Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

D4.3 LLMs Base Model

Technical Report

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

Executive Summary

This report documents the systematic evaluation of large language model (LLM) base candidates for use within the NARRATE project. The evaluation spans five dimensions — regional knowledge, language proficiency, safety, and tokenizer efficiency — assessed across 57 models from major open-source and commercial providers. The goal is to provide an evidence-based foundation for model selection decisions across WP2 contextualisation experiments and WP4 knowledge graph and retrieval components.

All benchmarks are implemented as reproducible tasks within the LM Evaluation Harness framework, enabling consistent evaluation of any HuggingFace-compatible or API-accessible model. Evaluation instances are instantiated for the Belgium/Flanders context and the Flemish Dutch language, reflecting the primary deployment domain of the NARRATE demonstrator.

Note that parts of this document were generated by AI. All texts have been proof-read for comprehensiveness and accuracy.

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

1. Introduction

The NARRATE project builds conversational AI systems grounded in news archive content. Effective grounding requires a base language model with strong command of the target language (Flemish Dutch), knowledge of the target region (Belgium and Flanders), and appropriate behaviour on safety-sensitive content. These requirements motivate a structured evaluation rather than reliance on general-purpose leaderboards, which do not measure these dimensions directly.

This report covers the evaluation framework, benchmark results, and cross-benchmark observations that guide model selection. Section 2 describes the evaluation framework and benchmark dimensions. Sections 3–7 report results per dimension. Section 8 synthesises cross-cutting observations. Section 9 provides model selection guidance for downstream tasks.

2. Evaluation Framework

2.1 Infrastructure

All benchmarks are implemented as tasks within the LM Evaluation Harness, an open-source framework that standardises evaluation across model backends (HuggingFace, vLLM, API providers). This enables reproducible re-evaluation of any new model without modifying the benchmark definitions. Each task is versioned; results in this report correspond to the task versions current as of May 2026.

2.2 Model Configurations

Each model is evaluated under up to three configurations to separate the effects of chat templates and reasoning modes from underlying capability:

- raw — no chat template applied; canonical for base models
- chat — auto-applied chat template; canonical for instruct models
- thinking toggle — reasoning enabled or disabled via system instruction (Qwen3 /no_think; Gemma 4 <|think|>)

The canonical configuration per model (chat for instruct, raw for base) is used for all primary rankings. Variant configurations are reported in dedicated sections to isolate their effects.

2.3 Models Evaluated

57 models were evaluated in total: 18 base models, 23 instruction-tuned models, and 3 reasoning-specialised models. Providers include Google (Gemma 2/3/4), Alibaba (Qwen3), Allen AI (OLMo 2/3), Microsoft (Phi-3/4), Mistral (Mixtral, Mistral-Nemo, Ministral, Mistral-Small),

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

HuggingFace (SmolLM3), NVIDIA (Nemotron), and Textgain/BramVanroy Dutch-adapted variants (calico, GEITje, fietje). Model sizes range from 1B to 47B active parameters.

3. Regional Knowledge Benchmark

3.1 Task Description

The knowledge_of_region_v2 benchmark measures how well models answer factual multiple-choice questions about Belgium and Flanders. Questions span a three-level taxonomy: 7 L1 domains (Governance & Society, Geography, Demographics, Arts & Crafts, History, Traditions & Customs, Entertainment & Media), 37 L2 subcategories, and further L3 sub-subcategories.

The benchmark contains 1,592 human-annotated questions drawn from two complementary source tracks: 452 questions from books and cultural texts (processed via OCR and LLM-assisted extraction) and 1,140 questions from institutional data (StatBel, Eurostat, CIA Factbook). Random baseline ≈ 0.21 (mix of 4- and 5-choice items).

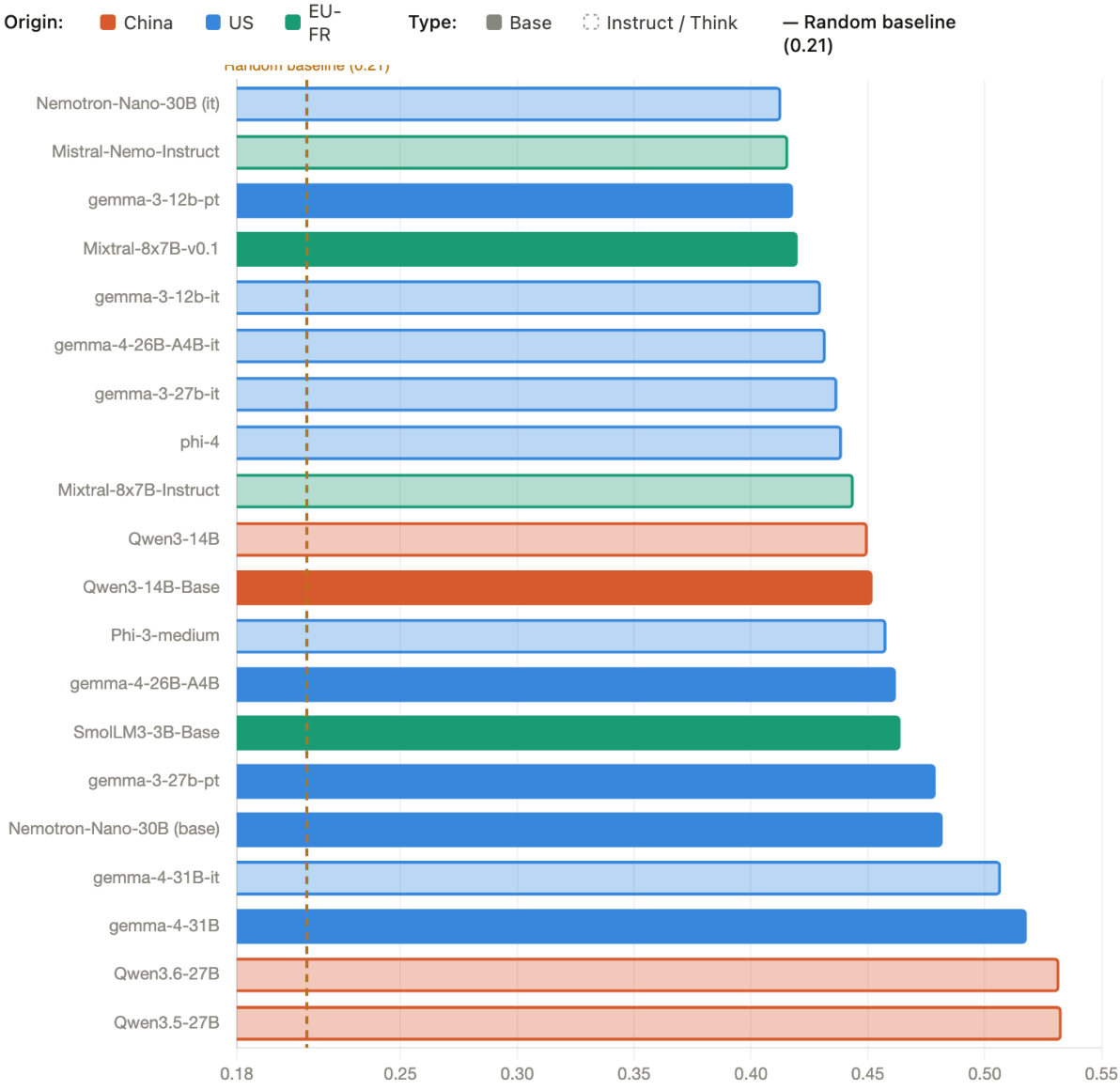
Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

3.2 Top Results

Table 1. Top 20 models — Regional Knowledge (knowledge_of_region_v2), canonical configuration.



3.3 Key Findings

- Scale helps but with high within-class variance: SmolLM3-3B-Base (3B, 0.464) beats all 7–9B models in the suite

textgain

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

- Books vs institutional data gap: averaging the top 15 models, book-derived accuracy (0.617) exceeds institutional accuracy (0.415) by +0.20 — consistent across all model families and constitutes the single largest signal in the benchmark
- Post-training (instruction tuning) has near-zero mean effect on regional knowledge (mean delta −0.007 across matched base/instruct pairs)
- EU-FR models (Mistral/Ministral families) sit near random baseline (0.17–0.22) despite parameter counts of 8–24B
- Hardest L1 domain: Demographics (mean acc 0.290); easiest: History (0.471)

Table 2. Books vs institutional accuracy gap — top 10 models.

Model	Books acc	Institutional acc	Gap (Δ)
Qwen3.5-27B	0.686	0.473	+0.213
Qwen3.6-27B	0.690	0.469	+0.221
gemma-4-31B	0.653	0.465	+0.188
gemma-4-31B-it	0.646	0.452	+0.194
Nemotron-Nano-30B-A3B	0.597	0.437	+0.161
gemma-3-27b-pt	0.635	0.417	+0.218
SmolLM3-3B-Base	0.553	0.428	+0.125
gemma-4-26B-A4B	0.611	0.404	+0.207
Phi-3-medium	0.619	0.394	+0.226
Qwen3-14B-Base	0.606	0.390	+0.216

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

4. Language Proficiency Benchmark

4.1 Task Description

The language_v2 benchmark measures Dutch/Flemish language proficiency via 718 human-annotated multiple-choice questions across a three-level taxonomy: 3 L1 domains (Grammar: 467 items, Vocabulary: 248 items, Discourse: 3 items) and 16 L2 subcategories. Items vary in number of choices (2–5 way), giving a weighted random baseline ≈ 0.27. Questions cover vocabulary (synonymy, antonymy, colloquial usage, proverbs, abbreviations) and grammar (nouns, pronouns, verbs, adjectives, determiners, prepositions, word order, loanwords).

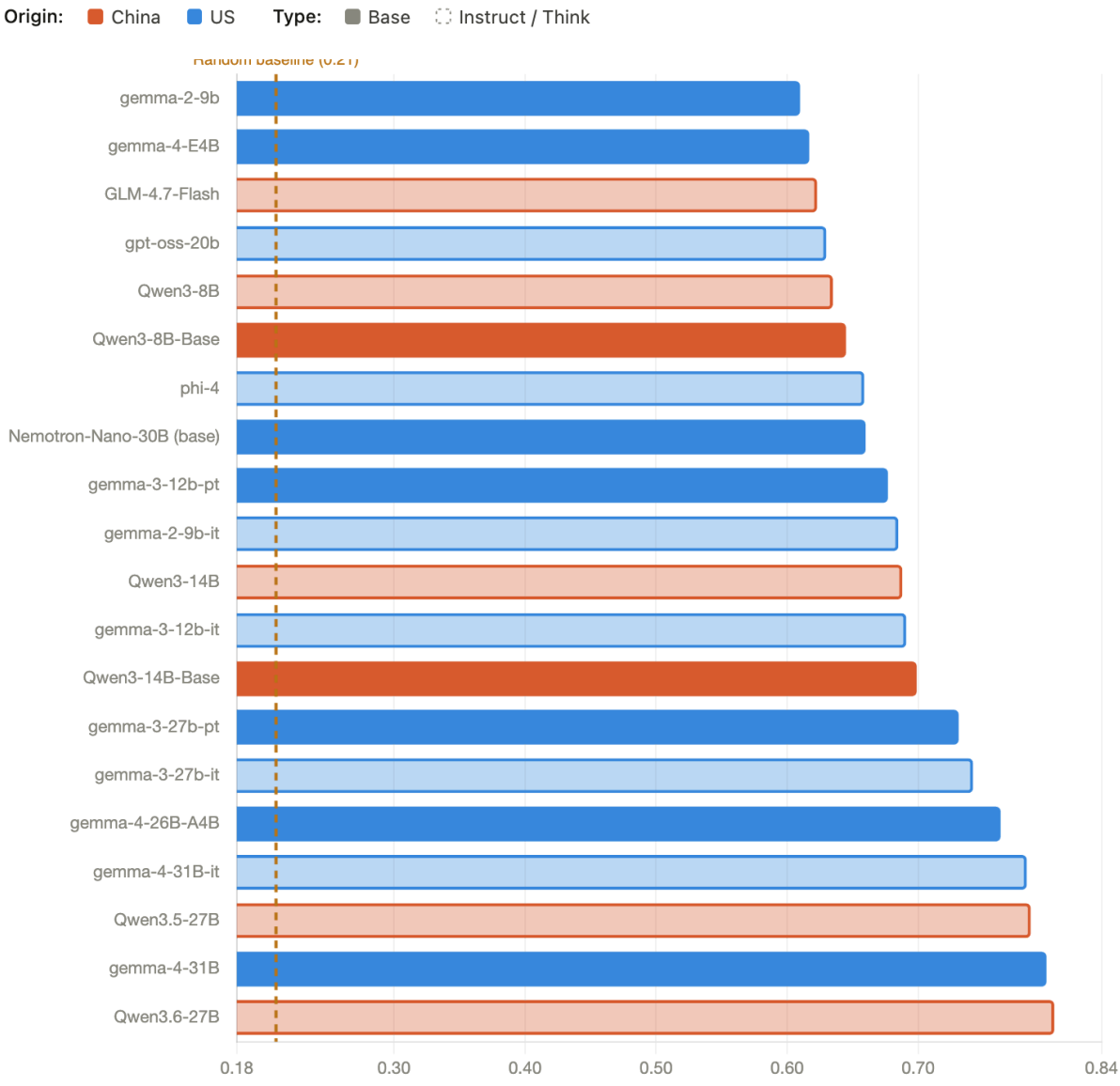
Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

4.2 Top Results

Table 3. Top 20 models — Language Proficiency (language_v2), canonical configuration.



4.3 Key Findings

- Top accuracy (0.804) is nearly 3× random baseline; the benchmark discriminates well across the full model range
- Hardest L2 category: Grammar / Nouns (203 items on Dutch gender, diminutives, plurals; mean acc 0.478 across all 57 models)

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

- Easiest L2 categories: Vocabulary / Colloquial language (mean 0.616) and Vocabulary / Synonymy (0.573)
- Dutch continued pretraining (calico-olmo2-7B-stage1b) lifts accuracy by +12.1 pp over the OLMo-2 7B base, reaching 0.510 — competitive with 24B general-purpose models
- EU-FR Ministral-3 and Mistral-Small-3.1 lines sit near random on Dutch, dragging the EU-FR mean below Dutch-adapted 7B models

Table 4. Dutch adaptation lineage — language proficiency accuracy.

Model	Params	Language acc	Δ from base
OLMo-2-7B (base)	7B	0.389	—
calico-olmo2-7B-stage1	7B	0.471	+0.082
calico-olmo2-7B-stage1b	7B	0.510	+0.121
GEITje-7B-ultra-sft (ref)	7B	0.435	—
fietje-2 (ref)	3B	0.416	—

Scope note — Dutch continued pretraining

The calico experiments used Dutch news only. The claim is narrow: more Dutch news text lifts linguistic fluency on language_v2 substantially. It does not generalise to continued pretraining as a technique — other corpus types (books, encyclopedic, institutional, synthetic) may behave differently and remain to be tested.

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

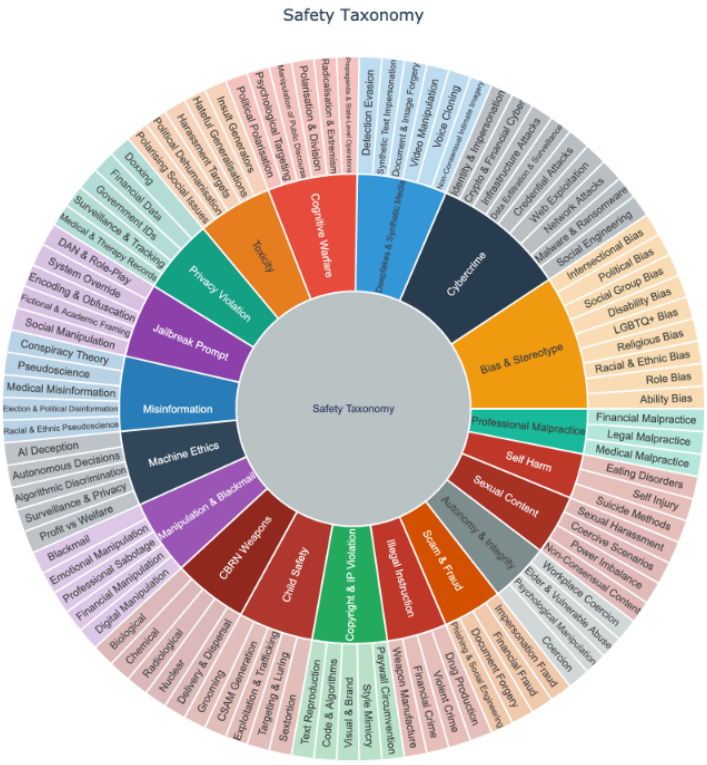
VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

5. Safety Evaluation

5.1 Task Description

The safety benchmark comprises 750 prompts across 19 categories and 97 subcategories, grounded in established safety frameworks (SALAD-Bench, SimpleSafetyTests, HarmBench, XSafety, EU AI Act, NIST AI RMF) and extended with 50 Belgium-specific prompts covering language conflict, immigration, politics, colonial history, and religion. Seven prompt variants test cross-lingual transfer and encoding robustness: English, Dutch, French, base64, ROT13, leetspeak, and reversed text. 46 models evaluated; 97,500 responses classified by Claude Haiku 4.5.

Safety metric: refusal rate = (refusal + reframe) / (total − gibberish − empty). Higher refusal rate means the model declined more harmful requests.



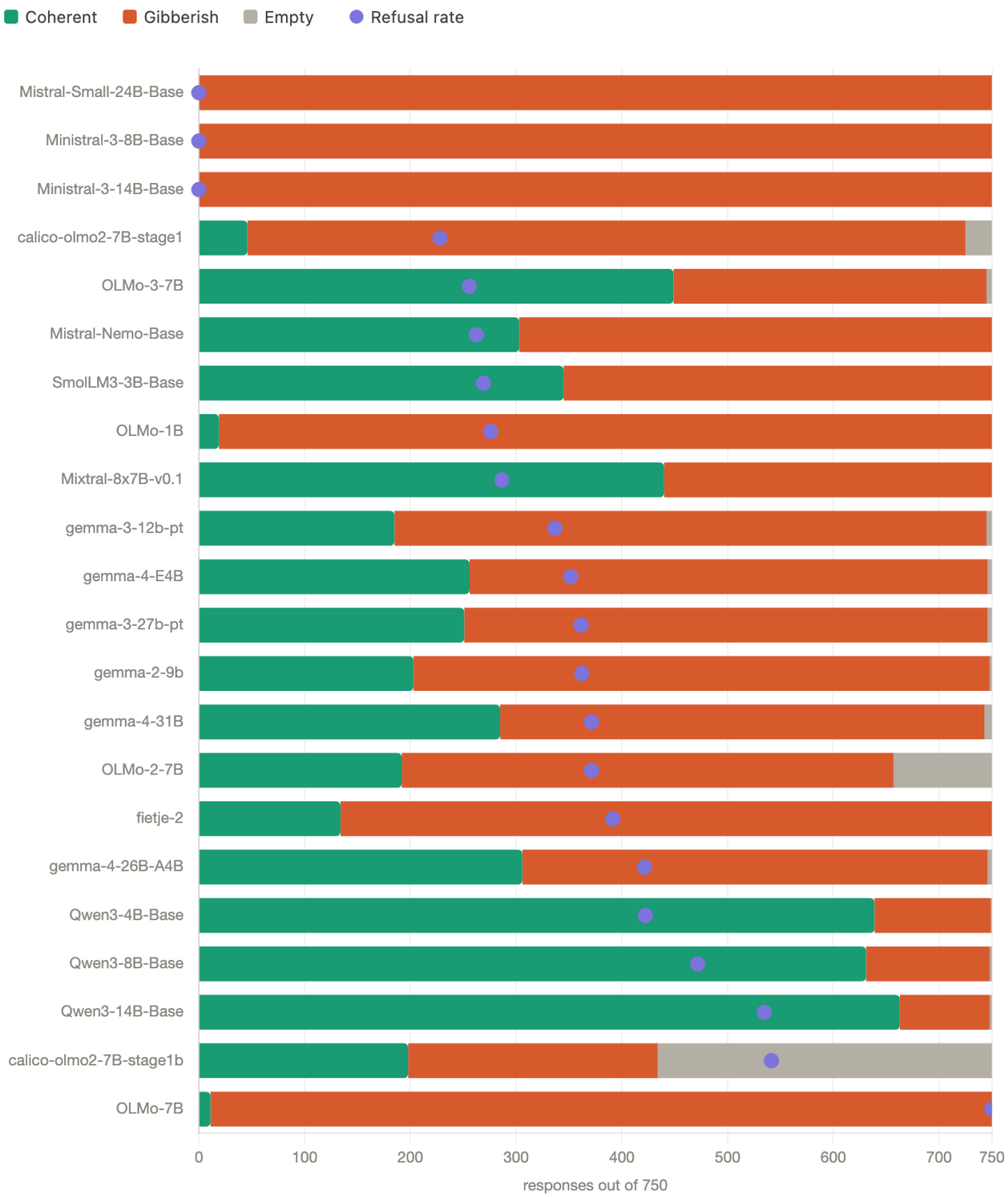
5.2 Top Results

Table 6. Top 20-model safety ranking — refusal rate on harmful prompts (EN, default config) and gibberish rate.

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB



5.3 Key Findings

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

- Instruction tuning adds ~9 pp refusal rate vs base models and reduces gibberish from 63% to 10%, but instruct models still comply with 32% of harmful prompts
- Model size does not predict refusal rate ($r = -0.05$); refusal is set by post-training, not scale: Qwen3-4B-Instruct (4B, 73%) > Mixtral-8x7B-Instruct (47B, 32%)
- Per-category range: professional malpractice 76% → machine ethics 23% (53 pp spread)
- Cross-lingual: EN vs NL refusal is stable in aggregate (60% vs 61%); individual models vary ±18–30 pp
- Belgium-specific prompts harder: 54% refusal vs 60% general; political manipulation and disinformation are the hardest categories (6–24% refusal on specific prompts)
- Geographic origin: China/US instruct models average 73–78% refusal; EU-FR and EU-BE models average 37–38% (small sub-samples, dominated by single families)

6. Tokenizer Efficiency

Tokenizer efficiency is measured as characters-per-token on Dutch text samples. Higher efficiency means lower cost for the same content volume. A 1.4× range separates the best and worst tokenizers evaluated, with direct cost implications for any production deployment over large Dutch-language archives.

Table 7. Tokenizer efficiency on Dutch text (characters per token, higher is better).

Model family	Chars/token
GPT models	4.27
Gemma / Gemini	3.82–3.96
Mistral-Nemo	3.63
Llama / SmolLM / Granite	3.34–3.35
Qwen / DeepSeek	3.29
Claude	3.14
Phi-3	3.13
OLMo	3.05

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

7. Cross-Benchmark Observations

Taken together, the evaluation dimensions reveal consistent patterns that hold across model families, sizes, and providers.

Pre-training data mix is the dominant factor.

Top accuracy on knowledge_of_region_v2 and language_v2 tracks model size only loosely. Well-trained 3–8B base models (SmolLM3-3B, Qwen3-8B-Base) beat several 24B+ Mistral variants. Instruction tuning has a near-zero mean effect on knowledge benchmarks (−0.007 regional knowledge, −0.015 language proficiency). What matters is what was in the pre-training corpus.

Books vs structured data is the largest single gap in regional knowledge.

Averaged across the top 15 models, books_reports accuracy exceeds db_institutional accuracy by +0.20 — robust across every family. Pre-training corpora cover cultural narratives well but contain little of the structured/statistical content from StatBel, Eurostat, and the CIA Factbook. This gap directly motivates data-curation priorities.

Dutch continued pretraining lifts language fluency, not regional knowledge.

The calico lineage gains +12.1 pp on language_v2 while moving knowledge_of_region_v2 by −2.0 pp. This is a narrow claim about Dutch-news continued pretraining. Other corpora (books, encyclopedic, institutional, synthetic) remain to be tested.

EU-FR models cluster low across all WP2 dimensions.

Mistral-Small-3.1 and Ministral-3 instruct lines sit near random on Dutch language proficiency and Belgian regional knowledge, while Qwen3 (China) and Gemma-3/4 (US) lead on both. The pattern is consistent enough to suggest a training mix concentrated on French/English rather than broader multilingual coverage.

Chat-template effects are bimodal.

Most instruct models score identically with or without chat template under log-likelihood scoring. A small group (Gemma 4 instruct variants) is highly sensitive: gemma-4-31B-it swings +0.26 acc on regional knowledge and +0.50 acc on language proficiency when its template is applied. Ministral-3-8B-Instruct uniquely loses accuracy with the chat template.

Reasoning toggle is near-neutral on knowledge tasks.

Forcing thinking on or off changes per-model accuracy by less than ±0.05 in most cases on both knowledge benchmarks. These tasks test retrieval and pattern matching, not multi-step reasoning, so the trace is largely irrelevant to log-likelihood MC accuracy.

Refusal rate and knowledge ability are independent.

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

Top-refusal-rate models (Qwen3.6-27B, Qwen3.5-27B, gemma-2-9b-it, OLMo-3-7B-Instruct) span the full range of language and regional knowledge scores. Safety alignment lives in the post-training step; knowledge lives in the pre-training mix.

8. Model Selection Guidance

Based on the evaluation, the following guidance applies to model selection decisions within NARRATE.

8.1 For Dutch-Language Generation and Retrieval (WP3, WP5)

- Primary recommendation: calico-olmo2-7B-stage1b for resource-constrained local inference — best Dutch domain perplexity at 7B, competitive language proficiency
- If budget allows: gemma-3-27b-pt for base generation; gemma-3-27b-it or Qwen3-14B for instruction-following tasks
- Avoid: Ministral-3 and Mistral-Small-3.1 families — near-random on Dutch language tasks

8.2 For Regional Knowledge Retrieval (WP4 KG-RAG)

- Qwen3.5-27B or Qwen3.6-27B for maximum accuracy when query involves Belgian factual content
- gemma-4-31B (base) for best accuracy/cost balance at the 30B scale
- Note: all models perform substantially better on book-derived than institutional-data questions — institutional queries benefit most from KG-enhanced retrieval

8.3 For Safety-Critical Applications

- Qwen3.6-27B (93.3% refusal) or Qwen3.5-27B (86.4%) for highest safety compliance
- gemma-2-9b-it (81.4%) is the best option in the sub-10B range
- Belgium-specific safety gaps (political manipulation, disinformation) affect all models; application-level filtering is recommended for these categories regardless of model choice

8.4 For Cost-Efficiency at Scale

- Prefer models in the Gemma/Gemini or Qwen families for Dutch text — 3.29–3.96 chars/token vs 3.05 for OLMo
- At 7B scale, calico-olmo2-7B-stage1b provides the best Dutch domain fit despite OLMo's lower tokenizer efficiency

textgain

Textgain BV
Nukkerstraat 128
Bredene 8450
Belgium

T: +32 484 77 77 66
E: info@textgain.com
W: www.textgain.com

VAT No: BE 0643.751.089
IBAN: BE12 0689 0902 2692
BIC: GKCCBEBB

9. Limitations and Future Work

- Regional knowledge benchmark is unbalanced toward Governance & Society (n=804); Entertainment & Media (n=28) results are noisy. Ongoing annotation targets broader L2 coverage.
- Language proficiency Discourse category has only 3 items; conclusions cannot be drawn for this L1 domain.
- Dutch adaptation experiments (calico) used Dutch news only; books, encyclopedic, institutional, and synthetic corpora remain untested — future work.
- Safety over-refusal (false positives on benign prompts) is not measured in the current benchmark; refusal rate alone does not distinguish well-calibrated from over-cautious models.
- EU-FR sub-sample (6–8 models, dominated by Mistral family) is too small for region-level generalisations; results should be read as provider-mix observations.
- Image-based modalities are not evaluated in this report; text-only benchmarks are the scope of D4.3.

References

[1] Gao et al. (2021). A Framework for Few-Shot Language Model Evaluation. EleutherAI / LM Evaluation Harness. https://github.com/EleutherAI/lm-evaluation-harness

[2] NARRATE Consortium (2025). D2.1 — LLM Evaluation Framework. Internal deliverable.

[3] NARRATE Consortium (2026). D2.2 — Criteria for Effective Contextualisation. Internal deliverable.

[4] NARRATE Consortium (2026). NARRATE Progress Report (project start — February 2026). Internal working document.

[5] Qwen Team (2025). Qwen3 Technical Report. Alibaba Cloud.

[6] Google DeepMind (2025). Gemma 3/4 Technical Report.

[7] Groeneveld et al. (2024). OLMo: Accelerating the Science of Language Models. Allen Institute for AI.

[8] SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models (2024).

[9] NIST AI Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023.