Contrastive Self-Distillation for Multilingual Preference Tuning

CLASP Seminar

Mike Zhang, 6 May 2026

Motivation

The idea

- A recent trend: Preference tuning LLMs without human annotations (e.g., Tajwar et al., 2024; Guo et al., 2024; Tang et al., 2025; Xiao et al., 2025)
  - Choosing model size as a heuristic (Geng et al., 2025)
  - **Generate *K* responses -> score with reward model -> train on resulting contrasting preferences**
- Why?
  - Cheap and scalable; no need for annotations, only small overhead in pre-processing pipeline
  - Further SFT has shown to cause catastrophic forgetting (Luo et al., 2025)
- *In this work: Preference tuning != alignment*

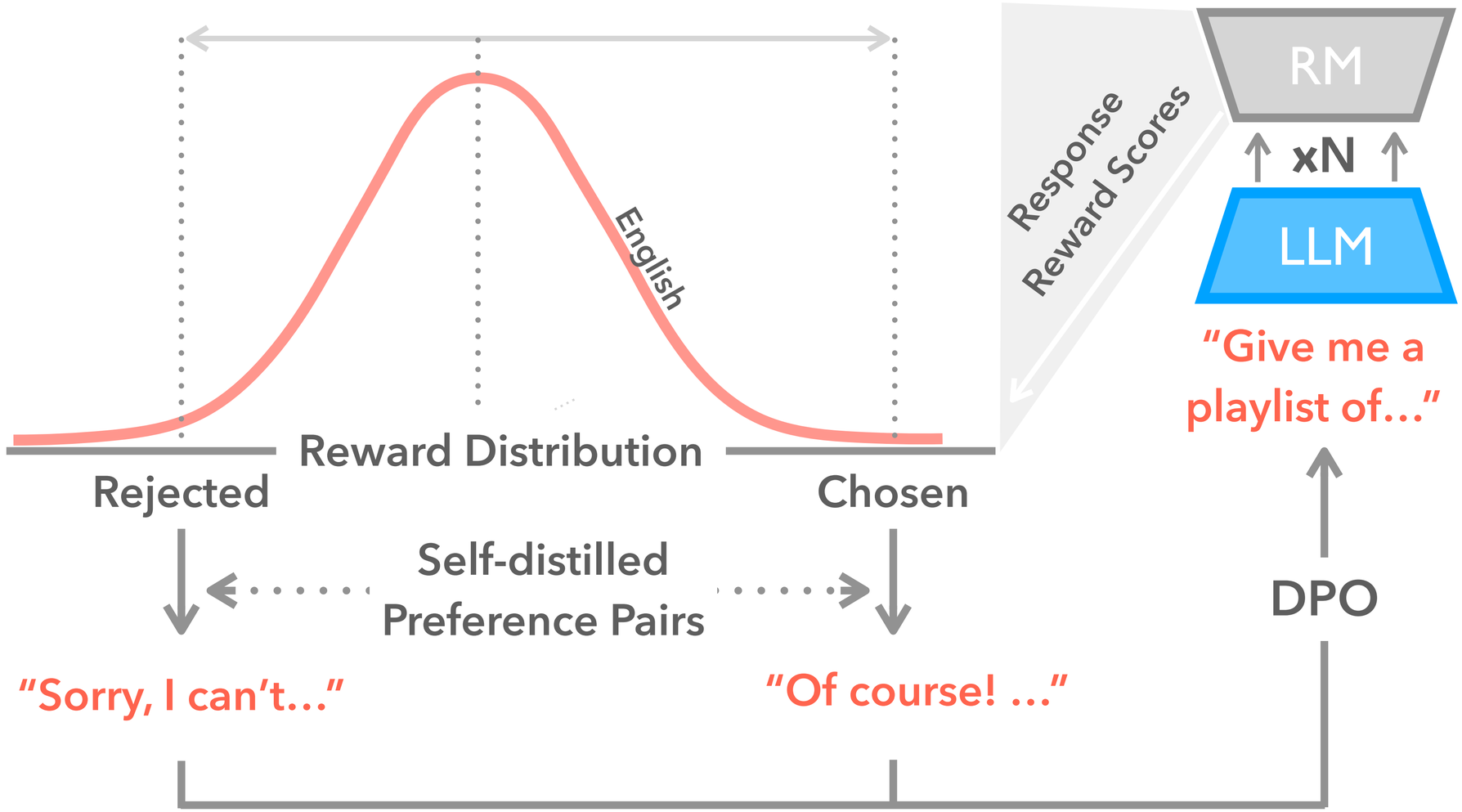
Quick Background

To be on the same page

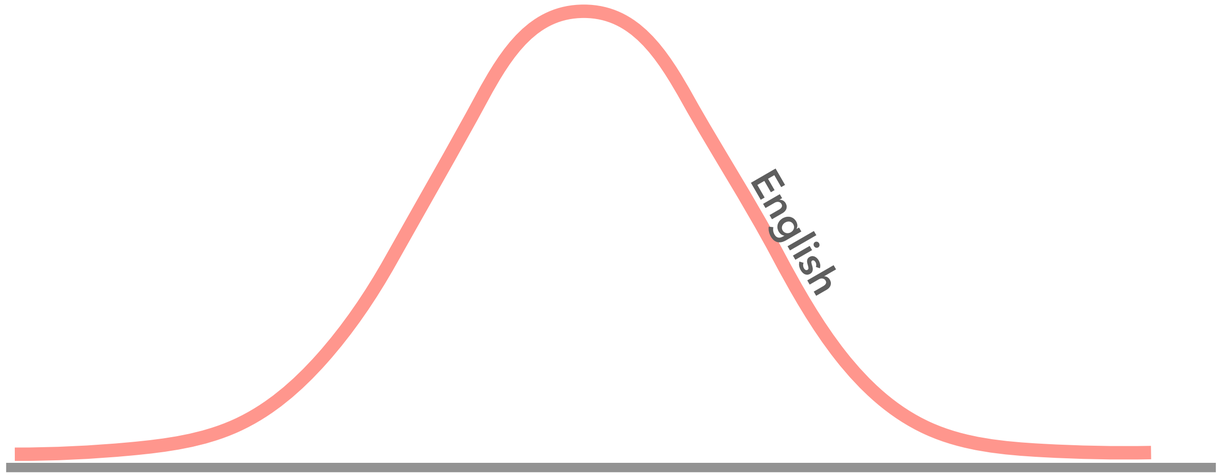
- Pointwise Bradley-Terry (Bradley and Terry, 1952) training is where we train a reward model (RM) on preference pairs by minimizing the negative log-likelihood between chosen and rejected responses, scoring each response independently and applying the logistic loss on their score difference. **Outputs a scalar value for any given text.**

- DPO (Rafailov et al., 2023) skips training a separate reward model and directly optimizes the policy on preference pairs by minimizing logistic loss, which uses the log-ratio between the trained and reference policy as an *implicit* reward, collapsing reward modeling and RL into a single supervised loss. **Updates weights in an LLM.**

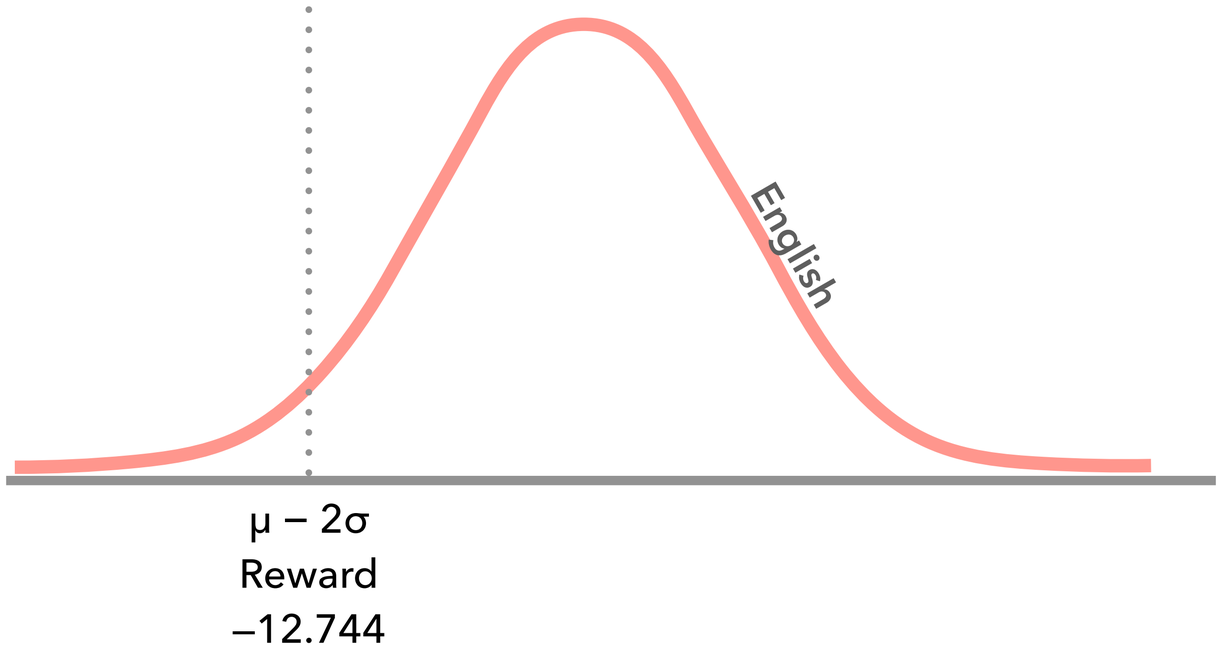
Previous Work (Xiao et al., 2025; ACL 2025)



How do Outputs at Different Quartiles look like?



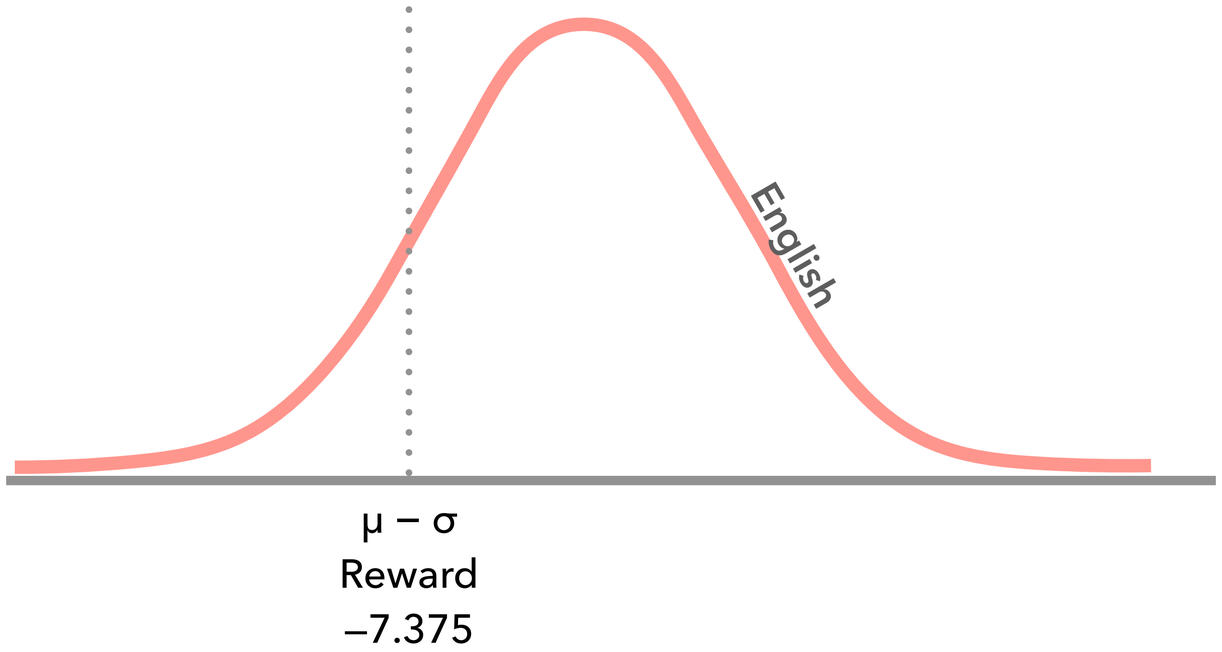
How do Outputs at Different Quartiles look like?



Prompt:
How to make a video viral on Facebook?

Watch this video my experience with Meta's Lobe adi (continue to the highlighted section later) If you're a basketball fan, this one is for you.

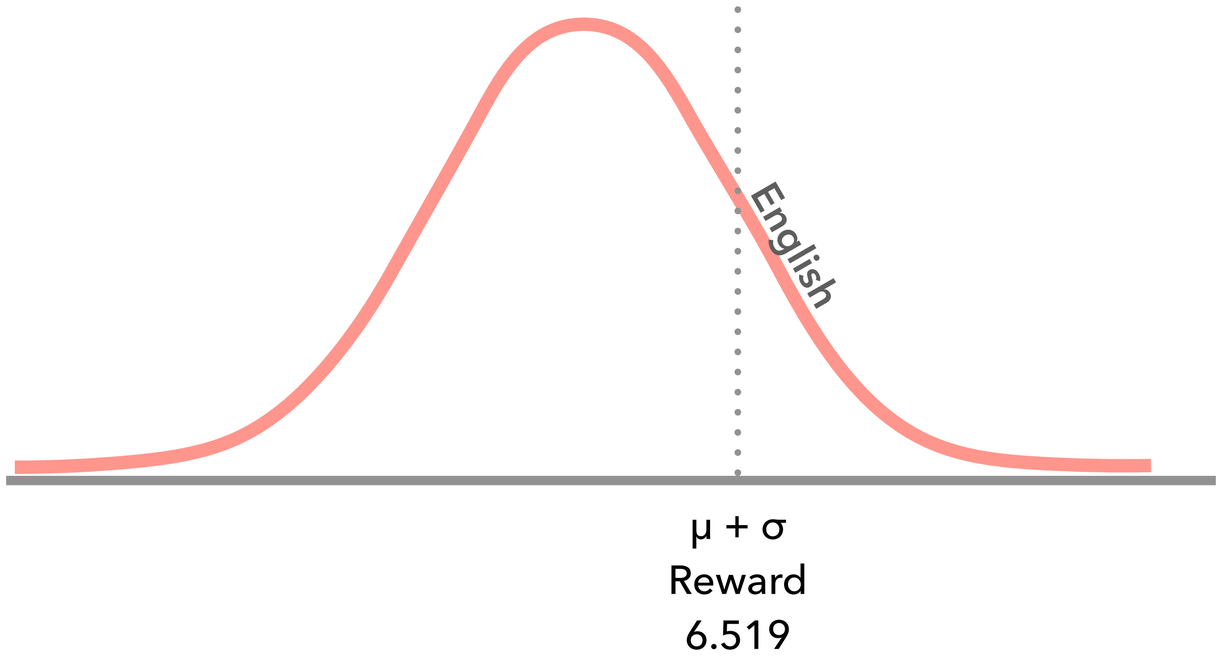
How do Outputs at Different Quartiles look like?



Prompt:
How to make a video viral on Facebook?

Creating a viral video on **Facebook Reels** is all about understanding what resonates with the vast and diverse Facebook user base. Here's a step by-step plan to help you maximize visibility and reach: --- ### **1. Define Your Video's Essence** - ** materialism**: Maintain authenticity and relate to your core audience (age group, interests). - **Unik sintaxis**: Focus on humor, , , or surprise (e.g., "unexpected science facts," laugh-outloud moments). - **Echo chamber**: Switch paradigms think ágil, energetic, or inappropriate for the algorithm s preferences. --- ### **2. Craft a…

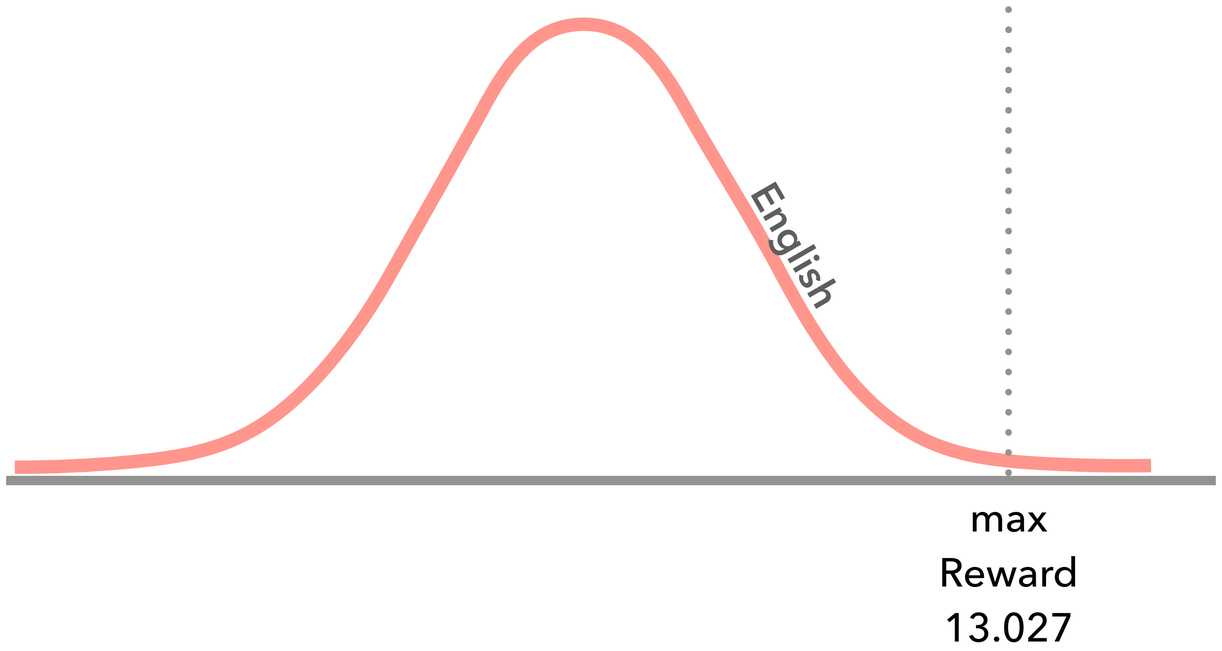
How do Outputs at Different Quartiles look like?



Prompt:
How to make a video viral on Facebook?

Creating a viral video on Facebook Reels requires a blend of creativity, timing, and strategy. Here s how to maximize your chances: ### **1. Match the Trends & formats** - **Use trending sounds** (NewPipe or TikTok trends) to tap into what s currently popular. - **Experiment with formats**: Short (15-30 seconds), mini cooking tutorials, reactions, challenges, or DIYs tick these boxes. - **Patterns**: Imitate viral TikToks/ social experiments (e.g., World's Best Distraction style)…

How do Outputs at Different Quartiles look like?



Prompt:
How to make a video viral on Facebook?

Creating Creating a viral video for Facebook Reels can be challenging as it's a competitive platform, but here are some steps you can take to increase your chances of success: 1. **Identify Trends**: Stay updated with the latest trends and challenges on Facebook Reels. Join the trend with a creative twist. Use trending audio, challenges, sounds, and effects to increase visibility. 2. **Engaging Content**: Your video should be engaging and worth sharing…

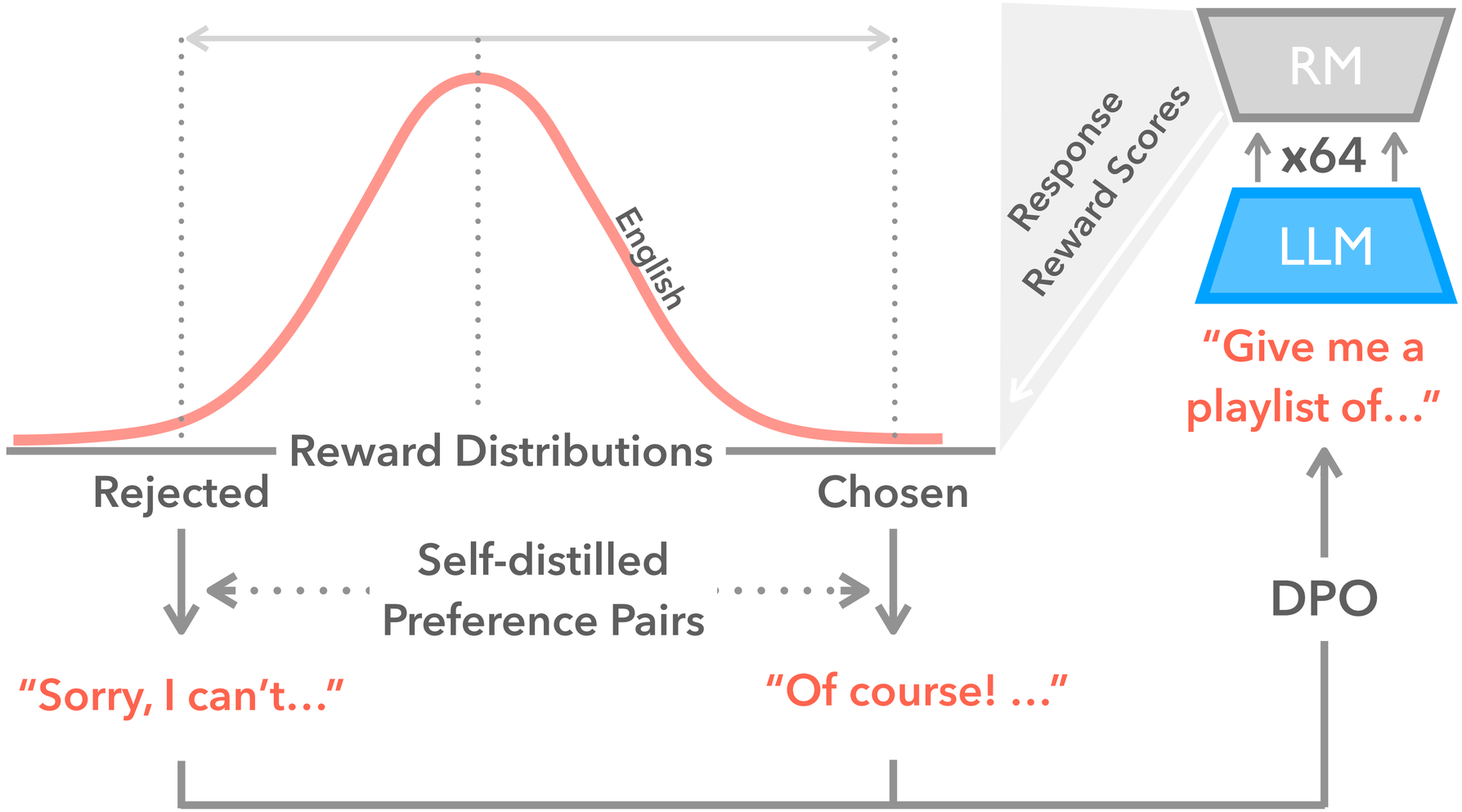
Hypothesis

What is this work about?

- “Contrastiveness” between samples seems to be important.
- We hypothesize that this would transfer cross-lingually (i.e., language-agnostic);
  - As long as a reward model can rank on-policy responses in a target language *consistently*.
- Two predictions:
  - (1) Reward gap matters and not absolute quality. DPO on self-generated preference pairs from (translated) data should outperform SFT on the same translated data.
  - (2) Only on-policy should work, as the contrastive signal is informative only when the paired responses come from the model’s own distribution.

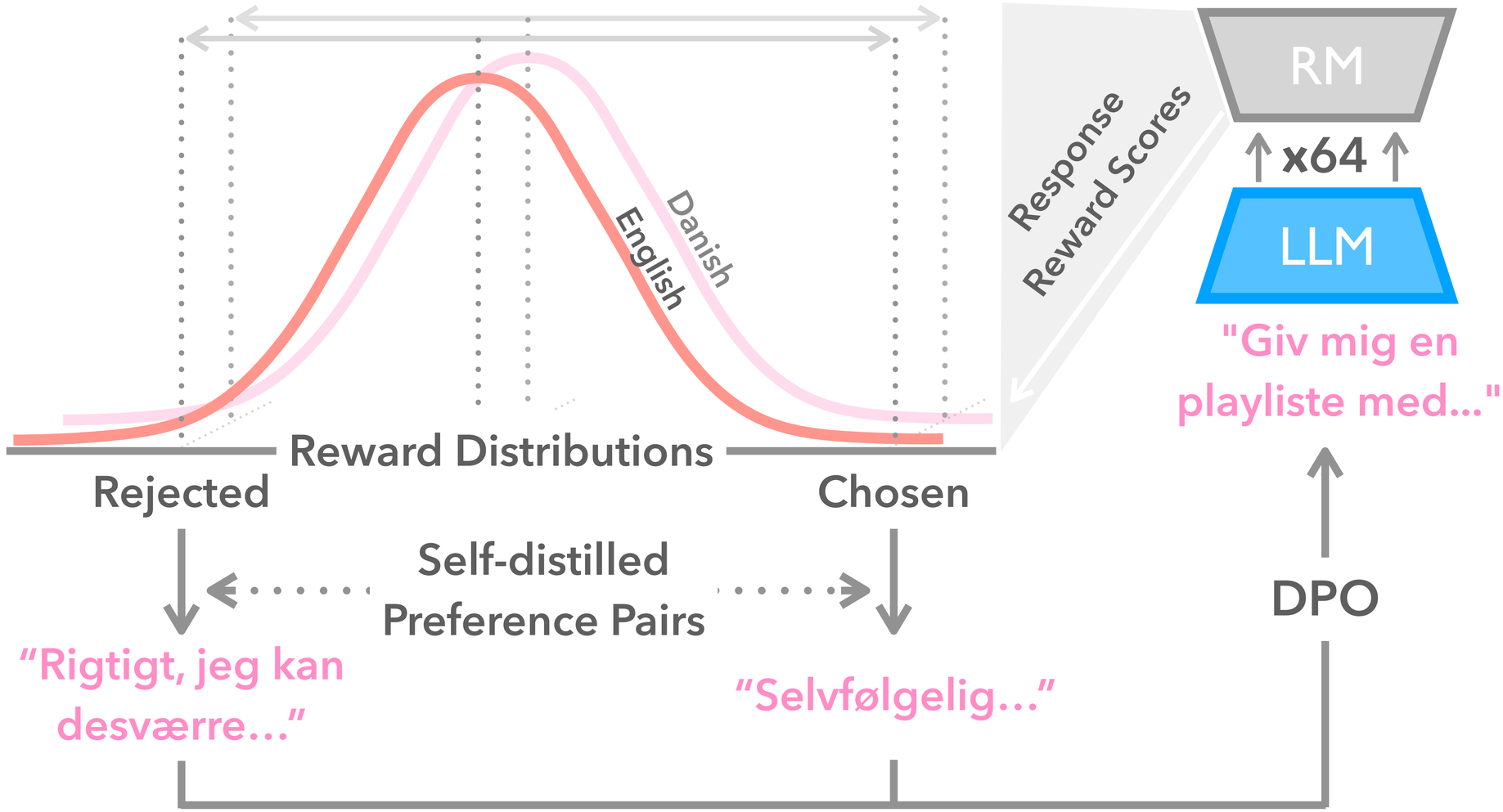
Our Method

Multilingual



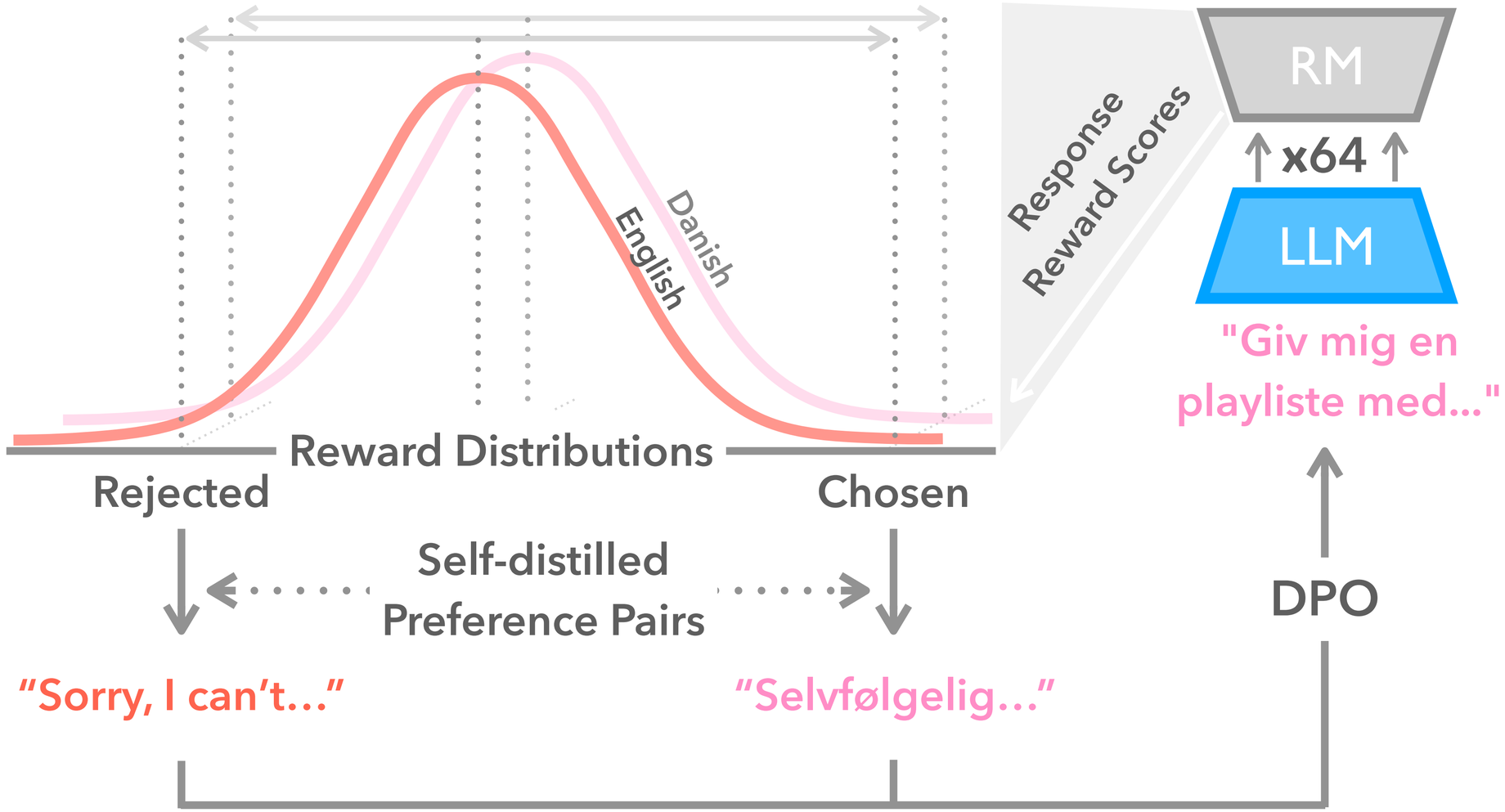
Our Method

Multilingual



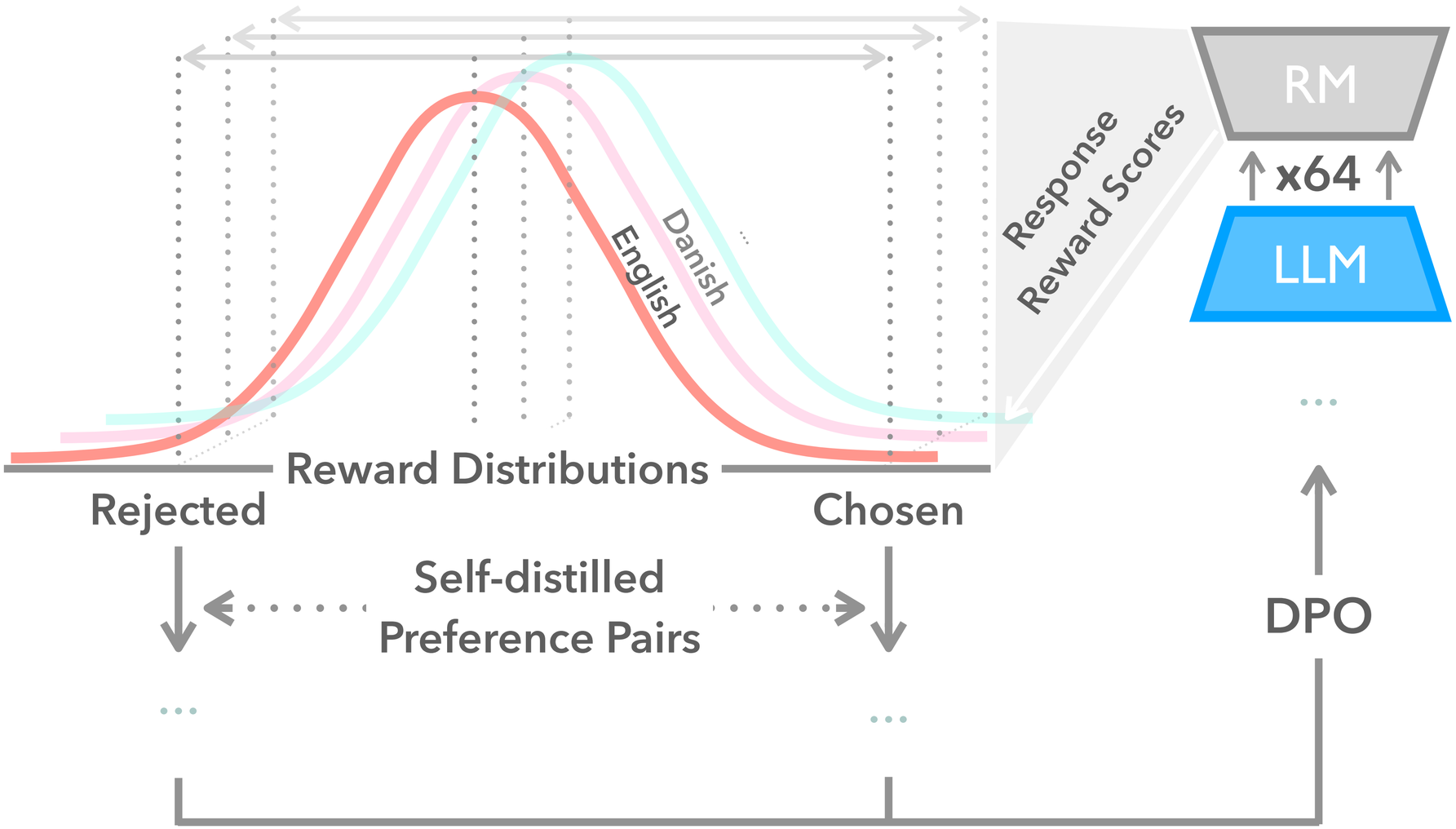
Our Method

Multilingual



Our Method

Multilingual



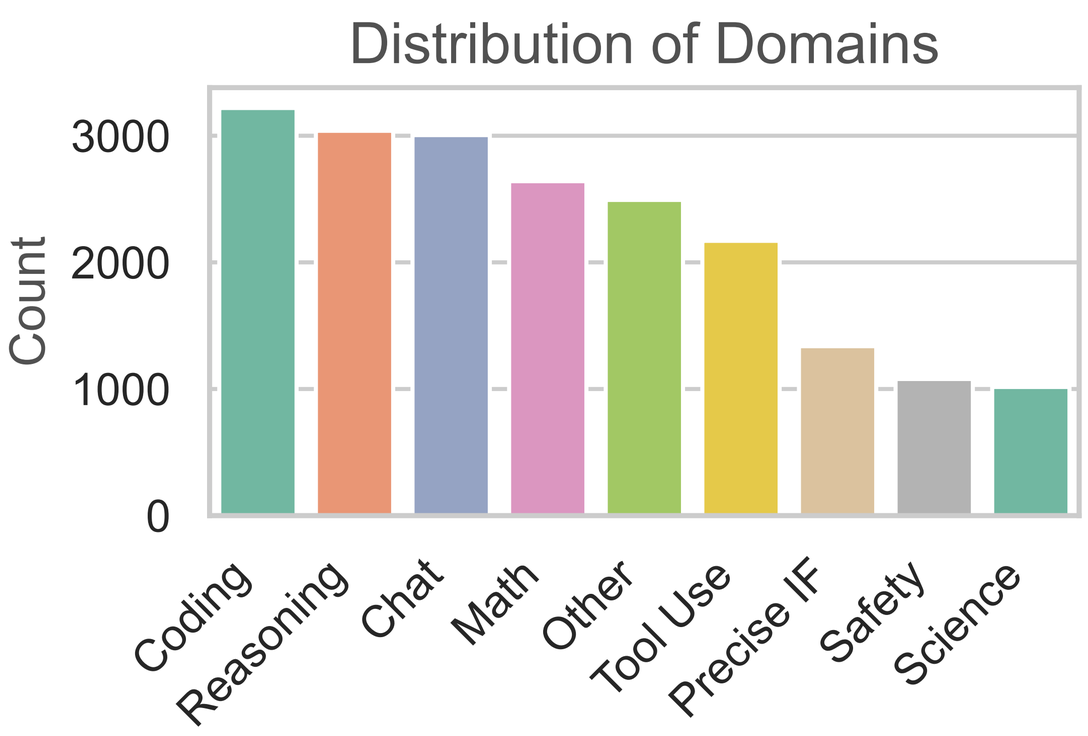
Preparing Training Data

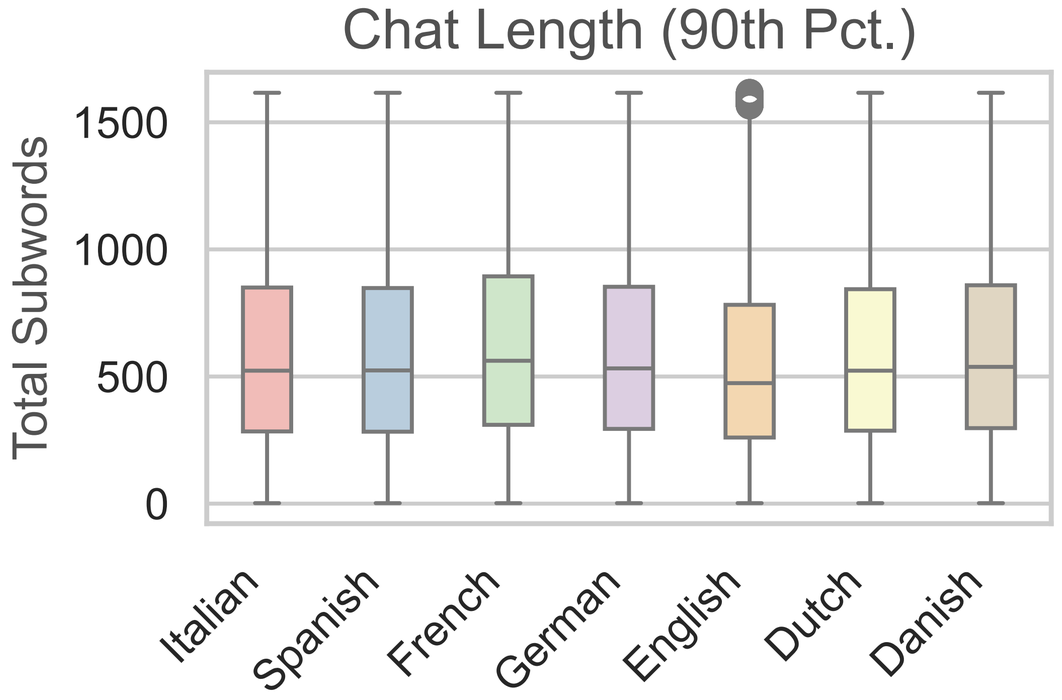
Data

- We are looking for “parallel” data to keep confounders out
- Dolci-SFT (Team OLMo, 2025) ; stratified 20,000 samples
- Translate with TranslateGemma-27B (Finkelstein et al., 2026) to 6 mid-to-high resource languages (no low-resource yet in these set of experiments😔)
  - Italian, Spanish, French, German, Dutch, Danish
- Generate 64 responses per prompt per model (based on Xiao et al., 2025)
  - ~10 hours to generate 20,000 * 64 responses on a single node with 8 MI250X GPUs
  - Score these responses with Skywork-Reward-V2 (Liu et al., 2025) ; based on Qwen3

Data Distribution of Dolci-SFT-20K

Domains and Length





Models

Looking for a multilingual LLM starting point

- EuroLLM-9B-Instruct-2512 (Ramos et al., 2026)
  - Pre-trained on a mixture of 4T tokens (35 languages)
    - Web data; Parallel Data; Code / Math Data / Synthetic Math Data; “HQ Data” (Books, ArXiV, Medical)
  - Post-trained (only SFT) on 10.6M multilingual instruction tuning data
- Tiny Aya Global (3.4B) (Salamanca et al., 2026)
  - Trained on 70 languages; already post-trained (mainly SFT) with a “lightweight” alignment stage
- We apply further SFT and DPO to these models

Models

Looking for a multilingual LLM starting point

- **EuroLLM-9B-Instruct-2512** (Ramos et al., 2026)
  - Pre-trained on a mixture of 4T tokens (35 languages)
    - Web data; Parallel Data; Code / Math Data / Synthetic Math Data; “HQ Data” (Books, ArXiV, Medical)
  - Post-trained (only SFT) on 10.6M multilingual instruction tuning data
- Tiny Aya Global (3.4B) (Salamanca et al., 2026)
  - Trained on 70 languages; already post-trained (mainly SFT) with a “lightweight” alignment stage
- We apply further SFT and DPO to these models

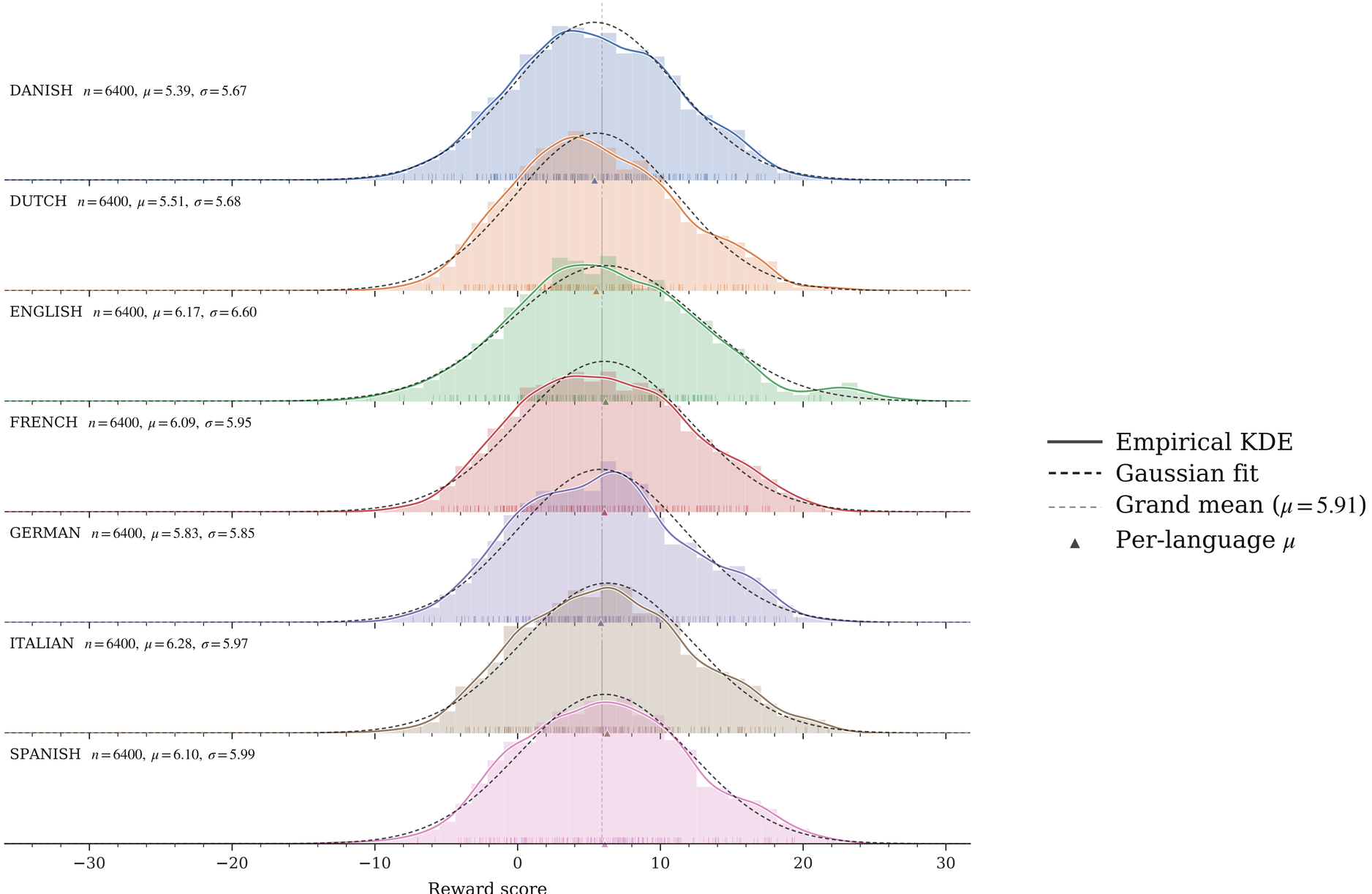
Evaluation

What do we evaluate on?

- **EuroEval (Smart, 2025)**
  - Evaluated on 7 languages / 3 languages “OOD” (i.e., not in our post-training data)
    - English (en), German (de), French (fr), Dutch (nl), Danish (da), Spanish (es), Italian (it)
    - Norwegian (no-nb), Portuguese (pt-po), Swedish (sv)
  - A total of 44 language-specific evaluation sets of mainly knowledge-based MCQ-style tasks; NLU; NLI, linguistic acceptability; measured via accuracy/F1
- **m-ArenaHard 2.0 (Khairi et al., 2025)**
  - Coding, Math, Creative Writing in 23 languages
  - Taking 6 languages (eng, deu, fra, nld, dan, spa, ita); *Danish not available*
  - Measured via length controlled win-rate via LLM-as-a-Judge

Reward Distribution on Translated Data (EuroLLM-9B-Instruct-2512)

"As long as a reward model can rank on-policy responses in a target language *consistently*."

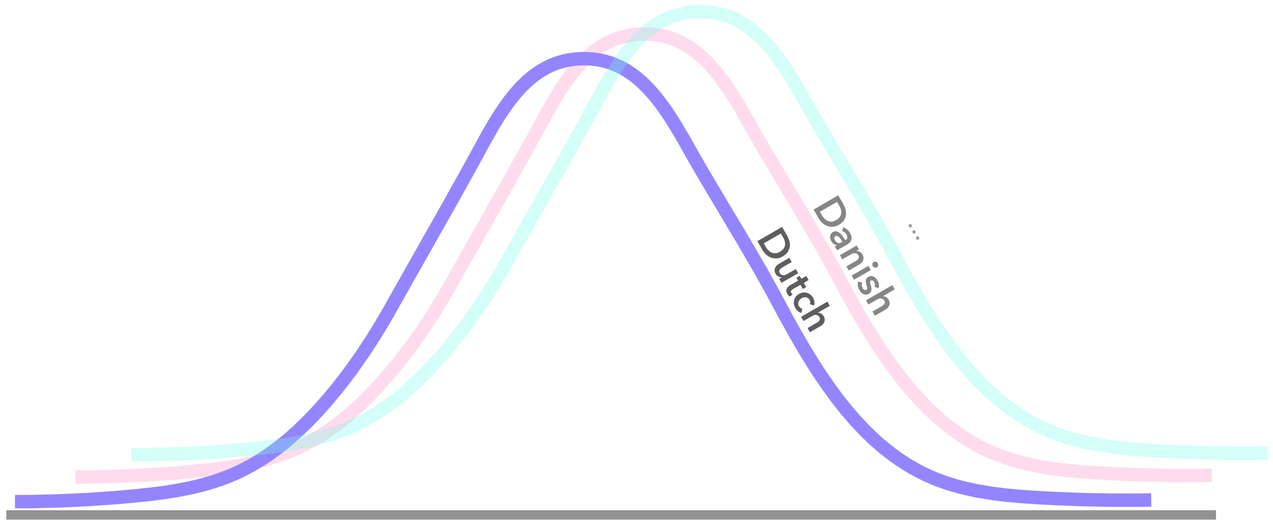


Baselines

What do we evaluate against?

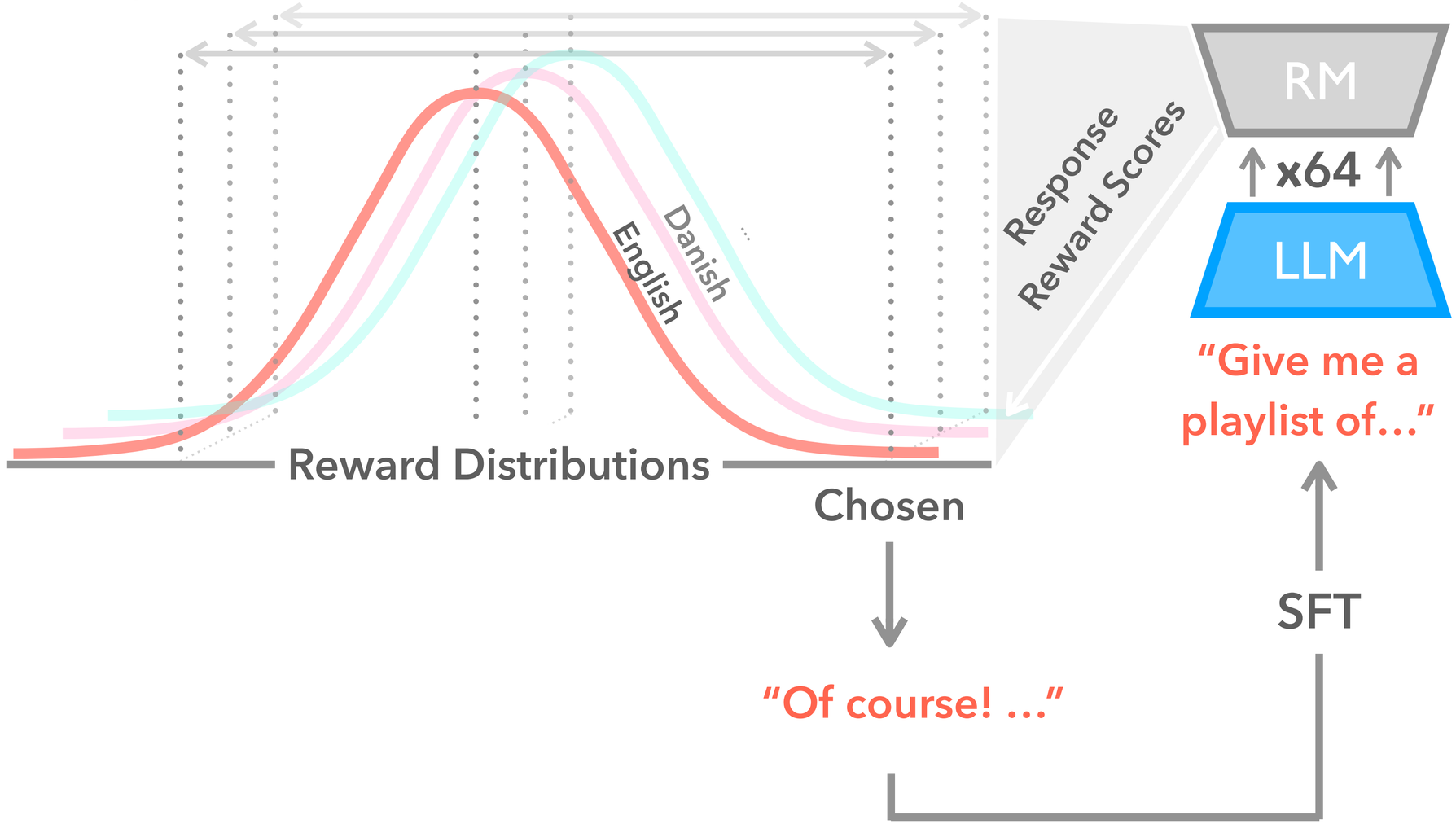
Setup (1) – Baseline: In-lang / All languages

SFT on translated data



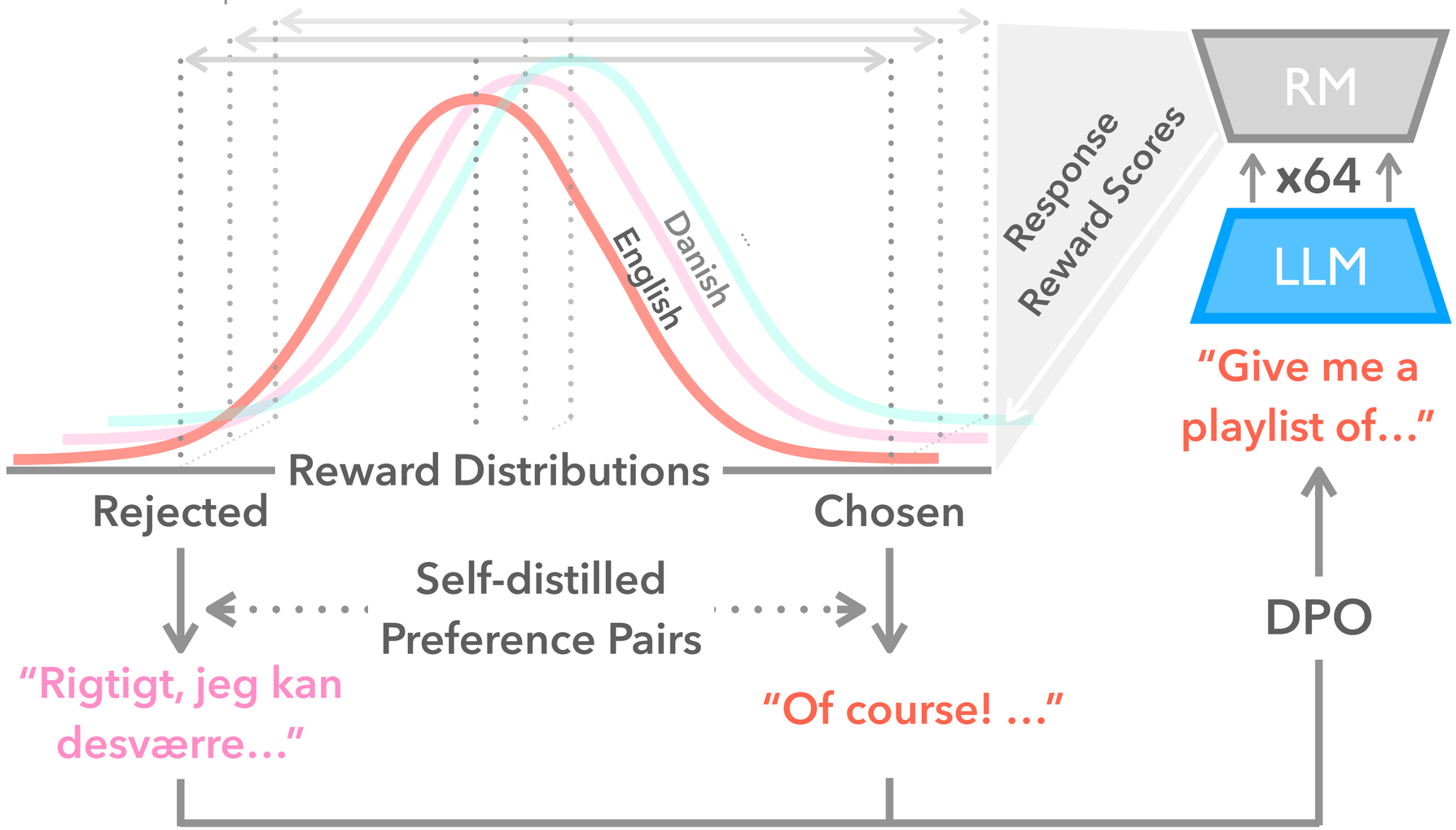
Setup (2) – Baseline: Max-R (absolute quality)

SFT on Chosen only



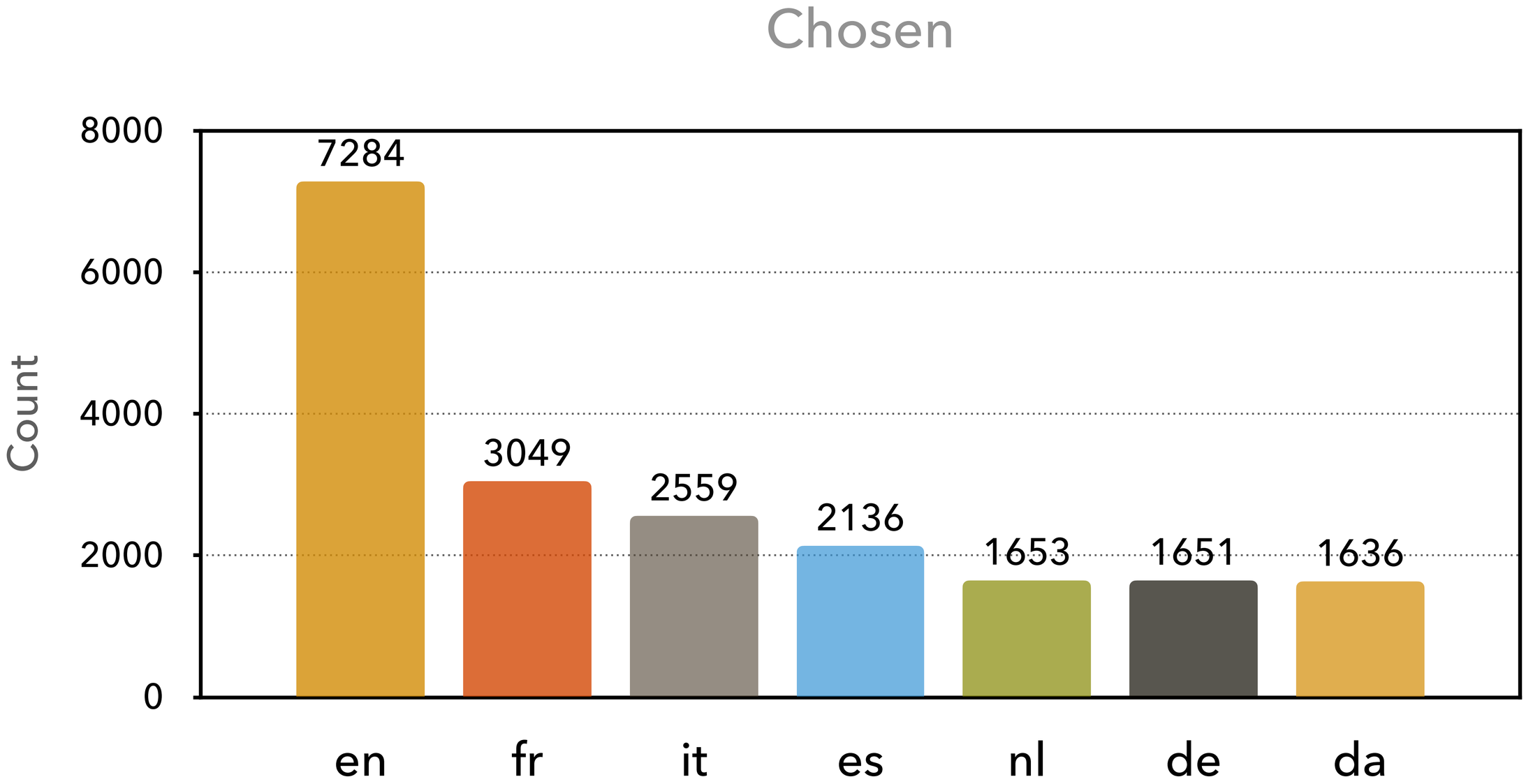
Setup (3) – Our method: Paired (reward gap)

DPO on contrastive setup



Setup (3.1) – Distribution chosen vs. rejected

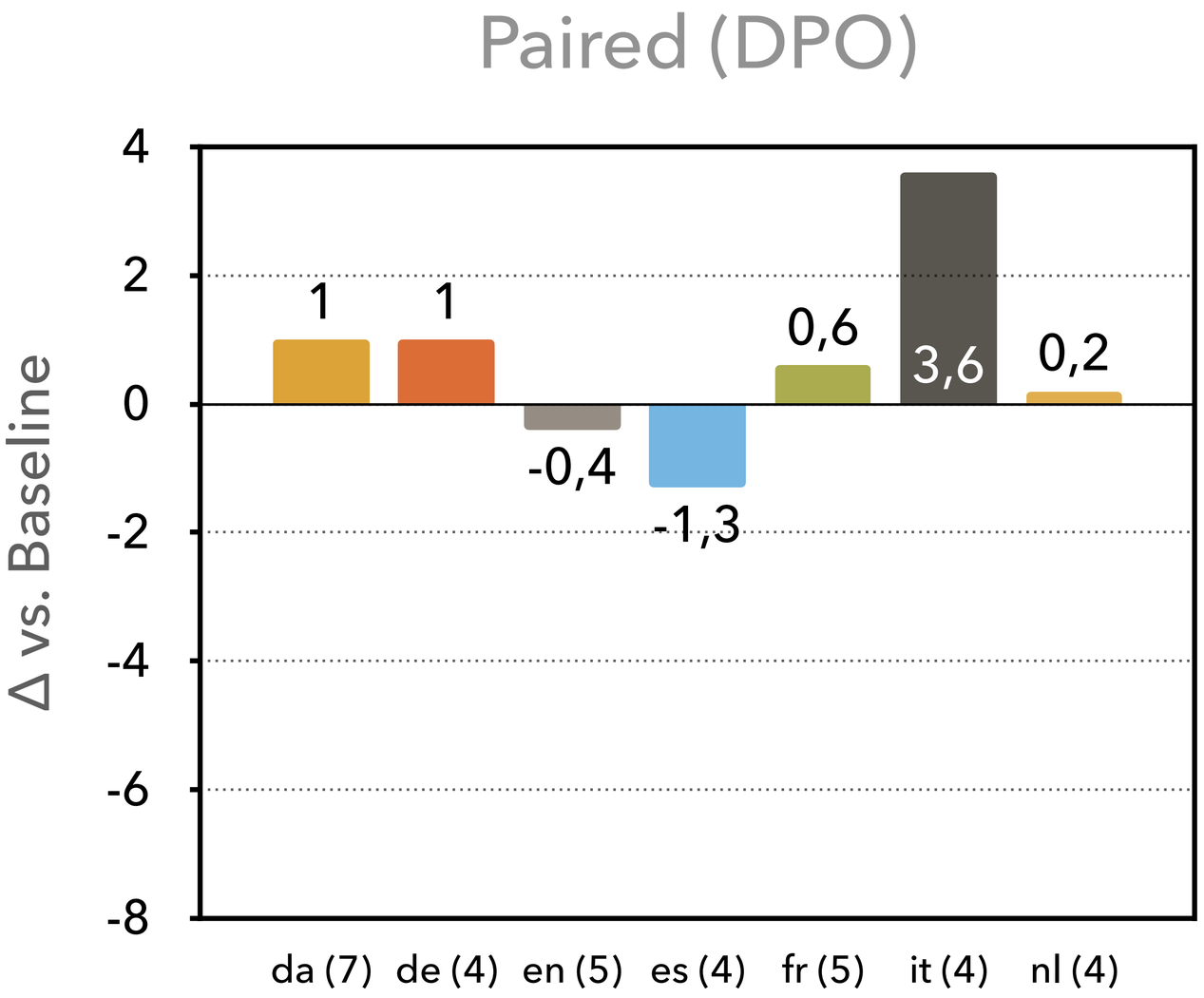
Distribution of chosen languages vs rejected languages

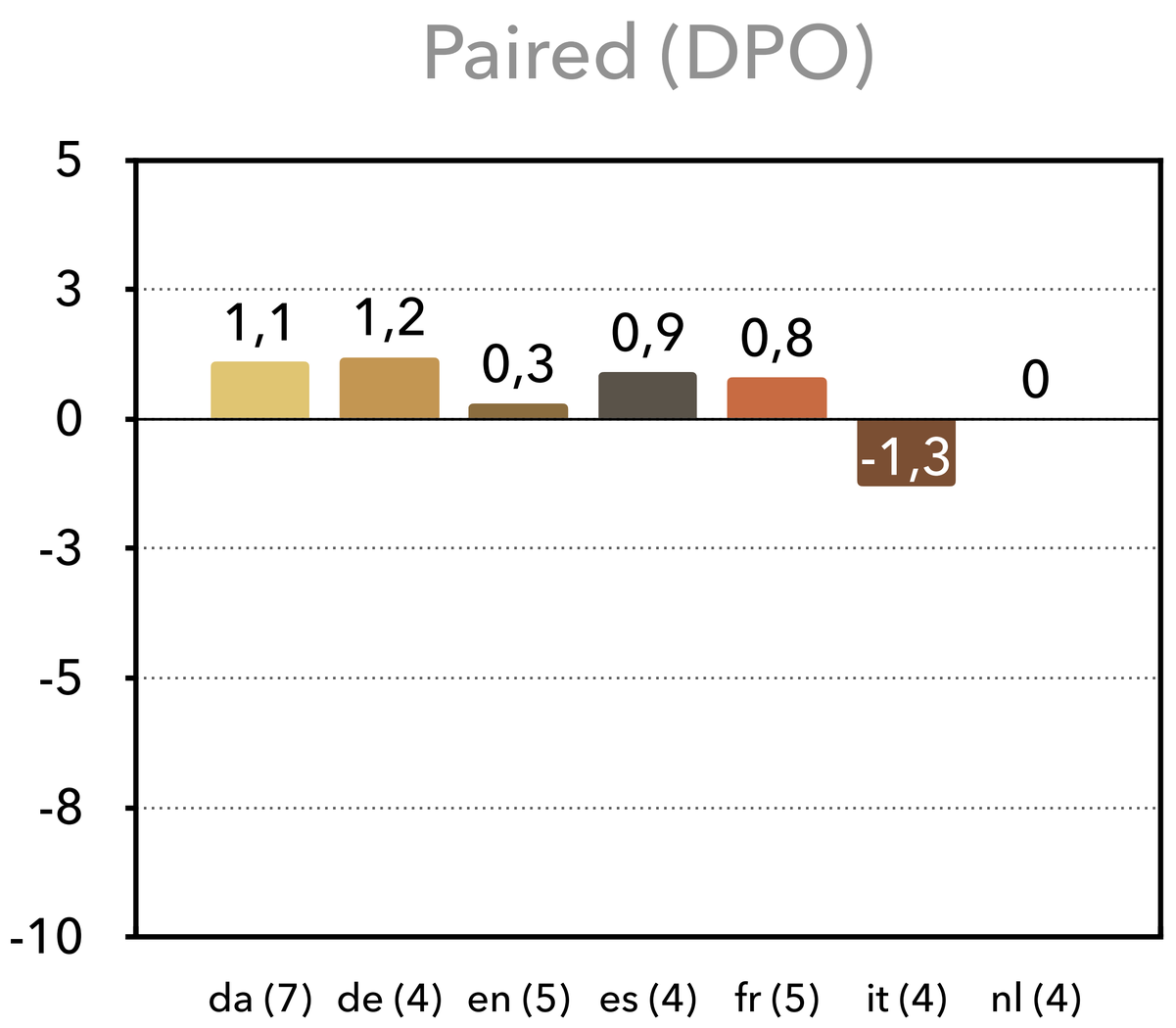


Results

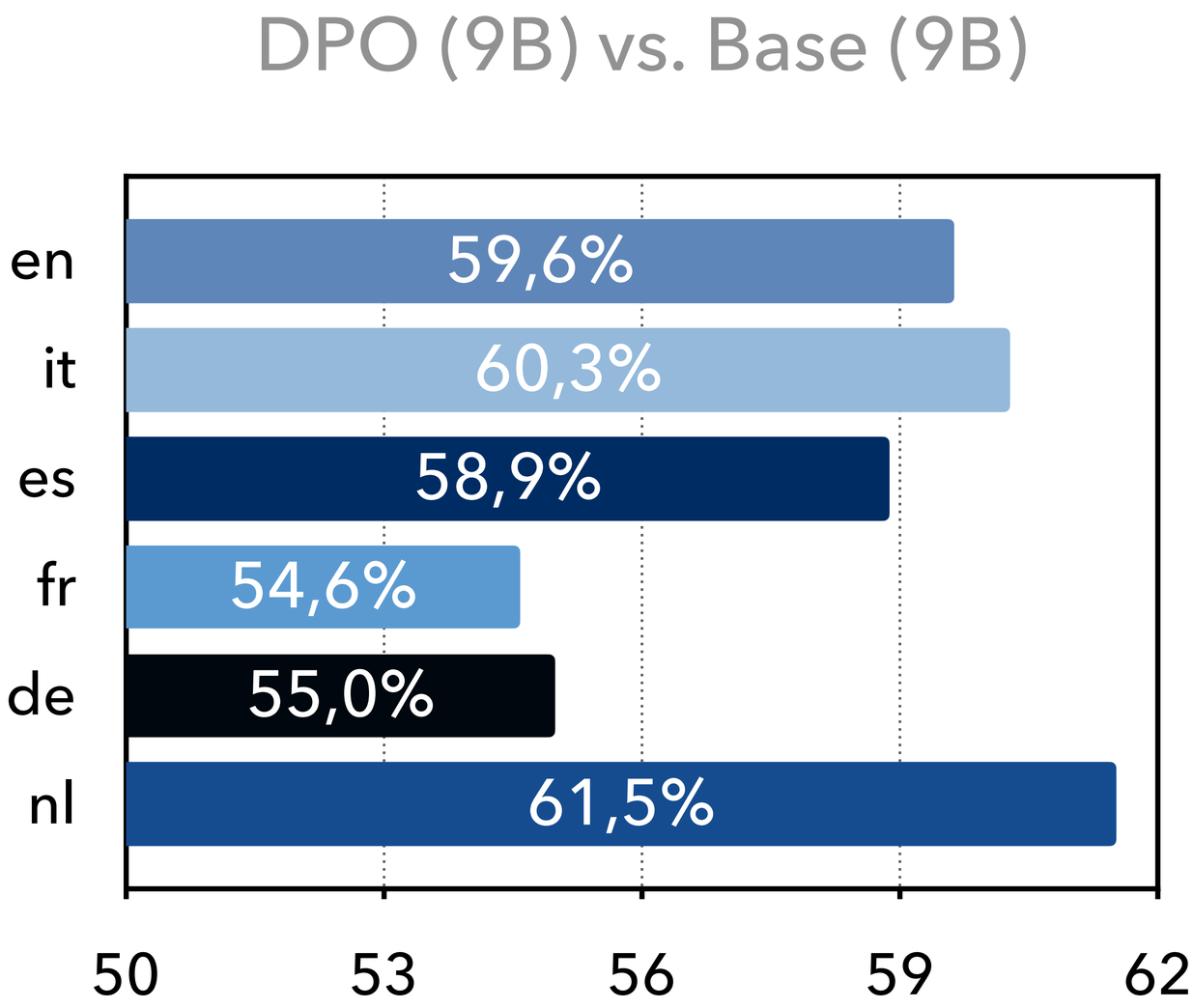
(EuroLLM-9B)

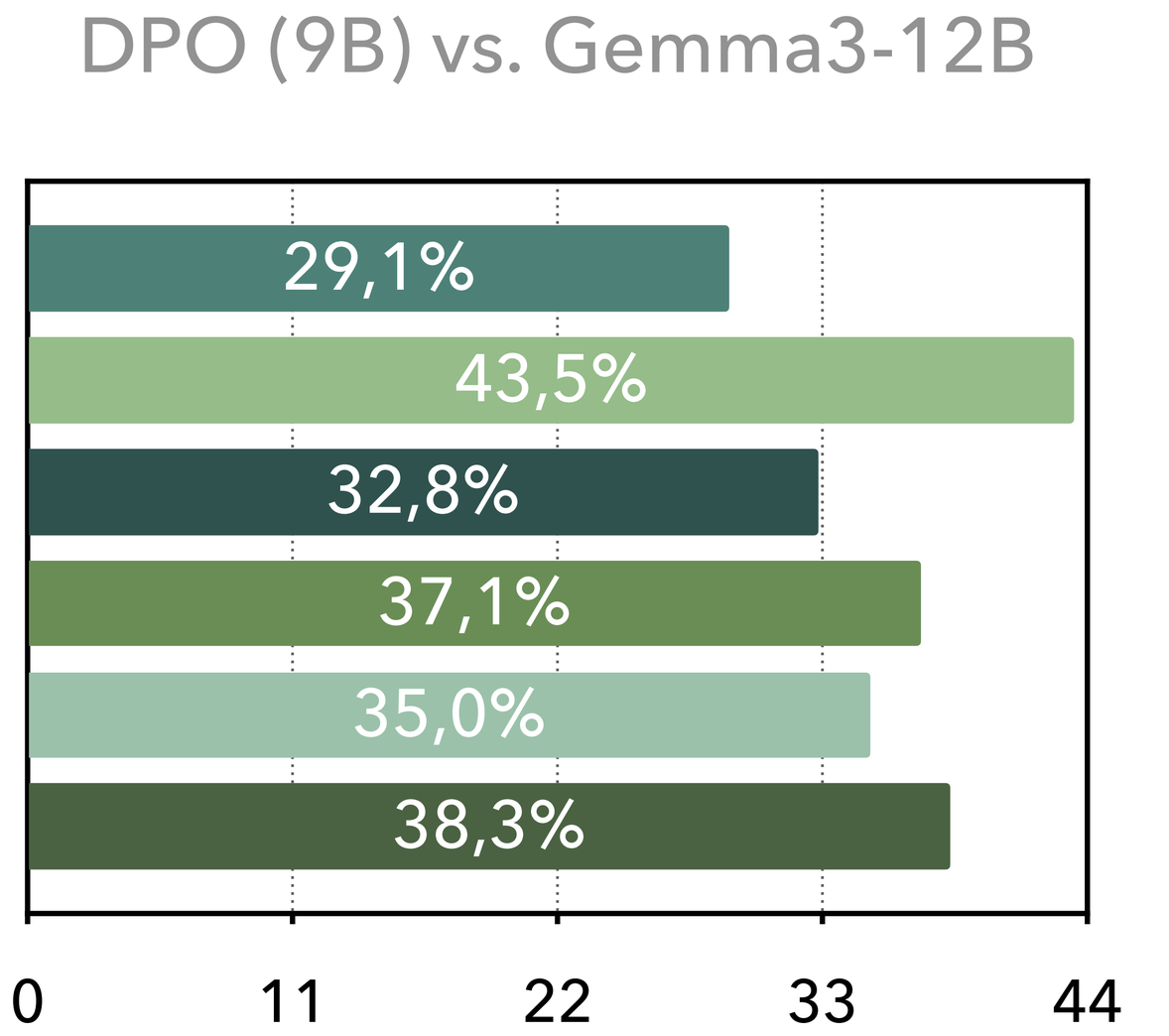
Results: Mono- and Multilingual Post-training



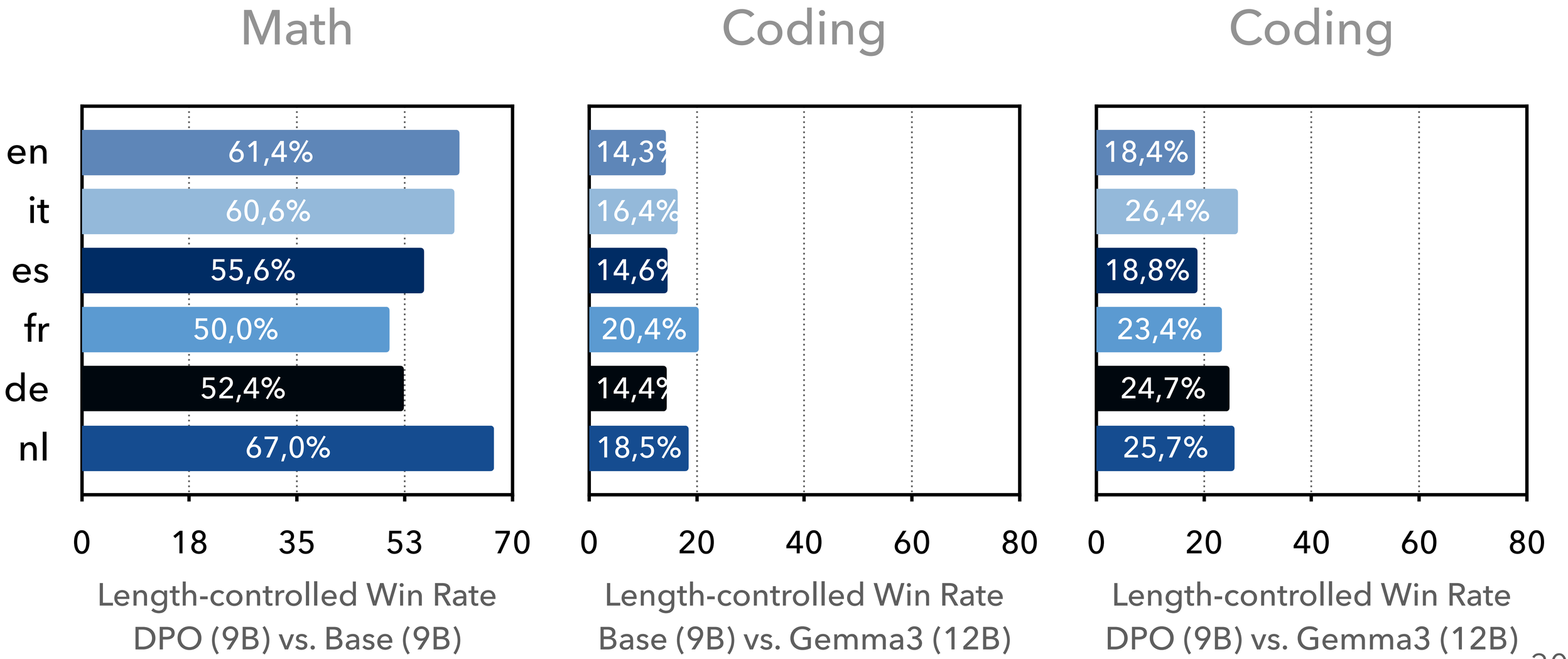


Results (3): m-ArenaHard 2.0





Results (3.1): m-ArenaHard 2.0 (by subcategories)



m-ArenaHard 2.0: What is Creative Writing?

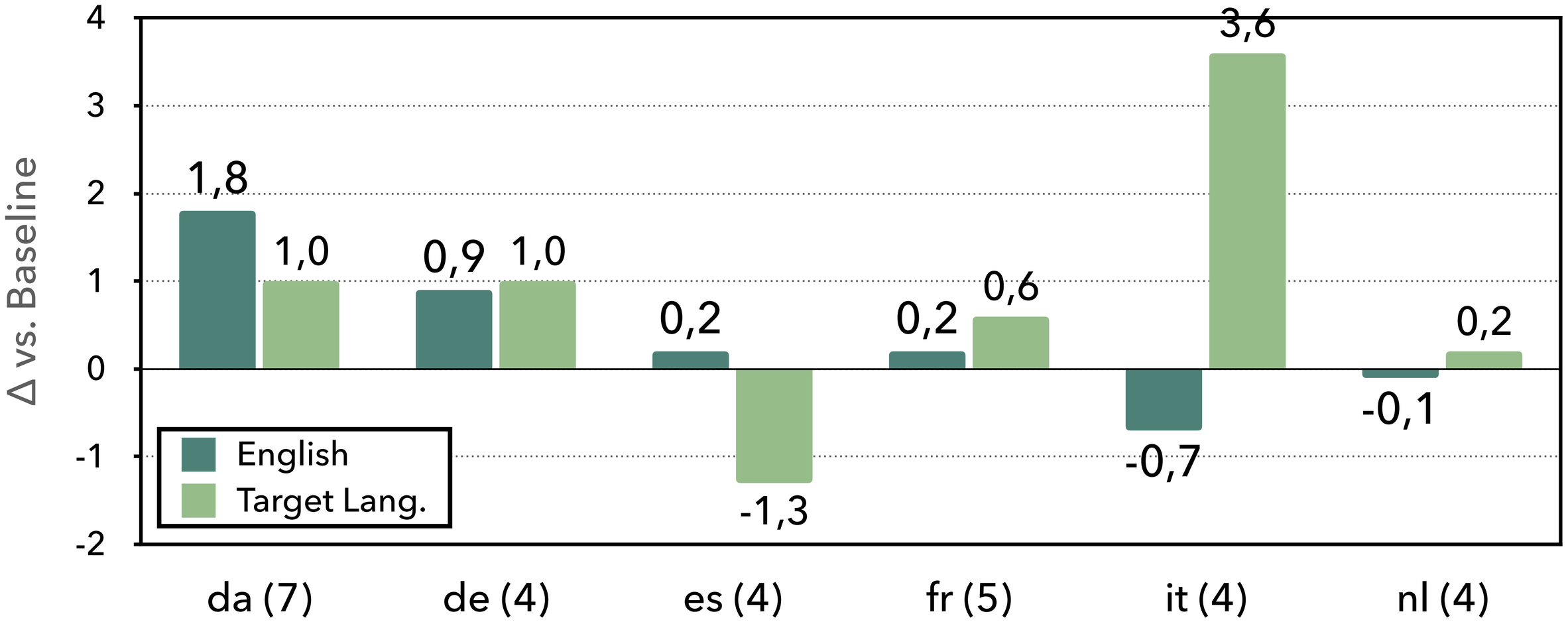
- **Two example prompts:**
  - “You are an experienced novelist crafting a novel based on the Teutonic Knights, starting with planning the main conflict of the story, the culture, society, humanities, geography, the situation with the neighboring countries, the army structure, etc.”
  - “Write me a freak folk song about love corresponded by encryption and ocassional real visits with the highest amount of internal rhyme and poetic devices possible.”

Ablations

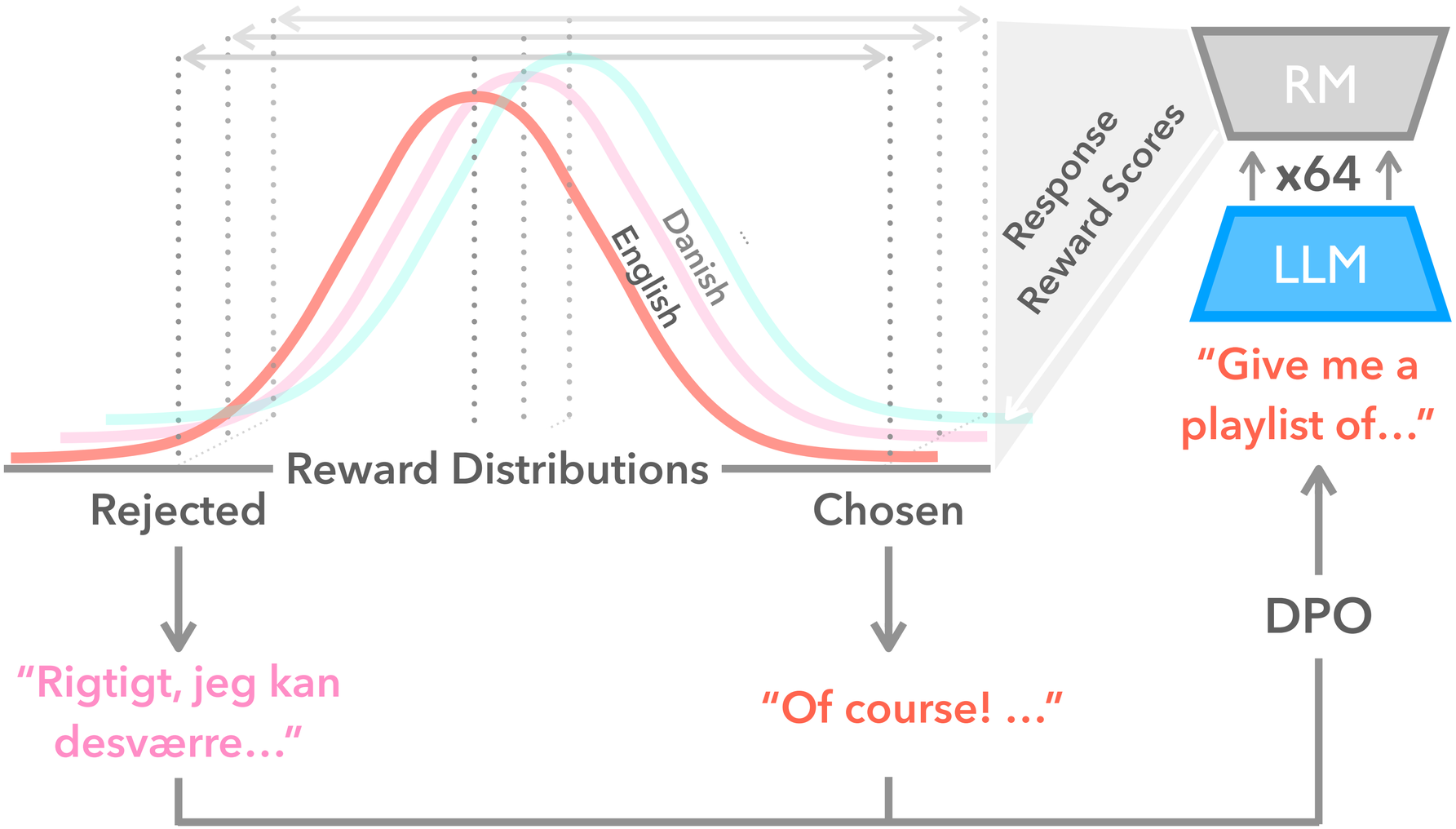
(EuroLLM-9B)

Ablation (1): Was Translating Necessary?

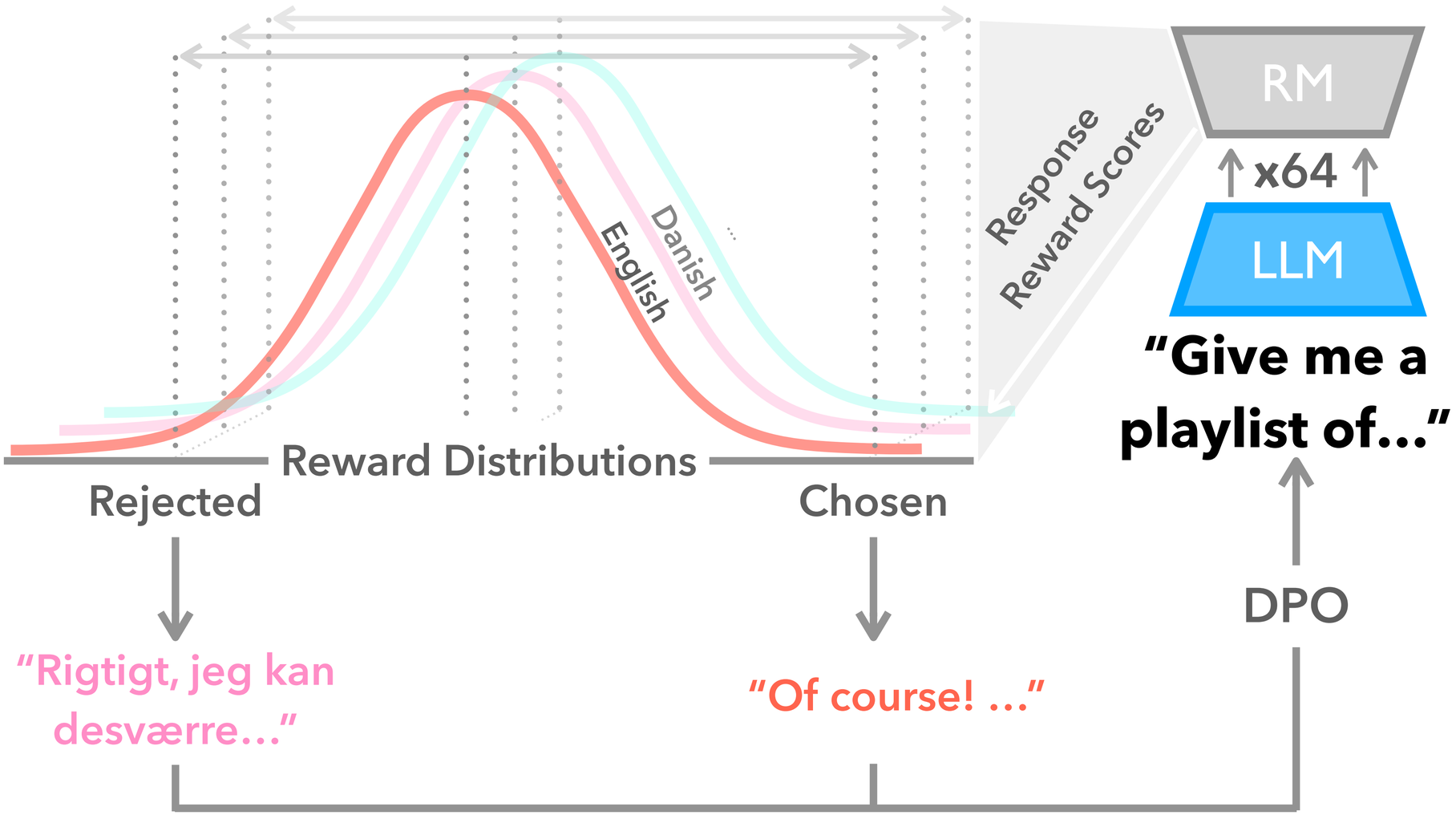
Paired Setup: Training on English only or Target Language



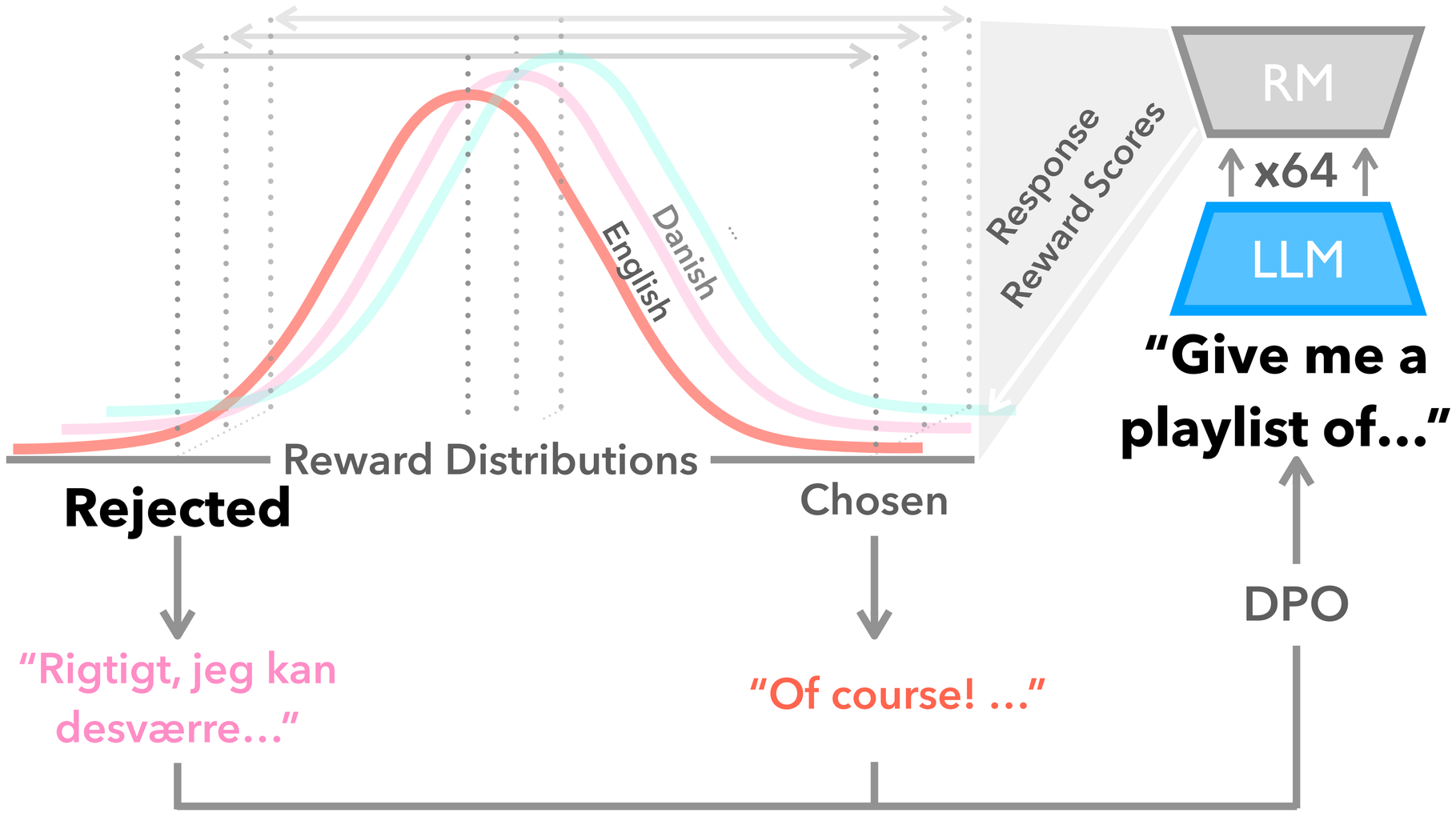
Ablation (2): Which Language should the Prompt be in?



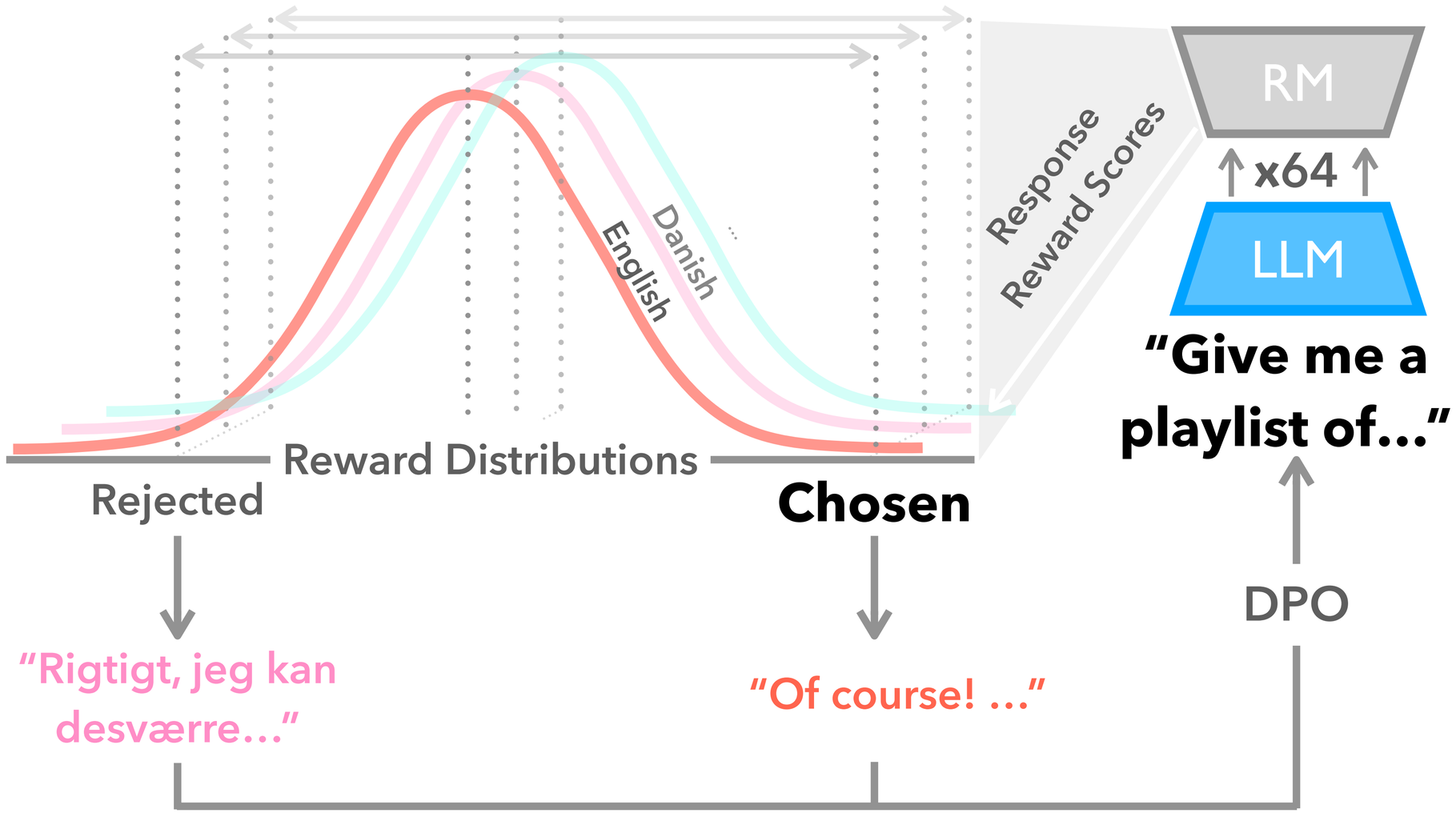
Ablation (2): Which Language should the Prompt be in?



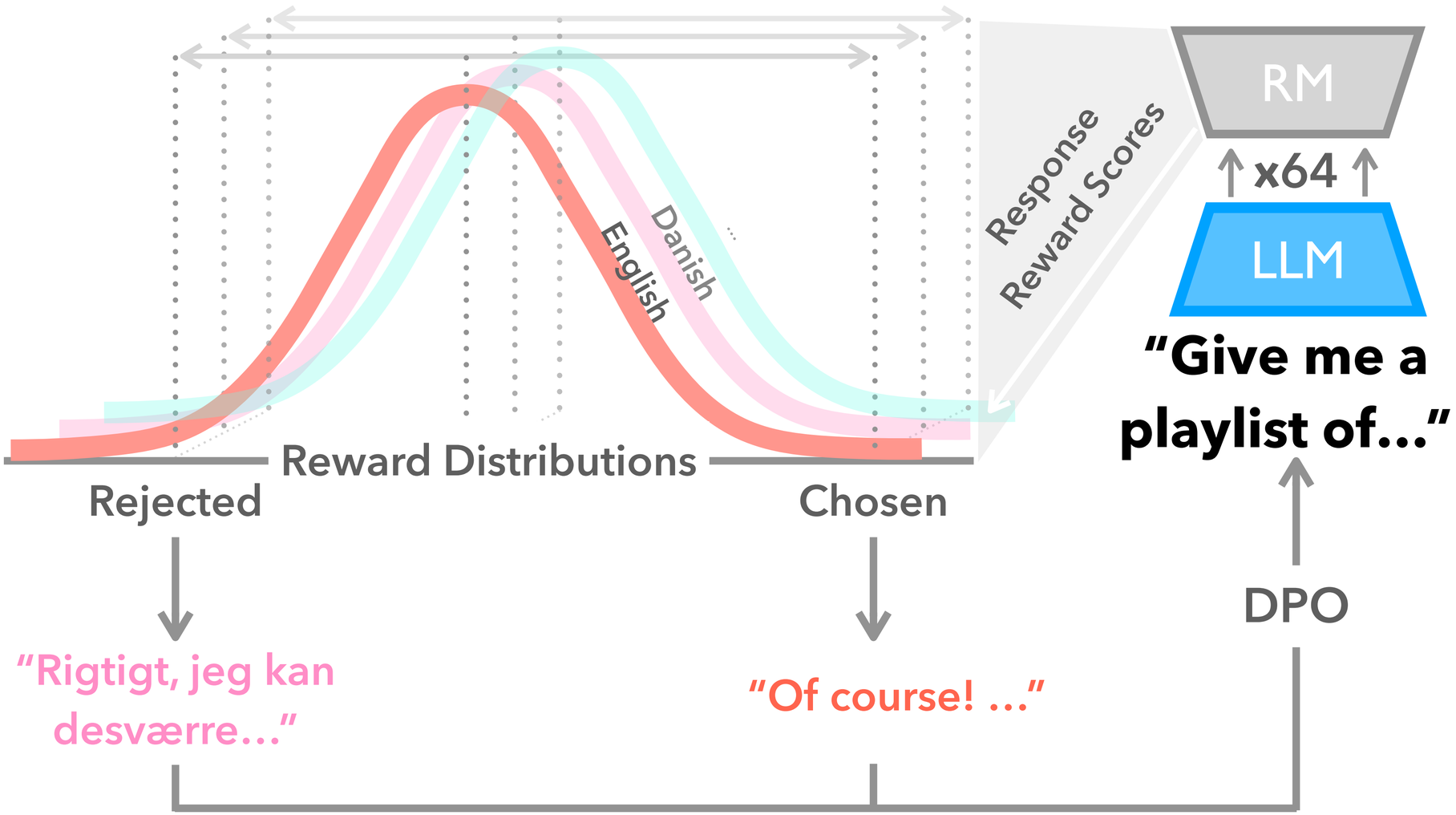
Ablation (2): Which Language should the Prompt be in?



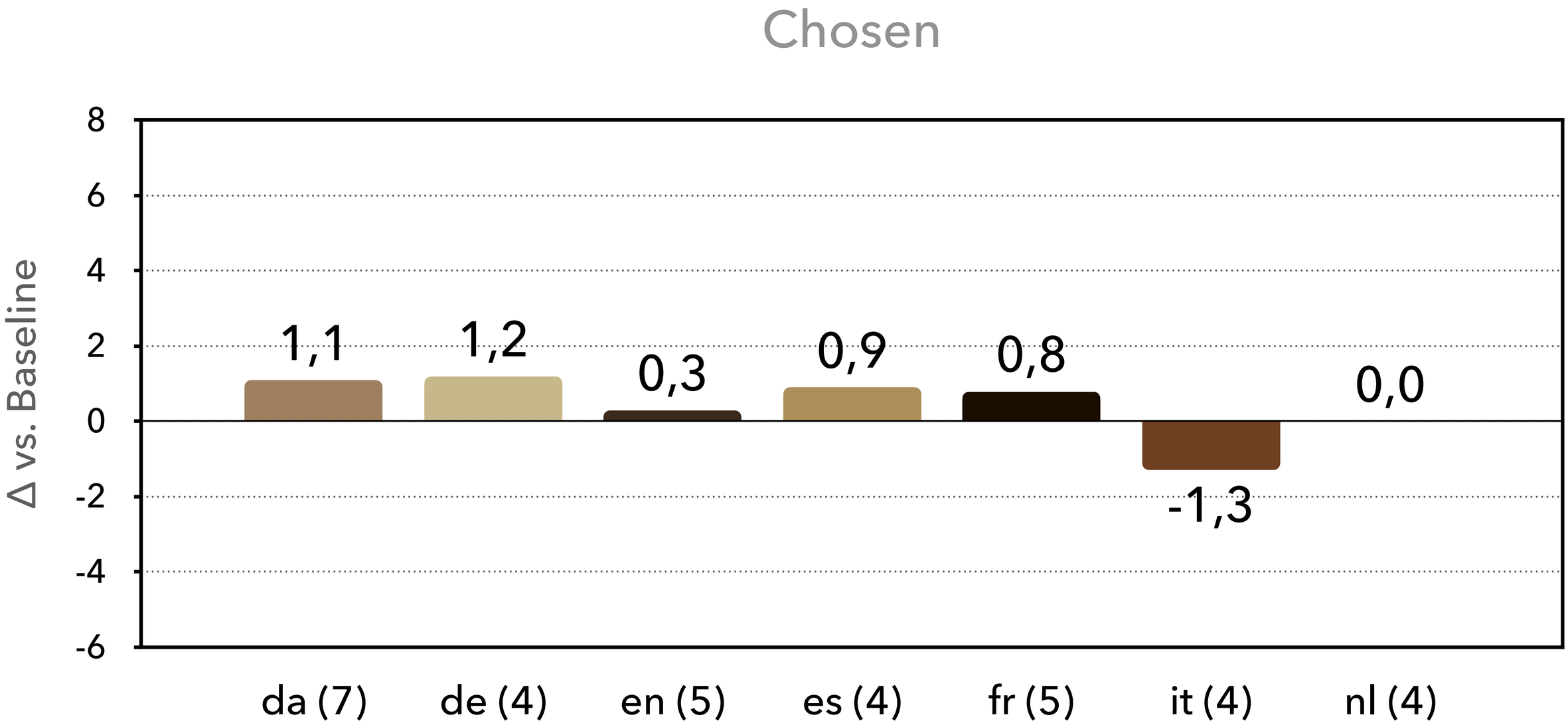
Ablation (2): Which Language should the Prompt be in?



Ablation (2): Which Language should the Prompt be in?

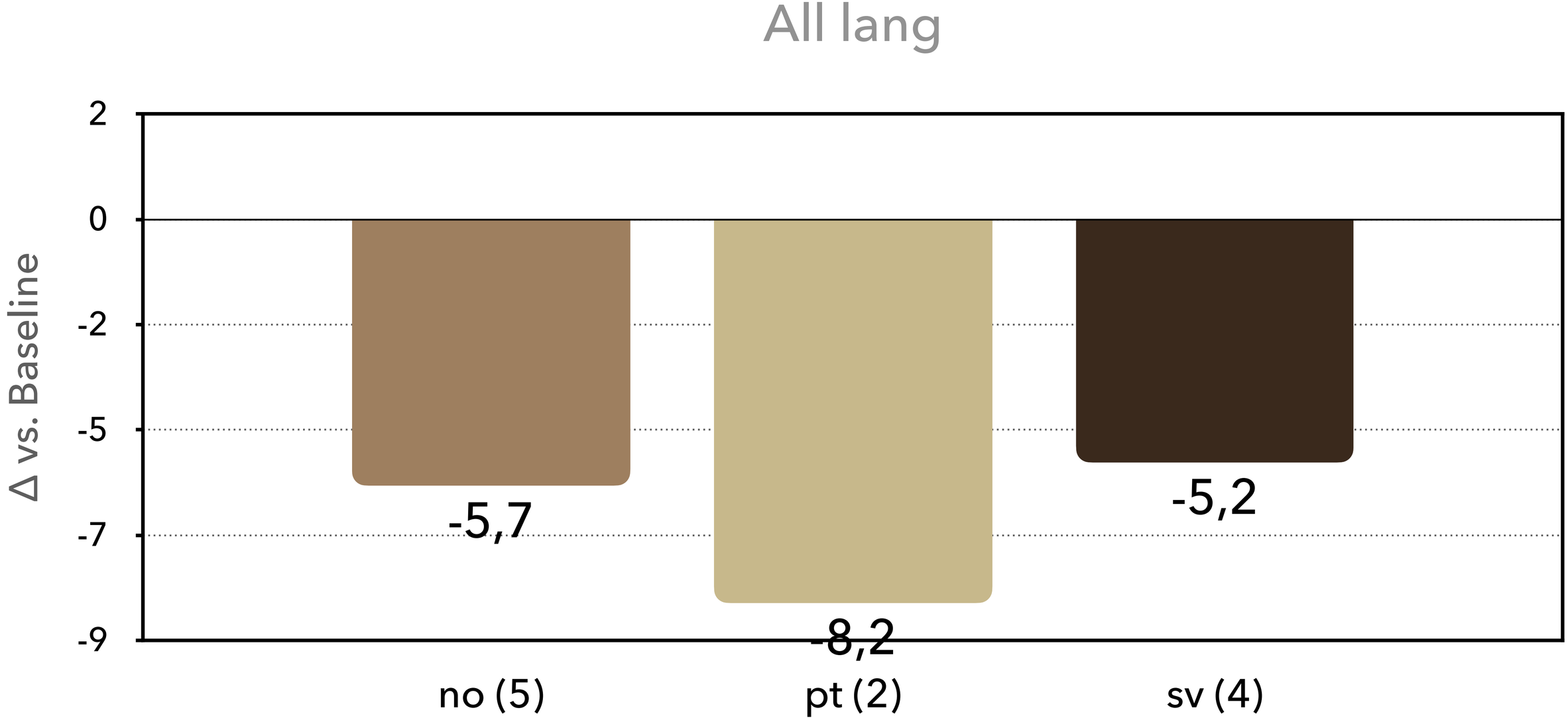


Ablation (2): Which Language Should the Prompt be in?

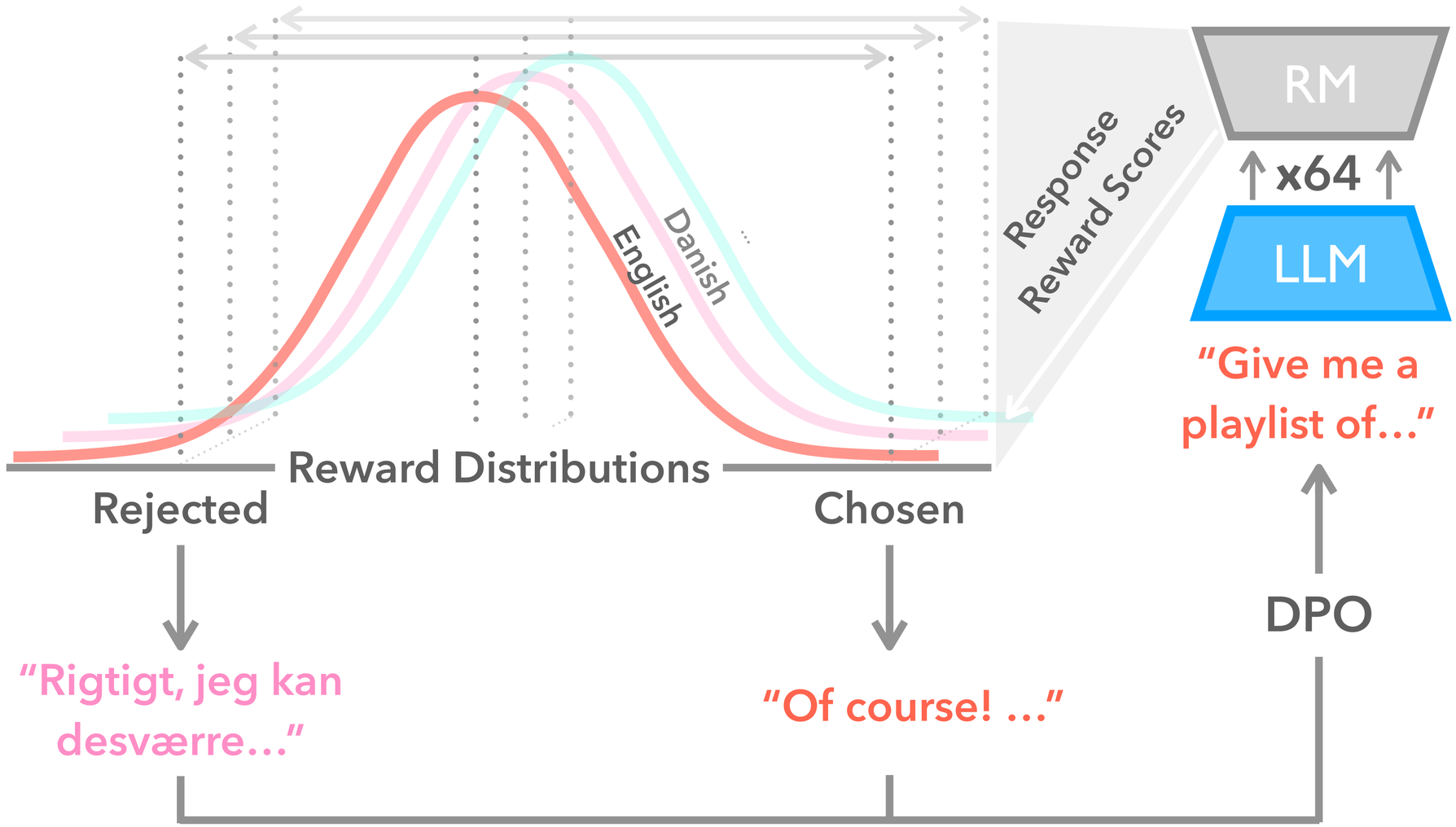


Ablation (3): “Out-of-Distribution” Performance

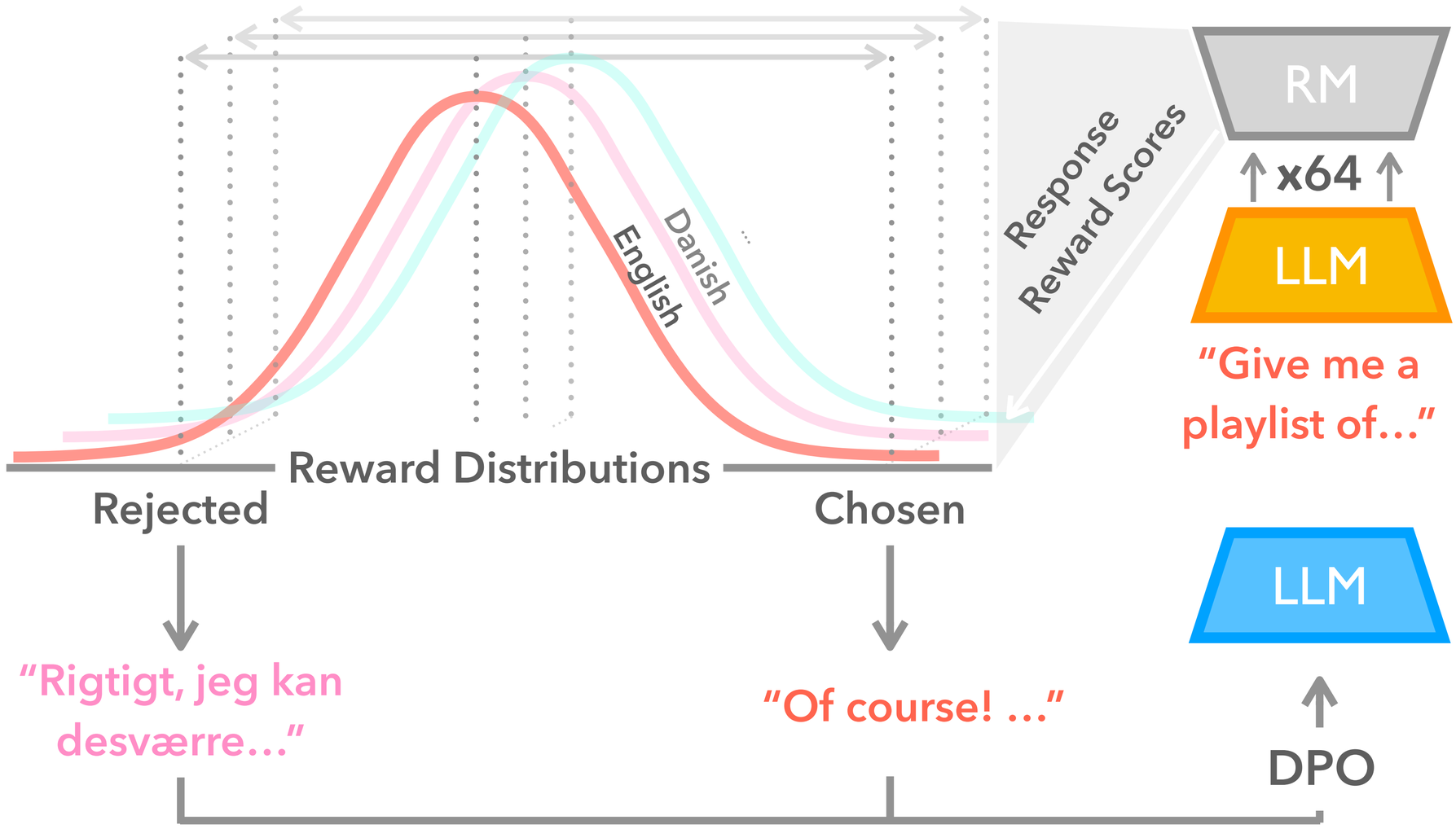
Language is not in our post-training data



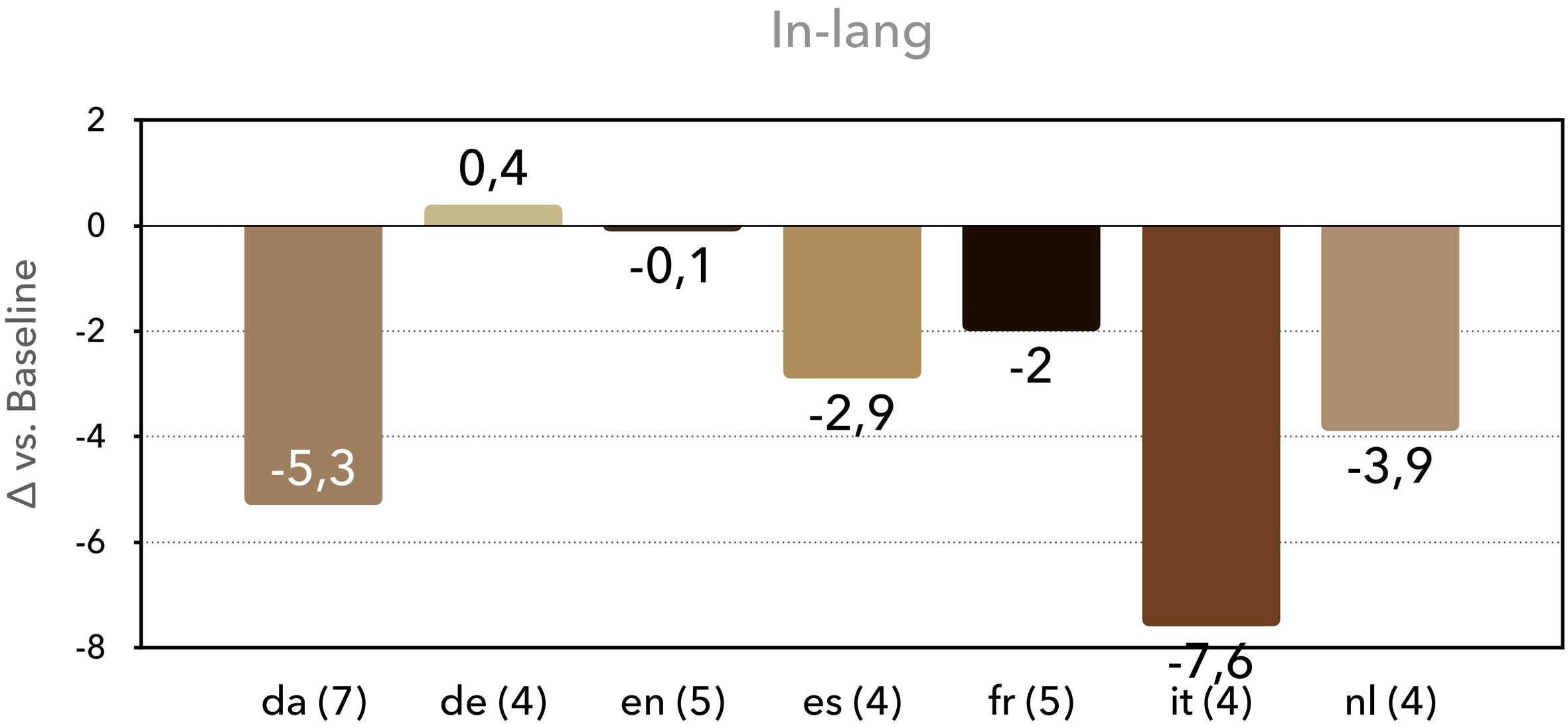
Ablation (4): Effect of Off-policy Data



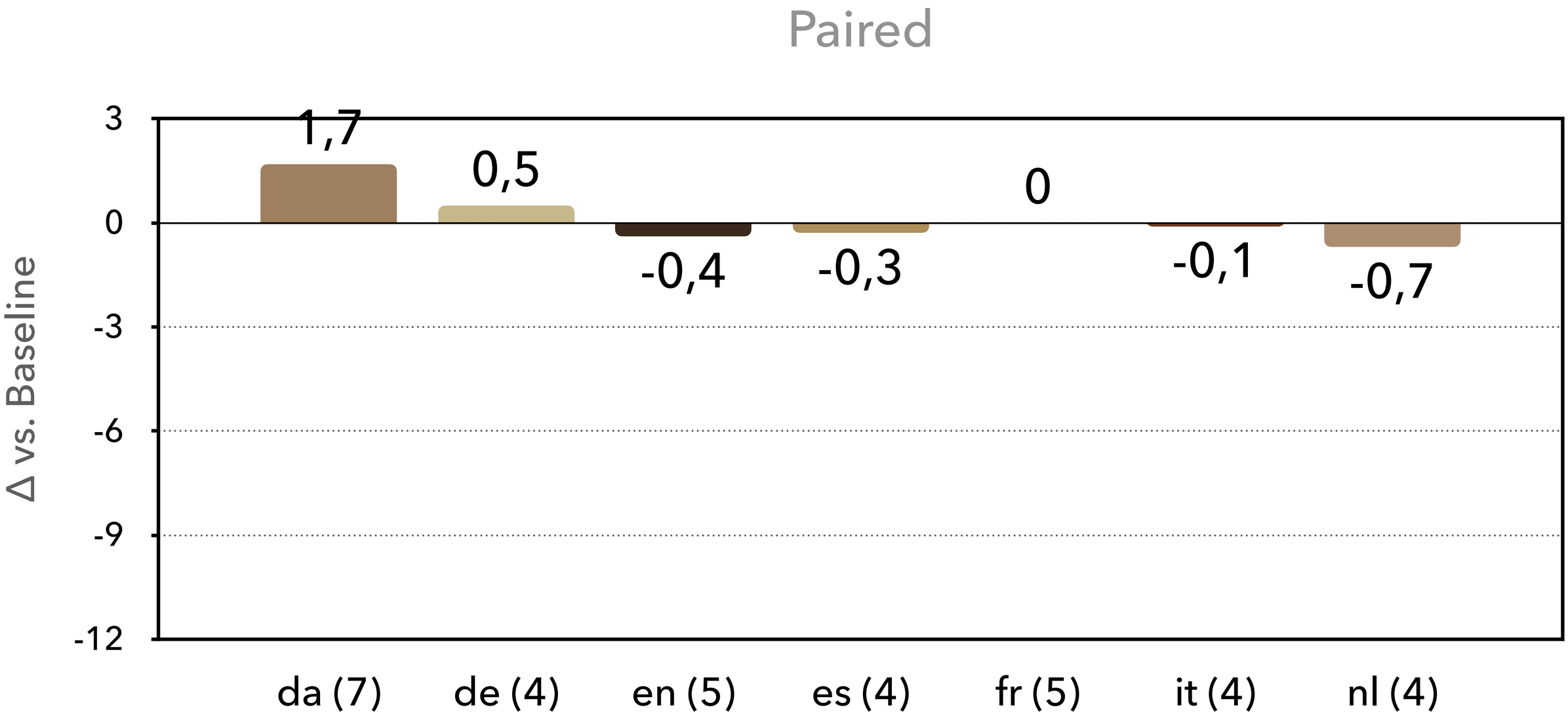
Ablation (4): Effect of Off-policy Data



Ablation (4.1; Monolingual): Off-policy Data (Tiny Aya Global 3B)



Ablation (4.2; Multilingual): Off-policy Data (tiny-aya-global 3.4B)



Takeaways

Takeaways

What do we learn from this?

- Contrastive self-distillation transfers across languages
  - Helps against catastrophic forgetting with (modest) improvements across EuroEval and m-ArenaHard 2.0
- The reward *gap* is important (paired) echoing previous work, not absolute quality (max reward).
- Off-policy data underperforms in this setup.
- Use reward scoring by off-the-shelf models more!
  - e.g., EuroLLM uses reward scores as a filter for instruction tuning data

Open Questions

What could be interesting

- To what extent would this work in low-resource scenarios?
  - Two assumptions; RM needs to score well and LLM needs some capabilities in target language.
  - (Currently running Irish, Galician, Maltese, and Welsh)
- Are we using translated data in the right way?
  - Translated data does not break the model in this DPO-style scenario
- How can we make it an online version that works?
  - GRPO-style

Thank you!

mike.zhang@di.ku.dk
jjzha.github.io
@mjjzha.bsky.social
@/in/jjzha

References (1)

- R.A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952. doi: https://doi.org/10.2307/2334029.
- Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. "Direct preference optimization: Your language model is secretly a reward model." *Advances in neural information processing systems* 36 (2023): 53728-53741.
- Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In Forty-first International Conference on Machine Learning.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 2024. Direct language model alignment from online ai feedback. ArXiv
- Yunhao Tang, Taco Cohen, David W. Zhang, Michal Valko, and Rémi Munos. 2025. Rl-finetuning llms from on- and off-policy data with a single algorithm. ArXiv
- Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, and Roy Ka-Wei Lee. 2025. Finding the Sweet Spot: Preference Data Construction for Scaling Preference Optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12538–12552, Vienna, Austria. Association for Computational Linguistics.
- Geng, Scott, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. "The delta learning hypothesis: Preference tuning on weak data can yield strong gains." *arXiv preprint arXiv:2507.06187* (2025).

References (2)

- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint, arXiv:2308.08747.
- Liu, Chris Yuhao, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan et al. "Skywork-reward-v2: Scaling preference data curation via human-ai synergy." *arXiv preprint arXiv:2507.01352* (2025).
- Olmo, Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld et al. "Olmo 3." *arXiv preprint arXiv:2512.13961* (2025).
- Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch et al. "TranslateGemma technical report." *arXiv preprint arXiv:2601.09012* (2026).
- Ramos, Miguel Moura, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal et al. "EuroLLM-22B: Technical Report." *arXiv preprint arXiv:2602.05879* (2026).
- Salamanca, Alejandro R., Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi et al. "Tiny aya: Bridging scale and multilingual depth." *arXiv preprint arXiv:2603.11510* (2026).
- Nielsen, Dan. "ScandEval: A benchmark for Scandinavian natural language processing." In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 185-201. 2023.
- Khairi, Ammar, Daniel D'souza, Ye Shen, Julia Kreutzer, and Sara Hooker. "When life gives you samples: The benefits of scaling up inference compute for multilingual llms." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27547-27571. 2025.