# DiaLLM: An Investigation into the Robustness–Generation Gap in English Dialect Adaptation

Jordan Painter[1], Dipankar Srirag[3], Adarsh Kappiyath[1], Diptesh Kanojia[1,2], Aditya Joshi[3], Lu Yin[1]

[1]NICE Research Group and [2]People-Centred AI, University of Surrey, UK [3]University of New South Wales, Sydney, Australia

## Motivation

**LLMs increasingly understand dialectal English, yet still produce standard, US-leaning English.**

**Three users chasing a late parcel:**

*Indian English:* My parcel is not **yet coming**; kindly **do the needful** and **prepone** the delivery.

*Northern British English:* **Me** parcel's not arrived; **summat** were up **wi'** the courier.

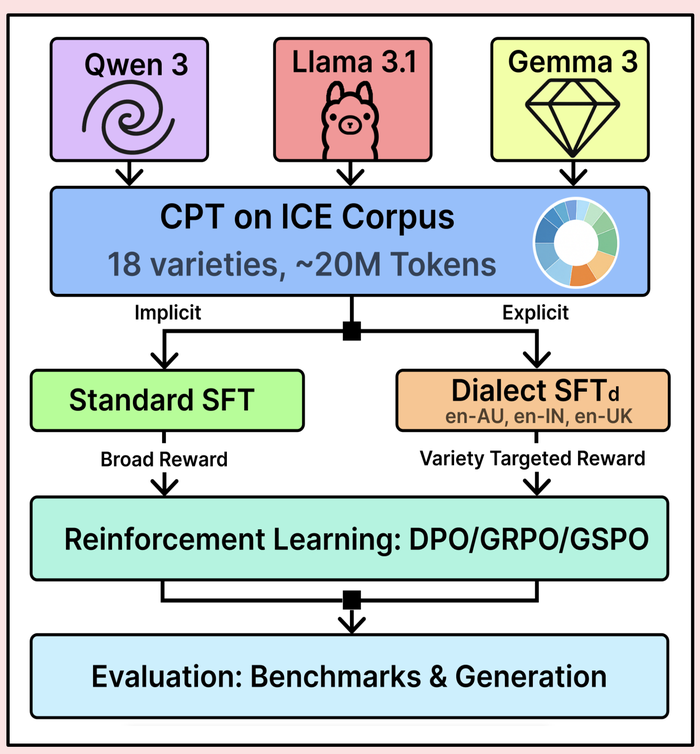
*Australian English:* Me parcel still hasn't turned up, **reckon** it's good this **arvo, eh?**

A modern LLM *understands* all three, yet answers each in flattened Standard American English.

A model that understands every variety but speaks only one has solved half the problem.

DiaLLM measures both *robustness* to dialectal input and *generation* of a variety, and asks which pipeline stage drives each.

## The DiaLLM Pipeline




We continually pretrain Llama-3.1-8B, Qwen3-8B and Gemma-3-4B-IT on the International Corpus of English (18 varieties, ~20M tokens). Two threads follow from each checkpoint. The **implicit** thread fine-tunes on standard preference data and aligns broadly. The **explicit** thread fine-tunes on data converted into one variety with Multi-VALUE (en-AU, en-IN or en-UK) and aligns towards that variety.

Both threads are aligned with DPO, GRPO and GSPO. GRPO and GSPO optimise a reward that weights eWAVE feature density against meaning preservation:

$$R = \lambda \cdot \varphi dial + \tfrac{1}{2}(1-\lambda)\cdot(\varphi comet + \varphi cos), \quad \lambda = 0.8$$

**ϕdial**: classifier over 135 eWAVE features. **ϕcomet, ϕcos**: meaning preservation against the SFT answer.

We evaluate on six benchmarks (BBH, GPQA, GLUE, VALUE, BESSTIE, DialectBench) and on 25 open prompts judged by six native or near-native annotators and by Phi-4.

## ① Can a Model Speak a Variety?

**Yes. Fine-tuning on dialect-specific data produces output that speakers recognise as dialectal.**

**Annotators picked the dialect-specific model over the instruct model (Llama-3.1-8B)**

| 84% | 87% | 48% |
|---|---|---|
| en-IN | en-UK | en-AU |

Australian English is the exception: annotators show no clear preference, consistent with its subtler features. *(Table 4, T1; ties excluded)*

**Same prompt, broad vs. variety-targeted model (selected examples, Table 3)**

**en-UK · Qwen3-8B · GRPO**

**Prompt:** *What's the best piece of advice someone's ever given you?*

| Broad model | Variety-targeted model |
|---|---|
| "Be kind to others, and treat them the way you would like to be treated." … a fundamental principle of many cultures. | This advice has helped **us** navigate through difficult situations and make decisions that align with **me** core values. |

**Features:** *possessive me; first-person singular us*

**en-AU · Llama-3.1-8B · DPO**

**Prompt:** *What are some small things that can brighten up a rough day?*

| Broad model | Variety-targeted model |
|---|---|
| … small things, depending on what makes you feel happy and energized… Treat yourself to a small indulgence… | … what makes you feel **like** happy… Kindness to **youself**: Treat **youself** to a small indulgence… |

**Features:** *like as a discourse particle; youself*

**en-IN · Gemma-3-4B-IT · GSPO**

**Prompt:** *What's your take on people who are always late?*

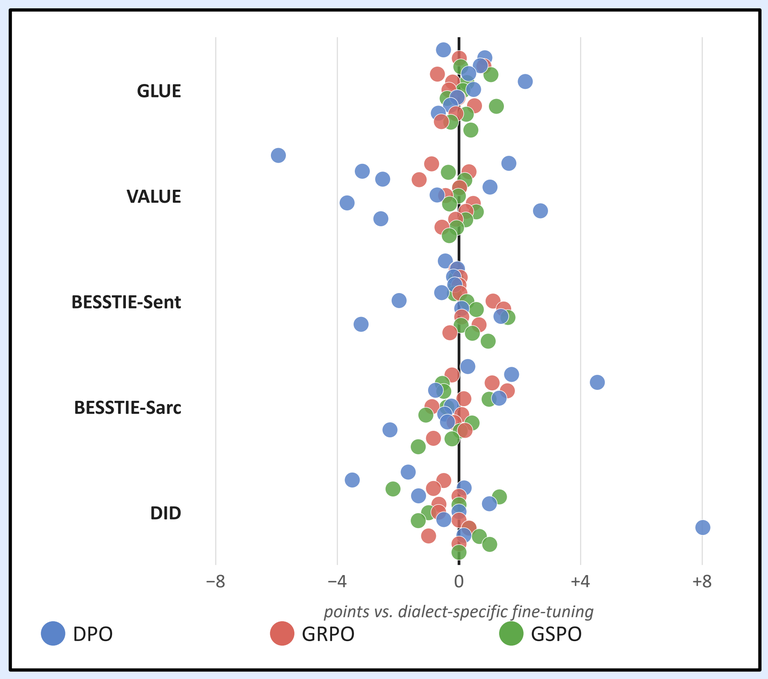
| Broad model | Variety-targeted model |
|---|---|
| … I understand that being late can be a frustrating experience for others… it's not always appropriate to judge someone solely based on their punctuality… | … **I am understanding** that being late can be frustrating, especially where **the** punctuality is highly valued… it's not appropriate to judge or **are criticizing** them without knowing the full contexts… |

**Features:** *stative progressive; article and progressive overextension*

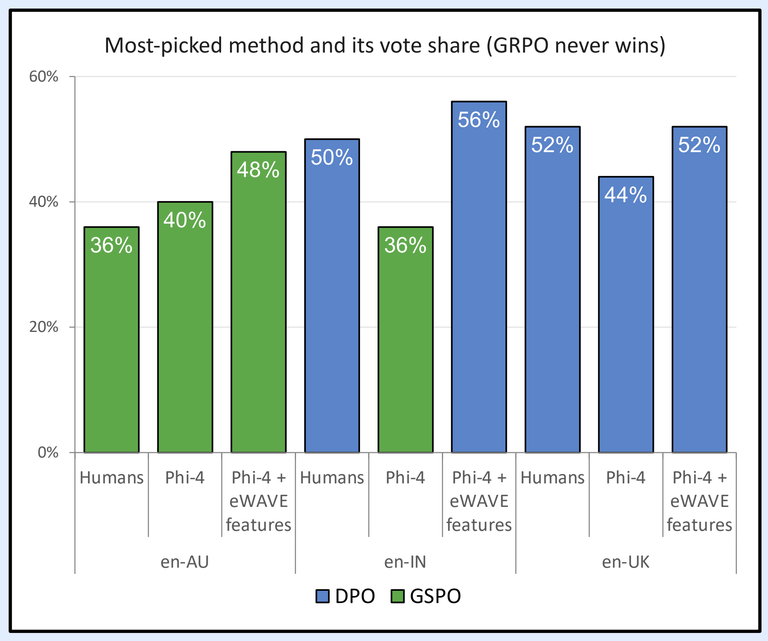
## ② Benchmarks Miss What Judges See

**DPO, GRPO and GSPO look alike on benchmarks. Judges tell them apart, and agree on a winner for en-AU and en-UK.**

**Benchmarks:** each dot is one model family and variety on one benchmark (Table 1, explicit thread).




**Judges** picked the most dialectal of the DPO, GRPO and GSPO outputs from Llama-3.1-8B (Table 4, T4): two native or near-native speakers per variety, and Phi-4 with and without the eWAVE feature list.




## ③ More Reward ≠ More Dialect

**The method the eWAVE reward likes most, GRPO, is never the one human or LLM judges pick as most dialectal.**

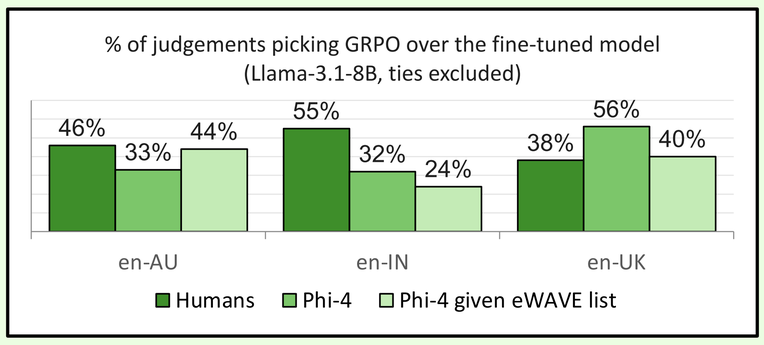
**By the eWAVE reward**

- **GRPO** ▲ most dialect features, in all three families
- **DPO** ▼ fewer dialect features than before alignment

**Judged most dialectal**

- **DPO** en-IN and en-UK (humans)
- **GSPO** en-AU
- **GRPO** never the most-picked

**Against the fine-tuned model it builds on, GRPO shows no consistent advantage with any judge**




## ④ Limits

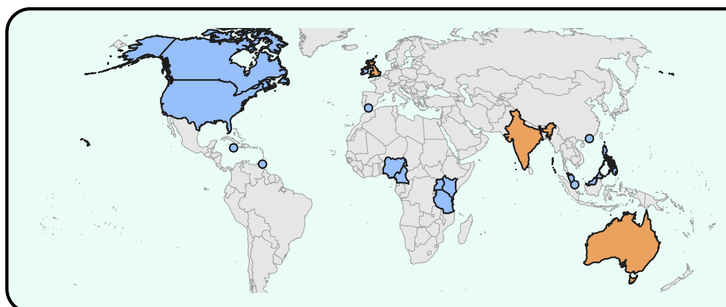
**eWAVE** lists the features a variety is known to use; it was not designed to measure how dialectal a text sounds. On en-AU, Phi-4 given the eWAVE feature list picks the dialect-specific output in 81% of judgements, while annotators are split at 48%.

**Some features** only surface in particular registers, so our generic prompts may not elicit them.

**Indian English** is consistently the most challenging variety, and has the most active features (38).

**What eWAVE features look like, with attestation ratings (A = pervasive, B = common; Appendix B, Table 6)**

| | eWAVE feature | Example | Rating |
|---|---|---|---|
| **en-UK** | Possessive me | *He's me brother* | A |
| **en-UK** | was/were generalisation | *he were thirsty* | B |
| **en-IN** | Progressive with statives | *I'm liking this* | A |
| **en-IN** | No wh- inversion | *What you are doing?* | A |
| **en-AU** | Progressive with statives | *I'm liking this* | B |
| **en-AU** | Sentence-final like | *it was good, like* | B |



**Target varieties:** Australian, Indian and Northern British English

**Pretraining:** 18 varieties from the International Corpus of English

**Variety targeting works, benchmarks do not show it, and a feature-density reward does not make output more dialectal.** Closing these gaps needs richer reward designs and more dialectal resources.

**All code, checkpoints and dialectal preference datasets are released:**

github.com/surrey-nlp/DiaLLM

**Questions:** j.painter@surrey.ac.uk