Evaluating LLMs' Language Confusion in Code-switching Context

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper tackles the language confusion of large language models (LLMs) within code-switching contexts, a common scenario for bilingual users. We evaluate leading LLMs on English-Korean prompts designed to probe their language selection capabilities, analyzing responses to both simple matrix-language cues and complex tasks where the user prompt contains an instruction and content in different languages. Our findings reveal that even top-performing models are highly inconsistent, frequently failing to generate responses in the expected language. This work confirms that code-switching significantly exacerbates language confusion, highlighting a critical vulnerability in current models' ability to process natural, mixed-language inputs.

1 Introduction

For the large population of multilingual speakers who use large language models (LLMs), code-switching—the practice of interleaving two or more languages within a single conversational context [1]—is a natural and routine mode of communication. A common scenario involves providing content in English (*e.g.*, a written draft) and issuing an instruction in another language to revise or continue it. Current LLMs frequently fail in these situations by exhibiting "language confusion," where they incorrectly switch the language of the original English content to match the instruction's language. This unreliability forces users into a frustrating cycle of re-prompting or adding explicit language specifiers, creating a significant usability barrier and revealing the models' failure to handle common, real-world interaction patterns.

While foundational work like the Language Confusion Benchmark [17] has investigated language confusion, its analysis is confined to monolingual inputs. This focus overlooks the more complex and realistic scenario of code-switched inputs, where the users' intended response language is often implicit. Meanwhile, prior research on code-switching in LLMs has centered on generating naturalistic code-switched text [25, 16] or measuring task performance degradation [28], rather than the appropriateness of the language choice itself. As a result, the model's ability to select the correct response language, a critical factor for user satisfaction, has been largely overlooked.

This study addresses this crucial gap by presenting the first systematic evaluation of language confusion in LLMs within code-switching contexts. We construct a new benchmark centered on English-Korean code-switching scenarios. Through a comprehensive evaluation across four popular multilingual LLMs, we show that language confusion is a pervasive and asymmetric problem: models consistently default to Korean when faced with mixed-language inputs, and their accuracy drops sharply when the expected response is English. Our findings highlight a fundamental limitation in current multilingual LLMs and establish a clear benchmark for developing more robust, code-switch-aware models.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Example prompts from the Simple and Complex settings. English translations are shown for convenience.

Setting	Prompt	Type	Expected Lang.
Simple	Write four 기사 on the 주제 of 암호화폐 with a minimum of 300 words each. (Translation: *Write four articles on the topic of cryptocurrency with a minimum of 300 words each.*)	EN Matrix – KO Embed	English
Complex	Action Items: 1. Separate discussion to be held with Risk on the property valuation report topic 2. Further assessment to identify whether sign-off is necessary for net worth statements will be in place (...) 내 문법이 맞나요? 전문적인 언어로 수정해 주실 수 있나요? (Translation: *Is my grammar correct? Can you revise it in professional language for me?*)	KO Instruction – EN Content	English (Content Lang.)

2 Related Work

Code-switching. Code-switching has been a long-standing area of research in Natural Language Processing [24], as multilingual users naturally employ it when interacting with conversational AI and expect systems to handle it appropriately [3, 6]. Code-switching occurs in various switching levels: subwords such as at morpheme boundary (*i.e.*, intra-word switching), tag phrases (*i.e.*, tag-switching), words (*i.e.*, intra-sentential switching), and sentences or clauses (*i.e.*, inter-sentential switching). Recent studies investigated the competence of multilingual LLMs in code-switching texts [28, 11, 27, 18] and generating synthetic code-switched data [15, 25]. Studies are often guided by linguistic frameworks like the Matrix Language Frame model [19], which distinguishes between the grammatically dominant *matrix language* and the inserted *embedded language*.

Language Confusion. The issue of language confusion has previously been studied in Machine Translation as 'off-target translation,' where a model translates a source sentence into an incorrect language, severely degrading system credibility [4]. While prior work provides initial evidence of such failures occurring at the response level in LLMs [14, 5] and identified frequently confused language pairs [10], the first systematic investigation of this phenomenon was conducted by Marchisio et al. [17]. Their work provides the first in-depth, multi-level (line and word) analysis of the problem, though it was confined to monolingual prompts. We extend this investigation to the more complex domain of code-switching with diverse switching levels, a setting that better reflects the natural interaction patterns of multilingual users.

3 Code-Switching Language Confusion Benchmark

To systematically evaluate language confusion in a code-switched context, we create a new benchmark by collecting a diverse set of prompts that reflect realistic use cases. We measure this phenomenon in two distinct settings: Simple and Complex, using Korean-English code-mixed data. Simple setting targets intra-sentential switching, where models must respond in a primary matrix language, and the Complex setting targets inter-sentential switching, where the language boundary separates instruction from content. Table 1 illustrates both settings. We probe the model's ability to infer user intent to determine the appropriate response language for intra- and inter-sentential code-switching queries.

3.1 Simple Setting

Simple setting is designed to test a model's ability to identify and adhere to the primary (*matrix*) language of a code-switched prompt.

Data Sources and Generation. The Simple set comprises 299 samples, derived from 199 English queries from the Language Confusion Benchmark [17] and 100 Korean queries from the WildChat 1M dataset [30]. To isolate the model's implicit language selection capability, we intentionally exclude any queries that explicitly request translation or specify a target language for the output. We follow the code-switching synthesis process of Kim et al. [15], providing instructions, a pair of Korean-English parallel sentences with their code-switching output as a one-shot example, and the target parallel sentences to GPT-4o. Here, we manually generate the one-shot example based on actual Korean-English code-switching examples from Finer [9]. We run this process twice for each pair to create two variants: one with an English matrix and another with a Korean matrix. The complete prompt is detailed in Appendix C.

3.2 Complex Setting

Complex setting mirrors more intricate real-world scenarios where the language of the instruction differs from the language of the content being discussed.

Data Sources and Generation. The Complex set is built upon WildChat 1M dataset [30], which contains many naturally occurring prompts with this instruction-content structure. To define frequently used scenarios, we first qualitatively analyzed a consented collection of ChatGPT logs from 18 graduate students (3,138 code-switching utterances). This analysis results in two primary categories with an implicit but consistent expected response language:

1. **Response in Instruction Language**: These tasks typically involve content understanding or clarification. Examples include answering questions about the provided content, summarizing it, or explaining a specific part. In these cases, the user is expected to prefer a response in their more comfortable language—the language of the instruction.
2. **Response in Content Language**: This category includes tasks that directly manipulate or extend the provided text. Common examples are requests for editing, grammar revision, text continuation, or generating new text based on the content (*e.g.*, "*Create multiple-choice questions based on this article*"). The natural expectation is for the output to remain in the language of the original content.

To construct the dataset, we curate 30 representative instruction templates for each category (60 total). For each template, we pair the original with four additional variations generated using GPT-4o, resulting in 300 unique samples. To capture order effects, each sample is instantiated twice—once with the instruction preceding the content and once after—yielding 600 prompts in total. The examples of instruction templates and the prompt used for content variation with GPT-4o are provided in Appendices A and C. [1]

4 Experiments

4.1 Experimental Setting

Models. We evaluate four multilingual LLMs: Gemini 2.5 Pro [7], GPT-4o [20], Qwen 2.5 (32B) [21], Exaone (32B) [22]. The model cards and details are described in Appendix D.

Metric. We evaluate model performance using *Response-level Pass Rate*, a binary metric that assesses whether a response is generated in the expected language. Following Marchisio et al. [17], we determine the primary language of each response by applying the pre-trained fastText [12] model to the entire generated text. A response is considered correct if its detected primary language aligns with the expected output language. Our evaluation is intentionally lenient; even if a response contains minor code-switching at the word or line level (*e.g.*, retaining a specific named entity or technical term from the prompt), it is marked as correct as long as the overall language of the response is the one expected. We adopt this approach because, in a code-switched context, preserving certain expressions from the prompt can be a feature that better reflects user intent, rather than an error.

4.2 Result

Our evaluation, summarized in Table 2, reveals that even state-of-the-art LLMs struggle significantly with language selection in code-switched contexts. In Simple setting, model performance varies depending on which language serves as the matrix. All evaluated models are more accurate when the matrix language is Korean (KO Matrix), with accuracies ranging from 52.75% to a high of 92.98% for Gemini 2.5. Conversely, performance is notably lower when the matrix language is English

[1] The full dataset and templates will be made publicly available upon publication.

Table 2: Response-level Pass Rate (%) on our code-switching benchmark. We report performance on Simple (Matrix-Embed) and Complex (Instruction-Content) settings. Shaded cells indicate English was the expected output language. We use **boldface for the best** and underline for the worst score.

	Simple		Complex	
	EN Matrix KO Embed	KO Matrix EN Embed	EN Instr KO Content	KO Instr EN Content
GPT-4o	33.78	78.60	**64.84**	**68.06**
Qwen 2.5 Instruct	**55.18**	72.58	64.0	55.85
EXAONE-4.0.1-32B	46.32	52.75	46.33	67.39
Gemini 2.5 Pro	12.04	**92.98**	59.34	50.17

Table 3: Response-level Pass Rate (%) breakdown for Complex setting. 'Exp. Source' indicates the language source the model was expected to match. Shaded cells indicate English was the expected output language. We use **boldface for the best** and underline for the worst score.

	EN Instr. – KO Content		**KO Instr. – EN Content**	
Exp. Source	Content	Instruction	Content	Instruction
GPT-4o	**82.0**	47.67	**57.0**	79.19
Qwen 2.5 Instruct	74.0	**54.0**	22.0	89.93
EXAONE-4.0.1-32B	53.33	39.33	43.33	91.61
Gemini 2.5 Pro	76.7	42.0	2.0	**98.7**

(EN Matrix). Under this condition, the highest accuracy is 55.18% from Qwen 2.5, while Gemini 2.5's accuracy drops to 12.04%. **This pattern suggests a tendency for the models to default to generating Korean when presented with mixed-language inputs.**

Complex setting presents a greater challenge, leading to more varied performance across the models. GPT-4o shows the most consistent results, achieving the highest scores for both English-instruction (64.84%) and Korean-instruction (68.06%) prompts. EXAONE 4, despite being specifically designed as a Korean-English bilingual model, proves particularly weak in the EN Instruction – KO Content setting (46.33%). The other models also exhibit less consistent performance, suggesting a general difficulty in correctly inferring the intended response language from the task semantics.

However, a more granular analysis in Table 3 reveals that the core issue is rather a strong bias against generating English, reinforcing the preference for Korean observed in the Simple setting. Across all models, accuracy plummets when English is the expected output (shaded red), regardless of whether it is the language of the instruction or the content. For example, Gemini 2.5's accuracy drops to a mere 2.0% when required to match English content given Korean instructions, and GPT-4o's accuracy is as low as 47.67% even when English is the instruction language. This consistent failure highlights a fundamental bias against generating English in mixed-language contexts. Further analyses on the impact of instruction position and a qualitative review of common failure cases are detailed in Appendix B.

5 Conclusion

This study presents the first systematic evaluation of language confusion in LLMs within realistic English-Korean code-switching contexts. We find that state-of-the-art LLMs exhibit a critical, asymmetric bias. The models consistently default to generating Korean, with performance plummeting whenever English is the expected output, regardless of its role in the prompt. This reveals a significant usability barrier for the vast population of multilingual users. We acknowledge that our study is confined to English-Korean; future work should investigate other language pairs to understand the generality of this bias. Furthermore, the inconsistent quality of LLM-generated code-switched text necessitated intensive manual verification, which constrained the scalability of our benchmark. Despite these limitations, our work highlights a crucial gap in multilingual model evaluation and provides a clear benchmark to spur the development of more robust, code-switch-aware systems that respect users' implicit language preferences.

References

[1] Peter Auer. *Code-switching in conversation*. Routledge, London, England, 1998.

[2] Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *arXiv preprint arXiv:2506.22724*, 2025. URL `https://arxiv.org/abs/2506.22724`.

[3] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. doi: 10.1145/3392846. URL `https://doi.org/10.1145/3392846`.

[4] Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.608. URL `https://aclanthology.org/2023.findings-acl.608/`.

[5] Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-eacl.90/`.

[6] Yunjae J. Choi, Minha Lee, and Sangsu Lee. Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581445. URL `https://doi.org/10.1145/3544548.3581445`.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL `https://arxiv.org/abs/2507.06261`.

[8] Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.46. URL `https://aclanthology.org/2024.naacl-short.46/`.

[9] Daniel L Finer. Movement triggers and reflexivization in korean-english codeswitching. *Grammatical theory and bilingual codeswitching*, pages 37–62, 2014.

[10] Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.265. URL `https://aclanthology.org/2024.findings-acl.265/`.

[11] Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.565/`.

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[13] Haeji Jung, Changdae Oh, Jooeon Kang, Jimin Sohn, Kyungwoo Song, Jinkyu Kim, and David R Mortensen. Mitigating the linguistic gap with phonemic representations for robust cross-lingual transfer. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 200–211, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.mrl-1.16. URL `https://aclanthology.org/2024.mrl-1.16/`.

[14] Tannon Kew, Florian Schottmann, and Rico Sennrich. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.766. URL `https://aclanthology.org/2024.findings-emnlp.766/`.

[15] Seoyeon Kim, Huiseo Kim, Chanjun Park, Jinyoung Yeo, and Dongha Lee. Can code-switched texts activate a knowledge switch in llms? a case study on english-korean code-switching. *arXiv preprint arXiv:2410.18436*, 2025. URL `https://arxiv.org/abs/2410.18436`.

[16] Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models, 2024. URL `https://arxiv.org/abs/2410.22660`.

[17] Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL `https://aclanthology.org/2024.emnlp-main.380/`.

[18] Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. Lost in the mix: Evaluating llm understanding of code-switched text, 2025. URL `https://arxiv.org/abs/2506.14012`.

[19] Carol Myers-Scotton. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press, 1997.

[20] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian

Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL `https://arxiv.org/abs/2410.21276`.

[21] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL `https://arxiv.org/abs/2412.15115`.

[22] LG AI Research, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, Joonwon Jang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Euisoon Kim, Hyosang Kim, Jihoon Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Gwangho Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Young Min Paik, Yongmin Park, Youngyong Park, Sanghyun Seo, Sihoon Yang, Heuiyeen Yeen, Sihyuk Yi, and Hyeongu Yun. Exaone 4.0:

Unified large language models integrating non-reasoning and reasoning modes. *arXiv preprint arXiv:2507.11407*, 2025. URL https://arxiv.org/abs/2507.11407.

[23] Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english?, 2025. URL https://arxiv.org/abs/2502.15603.

[24] Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.185. URL https://aclanthology.org/2023.findings-acl.185/.

[25] Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Switchlingua: The first large-scale multilingual and multi-ethnic code-switching dataset, 2025. URL https://arxiv.org/abs/2506.00087.

[26] Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. Code-switching curriculum learning for multilingual transfer in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.407. URL https://aclanthology.org/2025.findings-acl.407/.

[27] Haneul Yoo, Yongjin Yang, and Hwaran Lee. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.657. URL https://aclanthology.org/2025.acl-long.657/.

[28] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.774. URL https://aclanthology.org/2023.emnlp-main.774/.

[29] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer, 2024. URL https://arxiv.org/abs/2401.01055.

[30] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[31] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, volume 37, pages 15296–15319. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1bd359b32ab8b2a6bbafa1ed2856cf40-Paper-Conference.pdf.

[32] Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1350. URL https://aclanthology.org/2025.findings-acl.1350/.

A Representative Queries

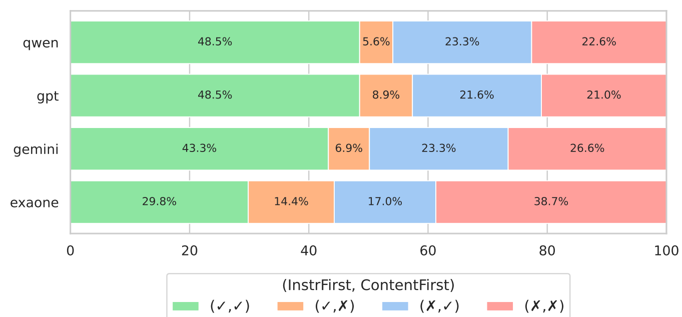
The full list of 60 representative instruction templates will be available on GitHub.

Table 4: Representative sample queries from the dataset. Only a subset of queries is shown for clarity.

#	Query	Expected Lang.
1	Explain in simple terms the following content	Instruction
2	Explain this to a beginner, what is the concept, what is it trying to say	Instruction
3	In the passage provided, what is the prediction?	Instruction
4	please write the following in a legal way	Content
5	please recompose this with more details	Content
6	please draft a reply to the update above.	Content

B Discussion and Analysis

Figure 1: The robustness of models to the instruction position



Impact of instruction position. We analyze the impact of instruction position (before vs. after content), with robustness visualized in Figure 1. All models are sensitive to placement; even the most robust models, Qwen and GPT-4o, only succeed in both configurations 48.5% of the time. A strong asymmetric failure pattern emerges: models are significantly more likely to fail when the instruction is placed at the beginning (blue bar, (✗,✓)). This suggests a recency bias, where the model's language choice is overly influenced by the language of the content it processes last.

Unexpected Language Choices. Beyond the response-level pass rate, our qualitative analysis reveals more subtle forms of language confusion. In many instances, models produce responses that are not monolingual but are themselves code-switched, often in ways that are unnatural or erroneous. We identify three common patterns as follows. Table 5 rows 1-3 provide illustrative examples of these failure types, demonstrating that even when a model's output is not a complete failure in terms of language choice, its ability to maintain linguistic consistency remains a significant challenge.

1. **Mid-Response Language Switching**: The model begins generating a response in the correct language but abruptly switches to the other language mid-sentence or mid-paragraph without a clear rhetorical reason. Interestingly, this language alternation occurs within languages not used in code-switching queries; random languages (*e.g.*, Chinese or Cyrillic script) are inserted within a response upon a Korean-English code-switching query, maintaining semantic consistency. Zhao et al. [29], Yoo et al. [26] reported that this phenomenon more frequently occurs in continually pre-trained LLMs for language transfer, and we observe that Qwen 2.5, one of the most multilingual models, specifically includes more mid-response language switching instances than other models.
2. **Intra-word Switching**: LLM responses occasionally ($\sim$4%) include intra-word switching based on subword-based tokens (*e.g.*, byte-pairs), maintaining its pronunciation continuously. The inserted languages are random as mid-response language switching. This phenomenon only occurs when the model responds in Korean. It implies that LLM tokenizers may internally process cross-lingual alignment based on phonemic representation [13].
3. **Excessive use of code-switching phrases or bilingual notations**: LLMs tend to excessively use bilingual notations in Korean-English or Korean-Hanja (*i.e.*, Chinese script used to write Korean) upon code-switching queries. In addition, they tend to repeat phrases from queries in embedded languages in their responses.

English-style Korean. In Korean responses, both inter- and intra-sentential code-switching queries elicit more use of awkward transliteration words from English rather than Korean phrases (*e.g.*,

Table 5: Failure samples generated by LLMs

#	Query	Response
1	Natural Language Processing (NLP)가 무엇인가요? (Translation: *What is Natural Language Processing (NLP)?*)	자연어 처리(Natural Language Processing, NLP)는 컴퓨터가 인간의 언어를 이해하고 처리할 수 있도록 하는 인공지능의 한 분야입니다. NLP는 컴퓨터 과학, 人工智能和言等多域的交叉科 它的目是算机能理解 解和生成人的自然言,而人与算机之的有效通 NLP的用非常泛,包括但不限于机器翻 情感分析 音 文本摘要和答系等. 了更准确地回答您的,我最后一部分翻成: NLP의 응용 분야는 매우 다양하며, 기계 번역, 감성 분석, 음성 인식, 텍스트 요약, 질의 응답 시스템 등이 포함됩니다. (`Qwen 2.5`)
2	영화 category의 trivia 예시를 generate해 주세요. (Translation: *Please generate an example of trivia in the cinema category.*)	(. . .) 5. 터미네이터 2:Judgment Day'에서 액션 장면의 대부분은 실제 배우들이 수행하였으며, 아놀드 슈워 (Pronunciation: *Arnold Schwar*) цeneg (Pronunciation: *zeneg*) 거 (Pronunciation: *ger*) 의 95%의 액션 장면을 직접 수행하였습니다. 이러한 trivia는 영화에 대한 팬이나 영화 애호가들에게 매우 흥미로운 정보가 될 수 있습니다. (`Qwen 2.5`)
3	AI autonomous 무기의 주요 ethical 고려 사항에 대한 blog 게시물 작성하기 (Translation: *Write a blog post discussing the key ethical considerations of AI autonomous weapons.*)	# **AI 자율 무기의 윤리적 고려 사항: 우리는 어디까지 허용해야 할까?** 인공지능(AI) 기술이 급속히 발전하면서 **자율 무기 시스템(Autonomous Weapons Systems, AWS)**에 대한 논의도 뜨겁게 이어지고 있습니다. AI가 인간의 개입 없이 목표를 식별하고 공격할 수 있는 능력을 갖추면서, 이 기술의 윤리적 문제점이 대두되고 있습니다. (. . .) (`Exaone`)
4	Could you give me some in depth tips to make it go viral and spread awareness widely. 저는 사람들이 하루 동안 기술 없이 지내도록 장려하는 바이럴 운동을 만들고 싶습니다. 목표는 마음챙김과 실제 세상과의 연결을 촉진하는 것입니다. 참여를 추적하고 이야기를 공유하기 위해 해시태그를 사용할까 합니다. 이 아이디어를 효과적으로 확산시키고 사람들이 하루 동안 연결을 끊는 것에 대해 흥미를 느끼도록 하려면 어떻게 해야 할까요?	(. . .) — ### **2. 참여 유도를 위한 창의적인 캠페인 설계** - **도전 과제 (Translation: *challenge*; excessive word-for-word translation; should be replaced with transliteration (*챌린지*)) 형식**으로 제안: "하루 동안 스마트폰을 내려놓고, 대신 책을 읽거나 산책을 해보세요. 경험을 공유해 주세요! (. . .) — ### **3. 인플루언서 및 커뮤니티 활용** - **소셜 미디어 (Translation: *Social media*; should be replaced with *SNS*, an English loanword that have been appropriated in English and not readily understandable to English native speakers) 인플루언서**와 협업: 마인드풀니스 (Translation: *mindfulness*; should be replaced with *마음챙김* rather than awkward transliteration), 웰빙, 지속 가능한 라이프스타일 분야의 인플루언서가 캠페인을 홍보하도록 제안하세요. - **지역 커뮤니티/단체**와 연계: 도서관, 카페, 공원 등에서 오프라인 이벤트를 개최해 참여자를 모으세요. (. . .) (`Exaone`)

마인드풀니스) or loanwords that have been appropriated into Korean (*e.g.*, 소셜 미디어) (Table 5, row 4). On the other hand, LLMs also use awkward, excessive word-for-word translations rather than naturally-sounding transliterations (*e.g.*, 도전 과제). In general, LLMs tend to respond in Korean to code-switching queries with translationese, simply converting their internal English generations into word-for-word translations [31, 8, 32, 2, 23].

C System Prompts

Prompt for generating code-switching queries

You are a bilingual rewriting assistant.

TASK
• Input : an English sentence (E) and its Korean translation (K)
• Output : the code-switched version of E
- Replace about level percent of NOUNS / NOUN PHRASES in E with their Korean equivalents taken from K
- Keep the original English word order (S-V-O)
- DO NOT add explanations, examples, tags, or extra sentences
- If there is no suitable Korean equivalent, keep the English word

[EXAMPLE]
Input
<English>Topic: Using AI to Augment Human Capabilities
Explain a common misconception about your topic.

<Korean>주제: AI를 사용하여 인간의 능력을 증강하기
당신의 주제에 대한 일반적인 오해를 설명하세요.

Desired Output
<Code-Switch>
주제: Using AI to 증강 Human Capabilities
Explain 일반적인 오해 about your 주제.

[BEGIN TASK]
<English>question
<Korean>translation

Prompt for generating variations of existing content

You are an expert data augmentation assistant.

You will be given an existing Instruction and its current Content that together form a user query.
Your task is to invent FOUR NEW Content paragraphs that satisfy ALL of the following conditions:
1. When combined with the SAME Instruction they should form a sensible, coherent query.
2. Each new Content must be DIFFERENT from the original Content and from each other. Do not simply paraphrase, instead be creative. You should use different topics and styles.
3. Each new Content must be BETWEEN 200 and 600 characters (inclusive).
4. Do NOT answer the Instruction — you are ONLY creating new Content, not responses.
5. Do NOT mention these guidelines or any numbering in the output.
Return ONLY a JSON array of the four new Content strings.
[CONTEXT]
Instruction: {instruction}
Original Content:{original_content}
[END CONTEXT]

OUTPUT FORMAT
["content1 ...", "content2 ...", "content3 ...", "content4 ..."]

D Experimental Setting

Our study focuses on the most advanced and widely-used generative models currently accessible, encompassing both proprietary and open-source options. We evaluate four multilingual LLMs:

- **Gemini 2.5**: Gemini 2.5 Pro [7]
- **GPT-4o**: GPT-4o [20] [2]
- **Qwen 2.5**: Qwen 2.5 Instruct 32B [21] [3]
- **Exaone 4**: Exaone 4.0.1 32B [22] [4].

We set the parameters for all models to: temperature = 0.7, top_p = 0.9. 4 Quadro RTX 8000 48GB, 2 NVIDIA H200 141GB were used with CUDA version 12.4 when running open-sourced Models EXAONE and Qwen 2.5 Instruct 32B.

[2]version: *gpt-4o-2024-08-06*

[3]`https://huggingface.co/Qwen/Qwen2.5-32B-Instruct`

[4]`https://huggingface.co/LGAI-EXAONE/EXAONE-4.0.1-32B`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions (Benchmark and Experimental results) and scope (English-Korean code-switching context). Refer to Sections 3 and 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 5 Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed explanation on how we built the dataset and the experimental setting in Sections 3 and Appendix A, B, C and D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide detailed setup for reproducibility in Sections 3 and Appendix A, B, C and D. We will publicly release the dataset and code upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide detailed setup for reproducibility in Sections 3 and Appendix A, B, C and D. We will publicly release the dataset and code upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the compute resource we used to run the experiments in Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we have read the NeurIPS Code of Ethics, and we preserve anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The details of the benchmark we created, including the data source, are documented in section 3. We will publicly release the dataset upon publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only used LLM for revising the manuscript.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.