-end Task-oriented Dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 727–735, Mexico City, Mexico.

Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2024. Curriculum-driven Edubot: A Framework for Developing Language Learning Chatbots through Synthesizing Conversational Data. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2024*, pages 400–419, Kyoto, Japan.

Siyang Liu, Trisha Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024a. The Generation Gap: Exploring Age Bias in the Value Systems of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL*, pages 19617–19634, USA.

Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024b. Customizing Large Language Model Generation Style using Parameter-efficient Finetuning. In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024*, pages 412–426, Tokyo, Japan.

Catherine Mackenzie. 2000. Adult Spoken Discourse: The Influences of Age and Education. *International journal of language & communication disorders*, 35 2:269–85.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500.

Henry B Mann and Donald R Whitney. 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. *The annals of mathematical statistics*, 18(1):50–60.

Alessandro Mazzei, Luca Anselma, Manuela Sanguinetti, Amon Rapp, Dario Mana, Md. Murad Hossain, Viviana Patti, Rossana Simeoni, and Lucia Longo. 2022. Anticipating User Intentions in Customer Care Dialogue Systems. *IEEE Trans. Hum. Mach. Syst.*, 52(5):973–983.

Robert McKee. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. ReganBooks.

Juliana Miehle, Isabel Feustel, Julia Hornauer, Wolfgang Minker, and Stefan Ultes. 2020. Estimating User Communication Styles for Spoken Dialogue Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, pages 540–548, Marseille, France.

Juliana Miehle, Wolfgang Minker, and Stefan Ultes. 2022. When to Say What and How: Adapting the Elaborateness and Indirectness of Spoken Dialogue Systems. *Dialogue & Discourse*, 13(1):1–40.

Elisa Di Nuovo, Manuela Sanguinetti, Pier Felice Balestrucci, Luca Anselma, Cristian Bernareggi, and Alessandro Mazzei. 2024. Educational Dialogue Systems for Visually Impaired Students: Introducing a Task-oriented User-agent Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC/COLING)*, pages 5507–5519, Torino, Italy.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9:907–824.

James W Pennebaker and Lori D Stone. 2003. Words of Wisdom: Language Use Over the Life Span. *Journal of personality and social psychology*, 85(2):291.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning Personality/Identity to a Chatting Machine for Coherent Conversation Generation. *CoRR*, abs/1706.02861.

Humaira Raslie. 2021. Gen Y and gen Z communication style. *Studies of Applied Economics*, 39(1).

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, MN, USA.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. *CoRR*, abs/1611.06216.

Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. 2021. PROTOTYPE-TO-STYLE: Dialogue Generation With Style-aware Editing on Retrieval Memory. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2152–2161.

Vithya Subramaniam and Norizan Abdul Razak. 2014. Examining Language Usage and Patterns in Online Conversation: Communication Gap Among Generation Y and Baby Boomers. *Procedia-Social and Behavioral Sciences*, 118:468–474.

Diogo Tavares, David Semedo, Alexander Rudnicky, and Joao Magalhaes. 2023. Learning to Ask Questions for Zero-shot Dialogue State Tracking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2118–2122.

Weilai Xu, Fred Charles, and Charlie Hargood. 2023. Generating Stylistic and Personalized Dialogues for Virtual Agents in Narratives. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023*, pages 737–746, London, United Kingdom.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. *CoRR*, abs/2007.12720.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized Dialogue Response Generation Using Stylized Unpaired Texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, pages 14558–14567, Virtual Event.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic, and Minlie Huang. 2023. ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–123, Singapore.

A Prompts

Generate 20 unique user profiles that represent a diverse range of socio-demographic characteristics, with a focus on gender and age groups. Each profile should demonstrate a distinct communication style, highlighting how individuals from different generations and genders might express themselves. Ensure balanced representation across the following age groups and genders: Boomers (born 1946–1964), Gen X (born 1965–1980), Gen Y (born 1981–1996), Gen Z (born 1997–2012). For each profile, provide:

- Gender
- Generation
- Communication Style

Figure 2: Prompt used for generating 20 user profiles equally distributed for generations and years of birth.

You are impersonating the following user: '{profile}' and asking for information from an assistant. Paraphrase the user's request '{user_request}' while maintaining only the original content and intent.
Do not introduce any additional details, such as communication style, personal experiences, or changes to the sentence structure.

Figure 3: Prompt used for paraphrasing the user request in MultiWOZ 2.2, where `profile` represents the characteristics of the user, `user_request` represents the user's request.

Given the following user profile: '{profile}' and the user request '{user_request}', you are an assistant who helps the user. Paraphrase the system output '{system_answer}' while maintaining only the original content and intent.
Do not introduce any additional details, such as communication style, personal experiences, or changes to the sentence structure.

Figure 4: Prompt used for paraphrasing the system output in MultiWOZ 2.2, where `profile` represents the characteristics of the user, `user_request` represents the user's request, and `system_answer` represents the system's response from the original MultiWOZ 2.2.

You are a dialogue act classification expert. Given the context: '{context}', analyze the sentence: '{sentence}'. Given the following domains: [domain list] and act types: [act type list + explanation], select the most appropriate dialogue act from the following options: [choice1, choice2, choice3].
Respond only with the dialogue act.

Example You are a dialogue act classification expert. Analyze the sentence: "I am in search of a hotel situated on the northern part of the town, and I do not require parking facilities."
Given the following domains: Attraction, Hospital, Police, Hotel, Restaurant, Taxi, Train and act types: Inform: The system or user provides information ...
Select the most appropriate dialogue act from the following options: 'Hotel-Inform', 'Train-Greet', 'Attraction-OfferBook'.
Respond only with the dialogue act.*

Figure 5: Prompt used for dialogue act classification with corresponding example.

You are a dialogue act classification expert. Given the context: 'context', analyze the sentence: 'sentence'. Answer the question: 'question' by selecting one of the following choices: [choice1, choice2, choice3].
Respond only with the slot.

Example You are a dialogue act classification expert. Analyze the sentence: "I am in search of a hotel situated on the northern part of the town, and I do not require parking facilities."
Answer the question: "In which area does the user want the hotel located?" by selecting one of the following choices: 'West side', 'North', 'East'. *Respond only with the slot.*

Figure 6: Prompt used for slot-value classification with corresponding example.

B User Profile Examples

Generation	Gender	Communication Style
Boomer	Male	Polite, formal, and detailed. Prefers structured conversations.
Gen X	Female	Balanced between formal and informal with a focus on getting things done. Uses a mix of professional and casual language depending on the context.
Gen Y	Male	Informal. Uses emojis and abbreviations but remains professional when necessary. Open to collaboration and seeks feedback.
Gen Z	Female	Casual and concise. Prefers messaging or voice notes. Frequently uses emojis or slang, and is very adaptable with tech.

Table 7: Sample user profiles with communication styles

C Human Review

During the human review to verify content consistency between P-MultiWOZ and MultiWOZ, two volunteer annotators with a good proficiency in English, who participated for free, were recruited and tasked with annotating the 104 generations produced by Llama 3.1 8B and Mistral 7B v0.3. The taxonomy from Kasner and Dusek (2024) was used to classify errors into four categories: *Incorrect*, where the text contradicts the data; *Not Checkable*, where the information cannot be verified; *Misleading*, where the text is deceptive given the context or missing information; and *Other*, for problematic cases that do not fit into the other categories. The annotation process lasted $\sim$ 3 hours. Table 8 shows an annotation example. We then calculated Cohen's κ for both annotators across both models and all generations. The annotators achieved a high score of 0.84, indicating near-perfect agreement.

System Output: The number there is 01223304705 and it's on 8 Mercer's Row. Anything else?		
Model	**Model Paraphrase**	**Label**
$\text{Llama}_{\text{Boomer}}$	That's the phone number, 01223304705, and it's located at 8 Mercer's Row.	Misleading
$\text{Mistral}_{\text{Boomer}}$	The number you mentioned, 01223304705, is located at 8 Mercer's Row. Is there any further information you need regarding this?	Ok

Table 8: Example of annotation: $\text{Llama}_{\text{Boomer}}$ output is labeled as *Misleading* due to the omission of the final part of the system output, *"Is there anything else?"*.

D P-MultiWOZ Example

User: Can you help me find a train going to Cambridge leaving on Saturday?
System: What city will you be departing from?

Generation	Paraphrased Request	Paraphrased Response
Boomer	Would it be possible for you to assist in locating a train departing for Cambridge on Saturday?	In what city will the train departure for Cambridge on Saturday be originating from?
Gen X	Could you please assist in locating a train to Cambridge departing on Saturday?	What is the departure city for the train to Cambridge on Saturday?
Gen Y	Hey there, can you assist in locating a train heading to Cambridge with a departure on Saturday?	What city are you departing from, for finding a train heading to Cambridge on Saturday?
Gen Z	Hey there! Got any info on trains heading to Cambridge, departing on Saturday? 😊	What city are you departing from, for trains heading to Cambridge on Saturday? 😊

Table 9: Generational Personalization on a Train Booking Request

E GeMoSC Annotation Guidelines and Overview

Screenwriting follows well-documented principles, with guidelines that emphasize consistency between characters and the context in which they operate (McKee, 1997; Dancyger and Rush, 2012). The selection of films was carried out by choosing well-known titles and specific genres, in accordance with the following guidelines, to ensure that the characters reflect realistic linguistic patterns as closely as possible.

The following guidelines outline clear instructions for the GeMoSC annotation process, detailing how to handle specific cases to ensure the integrity and reliability of the annotations.

- **Human Characters Only**: Include only human characters. Exclude non-human characters (e.g., animals in animated films) and voice roles.
- **Use Actor's Age When Unspecified**: If the character's age is not mentioned, use the actor's age as a reference (it is reasonable to assume that a 40-year-old actor would speak in a way consistent with a 40-year-old character).
- **Contemporary or Relevant Settings**: Focus on films set in modern or easily relatable settings where characters can be clearly assigned to a specific generation. Avoid films set in historical periods or ambiguous timelines where it's hard to determine a character's generational identity.
- **Clear Character Identification**: Ensure the character's role and name are clearly defined and can be directly matched to an IMDb entry. For example, if the script lists "Attorney," make sure it is specified whether it refers to roles like "Assistant Attorney" or "General Attorney" to avoid ambiguity.
- **Single Age Representation**: Exclude characters portrayed at multiple ages within the film. If a character is shown both as a child and an adult, do not include them.

F Experimental Settings for Perplexity Score

We used the following hyperparameters for fine-tuning each model by using the LoRA adapter for calculating the perplexity scores:

- R = 64
- α = 16
- no bias
- dropout: 0.05
- Target modules: Q-projections, V-projections

We loaded the adapter in 4 bits and did not use double quantization.

We trained on an A100 GPU, with a per-device batch size of 2. Fine-tuning took $\sim$ 2 hours per model. We used gradient accumulation steps with a learning rate of 1e-5 (with a linear scheduler).

Generation	#Character (M/F)	#Dialogues
Boomer	124 (88/36)	4679
Gen X	226 (151/75)	10827
Gen Y	240 (120/120)	15436
Gen Z	62 (23/39)	2998

Table 10: Overview of the GeMoSC dataset statistics showing generational group, the distribution of male (M) and female (F) characters across different generations, and the number of dialogues

G Qualtrics Interface

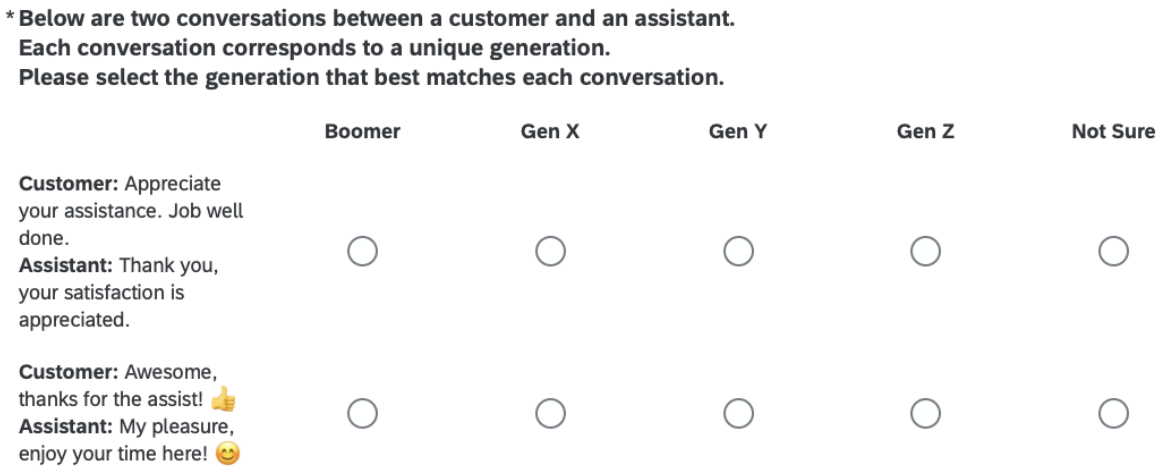

Figure 7: Example of question in Qualtrics