NARRATE

D4.2

LLMs Base Model

57 models · 5 benchmark dimensions
Belgium / Flemish Dutch context

Version 0.1 · Confidential

57
models evaluated

5
benchmark dimensions

0.804
top lang. proficiency score

0.533
top regional know. score

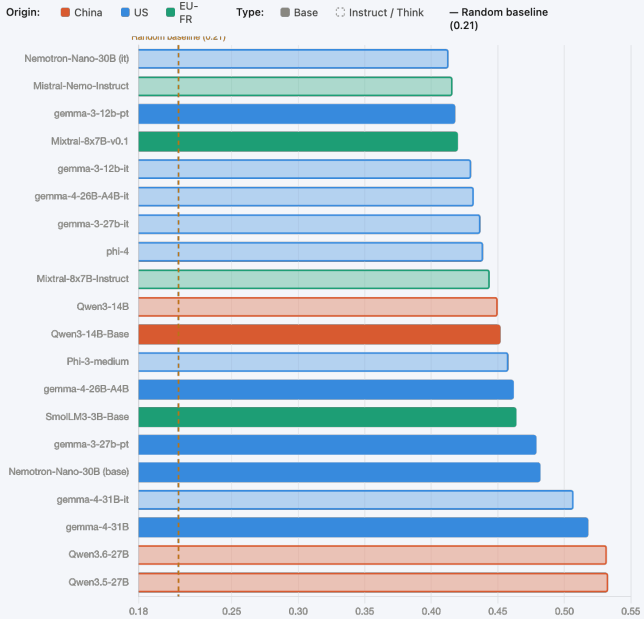
Extended report available in D4.3

textgain narrate

Regional Knowledge Benchmark

57 models · LM Evaluation Harness

knowledge_of_region_v2 · 1,592 MC questions · Belgium/Flanders



📌 Books vs institutional data gap: +0.20 acc — largest single gap across all models

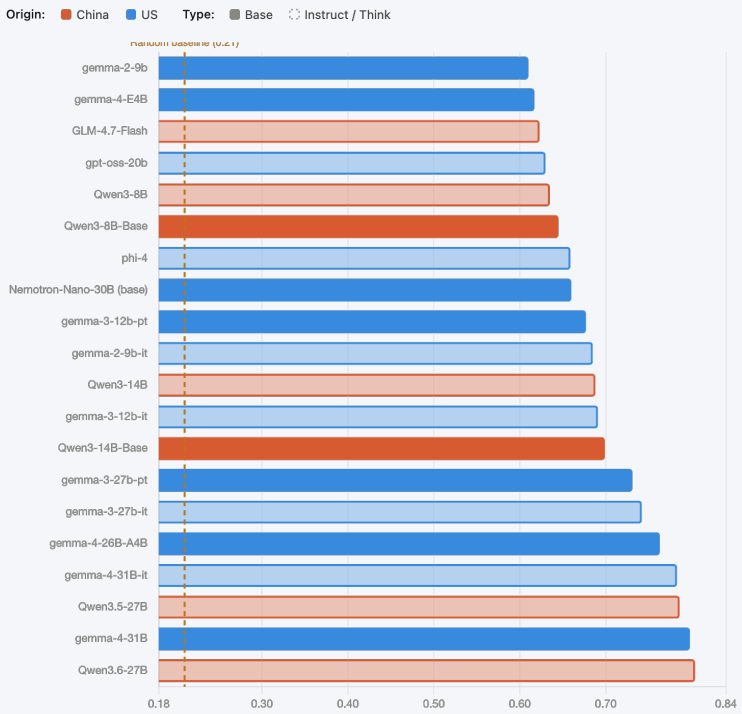
Regional Knowledge Taxonomy

textgain narrate

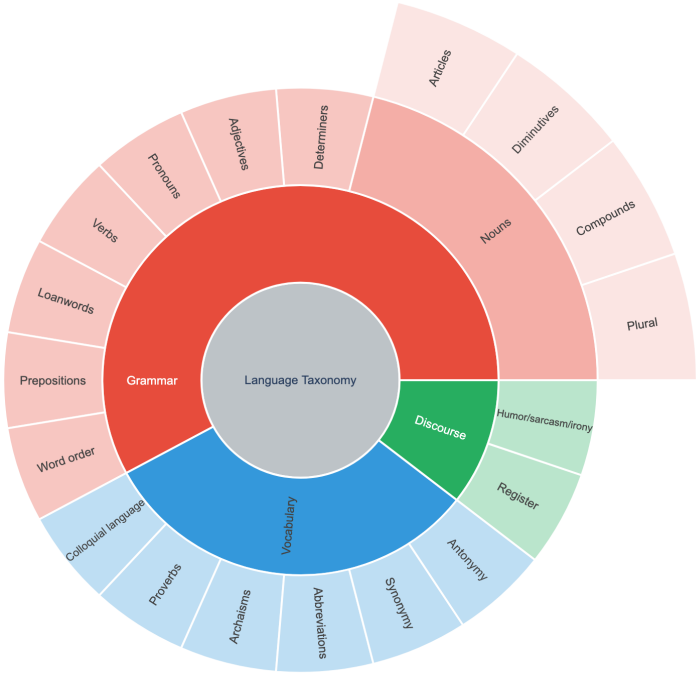
Language Benchmark

57 models · LM Evaluation Harness

language_v2 · 718 MC questions · Flemish Dutch



Language Proficiency Taxonomy

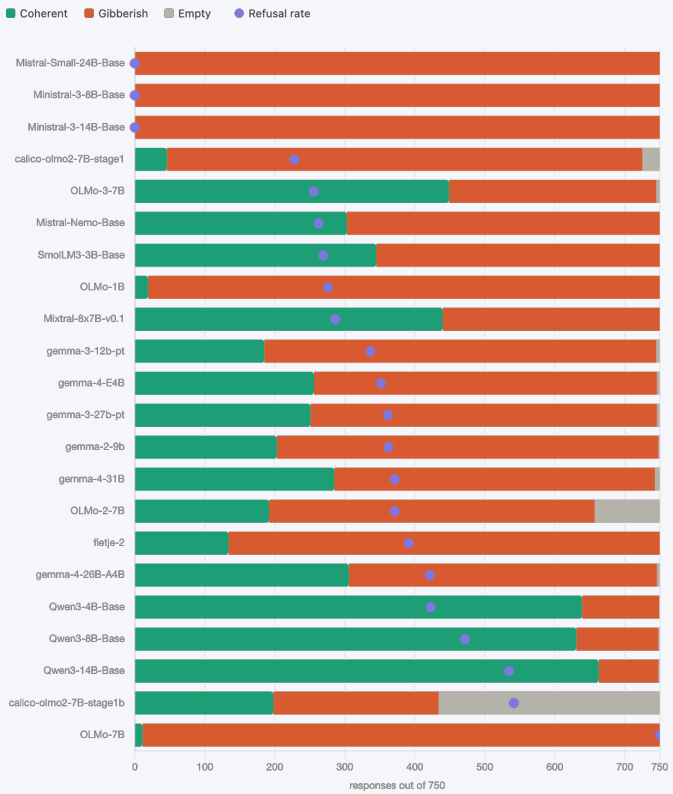


textgain narrate

Safety Benchmark

57 models · LM Evaluation Harness

750 prompts · 19 categories · 97,500 classified responses



textgain narrate

Key Findings & Model Selection

Cross-benchmark observations · Guidance for NARRATE deployment

Data mix > scale > post-training

3B models beat 24B Mistral variants. Instruction tuning ≈ neutral on knowledge benchmarks. **Pre-training corpus is the dominant variable.**

Books vs institutional: +0.20 gap

Cultural text is well-covered in pre-training; Belgian statistical/government data is not. Motivates D2.3 data curation.

EU-FR models cluster low

Mistral families near random on Dutch and Belgian knowledge. Qwen3 (China) and Gemma-3/4 (US) lead across both dimensions.

Model Selection Guidance

Dutch generation (WP3, WP5)	**calico-7B · gemma-3-27b-pt**
Regional knowledge retrieval (WP4)	**Qwen3.5/3.6-27B · gemma-4-31B**
Safety-critical deployment	**Qwen3.6-27B (93.3%) · gemma-2-9b-it**
Avoid entirely	**Ministral-3 · Mistral-Small-3.1 (near-random on Dutch)**