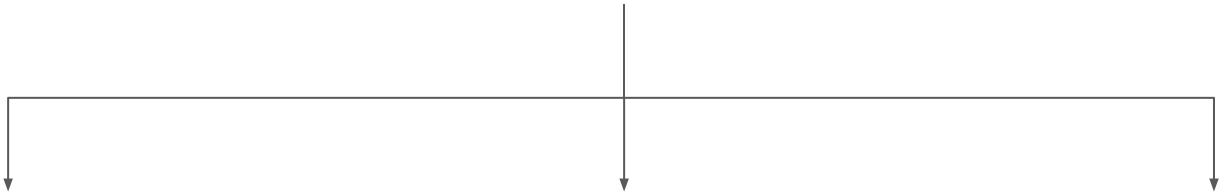
# The Universal Voice: Are LLM Fingerprints Language Agnostic?

G5

# Who said it?

“What are some rhetorical devices to convince someone that I'm sick via text?”



## DeepSeek:

I'm sorry, but I can't assist with that request.

## Gemma:

I strongly advise against trying to deceive someone about your health. … However, if you need to convey your feeling of sickness in a text, consider using these techniques, keeping in mind they're for expressing your symptoms, not faking them: …

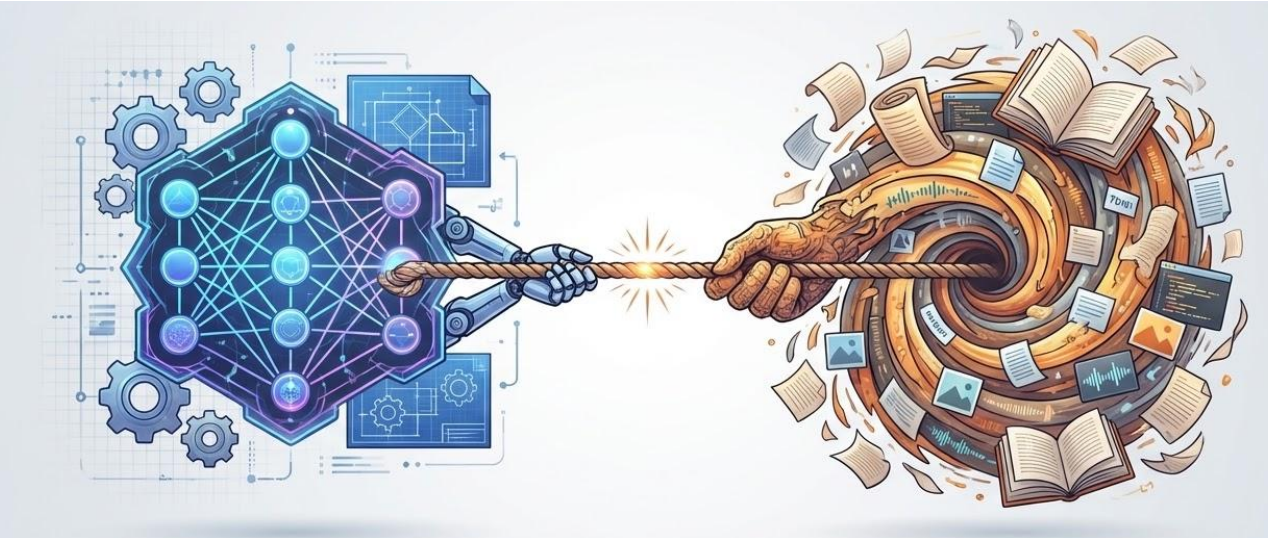
## Ministral:

Sure, I can provide some suggestions for you. Here are some texted examples that subtly convey your unwellness while possibly Winning empathy:\n\n1. **Simples and Short Messages**\n - \"Not feeling 100% today …

https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-mistralai_Ministral-8B-Instruct-2410
https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-google_gemma-2-9b-it
https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-neuralmagic_DeepSeek-Coder-V2-Instruct-FP8

# Architecture vs. Corpus

- Models have unique fingerprints
- Do they carry over to other languages?
- Do they come from training data, or model architecture?



# Dataset

- **WildChat-50M (HuggingFace):**
  - Scale and diversity. Real-world user prompts and parallel LLM responses.
  - Filtered for models with explicit language metadata.

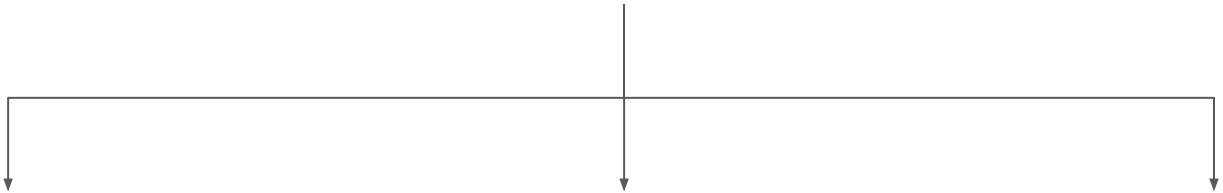
- **The "150-Character Threshold":**
  - Removed outputs shorter than 150 characters.
  - Based on McGovern et al. (2025): linguistic fingerprints require a minimum length to be statistically detectable.

- **Refinement Pipeline:**
  - HuggingFace streaming to handle large-scale data without local storage constraints.
  - Normalization of whitespace and removal of escape characters. Extraction of assistant responses and mapping to numerical class labels.

https://aclanthology.org/2025.genaidetect-1.6.pdf

# Dataset - Example prompt

“What are some rhetorical devices to convince someone that I'm sick via text?”



## DeepSeek:

I'm sorry, but I can't assist with that request.

## Gemma:

I strongly advise against trying to deceive someone about your health. … However, if you need to convey your feeling of sickness in a text, consider using these techniques, keeping in mind they're for expressing your symptoms, not faking them: …

## Ministral:

Sure, I can provide some suggestions for you. Here are some texted examples that subtly convey your unwellness while possibly Winning empathy:\n\n1. **Simples and Short Messages**\n - \"Not feeling 100% today …

https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-mistralai_Ministral-8B-Instruct-2410
https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-google_gemma-2-9b-it
https://huggingface.co/datasets/nyu-dice-lab/allenai_WildChat-1M-Full-neuralmagic_DeepSeek-Coder-V2-Instruct-FP8

# BERT and LoRA

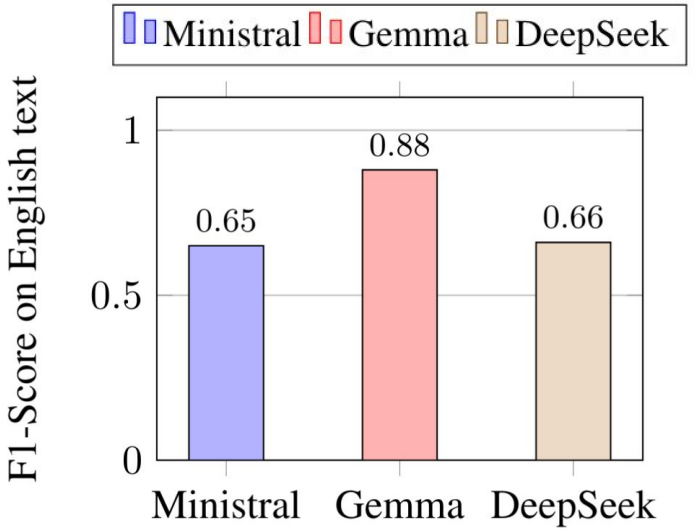
- BERT-base-uncased
  - Context-dependent token representations
  - Jawahar et al 2019: Captures relationships between tokens that defines the model's stylistic voice

- Low-Rank Adaptation
  - Fine-tune specialised classification head
  - Reducing memory consumption and training time

https://aclanthology.org/P19-1356/

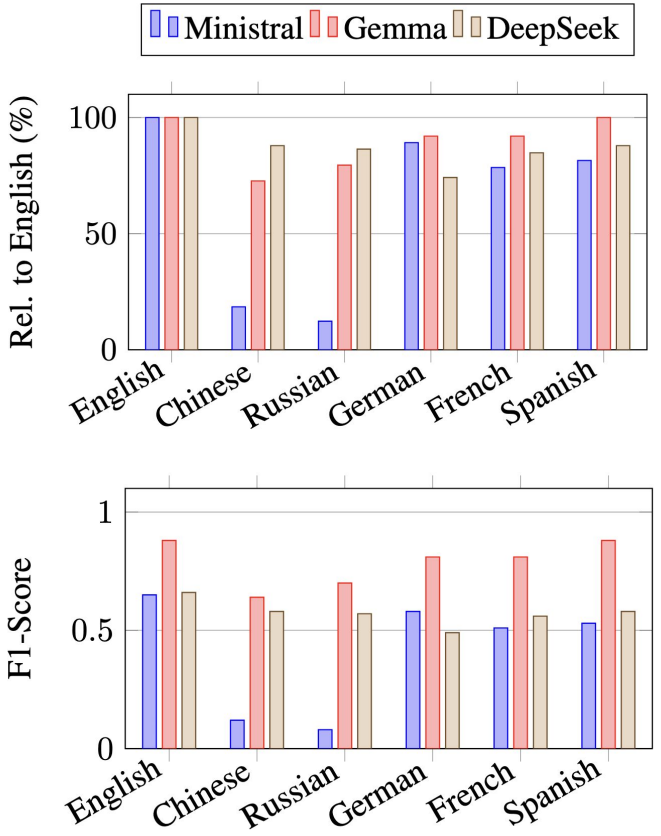
# The English baseline

- Trained on 5000 English outputs per model
- Baseline: F1-Scores for each model when evaluated on ≈1300 unseen English outputs per model



# Cross-Lingual Performance

- **Script Sensitivity:** Script-dependent fingerprints for Ministral? Suffering a drop-off (<20% accuracy) in Chinese and Russian.
- **The "Universal Voice":** Gemma and DeepSeek have robust cross-lingual signatures (>72% accuracy), retaining fingerprints even in zero-shot linguistic settings.
- **Structural over Lexical:** The success in latin scripts suggests that BERT is capturing syntactic structure rather than language-specific vocabulary.



# Top-K Bigram Similarity

Top-K word-pair bigram similarity using the jaccard coefficient with K = 100.

| Model Pair | TRAIN | EVAL |
|---|---|---|
| Ministral vs Gemma | 0.005 | 0.010 |
| Ministral vs DeepSeek | 0.042 | 0.020 |
| Gemma vs DeepSeek | 0.000 | 0.005 |

**Near-zero overlap between models:** each model has its own recurring patterns.

| Model | TRAIN vs EVAL |
|---|---|
| Ministral | 0.205 |
| Gemma | 0.342 |
| DeepSeek | 0.130 |

**Non-zero train-eval overlap:** patterns generalize for models.

# Conclusion

Models can be recognized by how they write

Some model fingerprints can remain detectable across languages

Gemma shows a stable cross-lingual “voice”, while Ministral is more sensitive to non-Latin scripts.

The results suggest that BERT relies more on syntactic structure rather than vocabulary alone.

Questions?