AVIGNON
UNIVERSITÉ

Evaluating LLM-Generated Wikipedia Content

Political Topics in a French Setting

Presented by **Jeanne Vermeirsche**, **Eric San Juan**, and **Tania Jiménez**
Avignon Université | LIA & JPEG | France

NLDB_2026

Structure Fédérative de Recherche Agorantic

Table of Contents

Autonomous Knowledge Generation?

The Closed-Book Challenge

Isolating LLMs from real-time data allows us to inspect their **internalized parameters** directly, probing parametric biases without the corrective mechanism of RAG.

We analyze whether open-weight models can serve as standalone, neutral knowledge bases for politically sensitive tasks.

The "Invisible Editor"

An LLM's internal weights behave as a strong **statistical prior**. Corrective prompts require constant "force" to overcome this implicit baseline.

By discovering weight-level bias, we identify the hidden editorial logic the model applies during generation.

Contrasting Mediation Regimes

Wikipedia Regime

Horizontal and deliberative. Fully traceable history and public talk pages make disagreements accessible, auditable, and subject to collaborative consensus.

Algorithmic LLM Regime

Invisible and non-discursive. Instantly outputs highly fluent summaries, but completely lacks structural metadata, negotiation histories, or contextual indicators of dispute.

Wikipedia Community Bias

Demographically unequal (primarily highly educated, male editors) but transparently auditable by the public.

LLM Weight-level Bias

Stored directly within millions of silent matrix parameters. Acts as an invisible, permanent gravitational pull.

The Cascaded Evaluation Pipeline

A robust, language-independent framework transforming surface-level lexical similarities into deep thematic structural alignments.

1. Extraction

GIN queries on 2022 French Wiki snapshot to prevent data contamination.

2. ROUGE

Calculating F1–score to capture basic lexical overlap and topic boundaries.

3. LDA

Building topic probability matrices to smooth stochastics of generation.

4. LSA

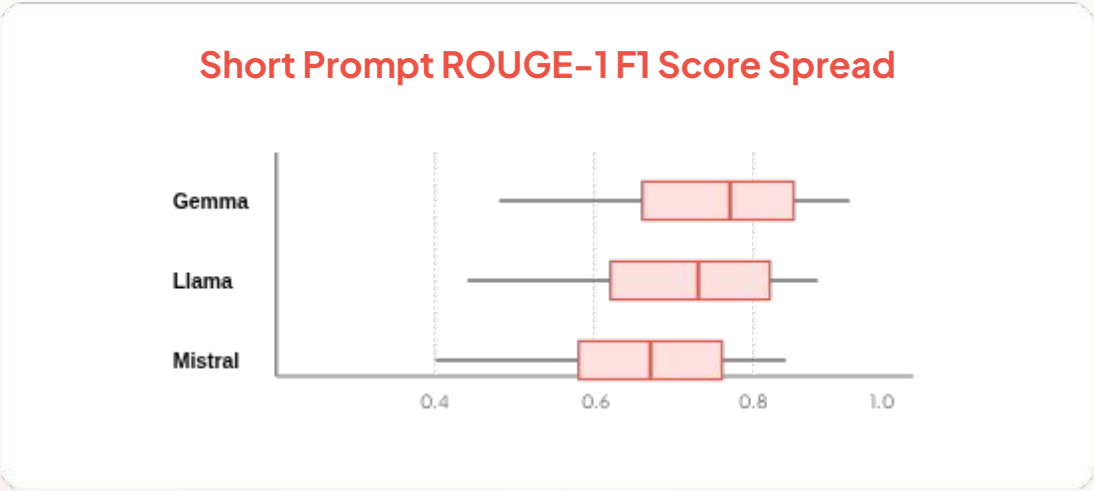
Using SVD to compute and map latent semantic coordinates of political discourse.

Content Quality: ROUGE Metrics

Model Name	Param Size	ROUGE-1 (Recall)	ROUGE-1 (Precision)	ROUGE-1 (F1-Score)	ROUGE-2 (F1-Score)
Gemma	4.3B	1.0000	0.8122	**0.8964**	**0.4574**
Gemma 27b	27.4B	0.9699	0.8318	0.8955	0.4145
Llama	8.03B	0.9004	0.8637	0.8817	0.4554
Mistral	7.25B	0.8354	0.8118	0.8604	0.4093
Qwen	8.19B	0.2360	0.8456	0.3690	0.1868

All generations were truncated to the median length of Wikipedia abstracts (628.5 characters) to ensure size-neutral calculations. Under short prompts, Gemma displays a slight performance edge.

ROUGE Distributions



The French Context Advantage

While models show high variance overall, **Mistral** is the only model that exhibits an increased ability to process long prompts without losing syntactic alignment.

This localized advantage reflects its robust training data curation, allowing it to navigate complex French syntactical instructions effectively.

Proximity: Wikipedia vs Gemma

#	Correlated Term Pair in French Lexicon	Wikipedia Reference Proximity	Gemma Generation Proximity	Alignment Quality
1	*autonomie — indépendance*	0.6178	**0.9961**	Over-correlated (+0.38)
2	*autonomie — peuple*	0.5929	0.9033	Over-correlated (+0.31)
3	*défense — sécurité*	0.5821	0.9951	Over-correlated (+0.41)
4	*démocratique — souveraineté*	0.6890	0.9899	Over-correlated (+0.30)
5	*gauche — union*	0.6633	0.9286	Over-correlated (+0.26)
6	*extrême — gauche*	0.9782	**0.9705**	Highly Accurate (-0.01)

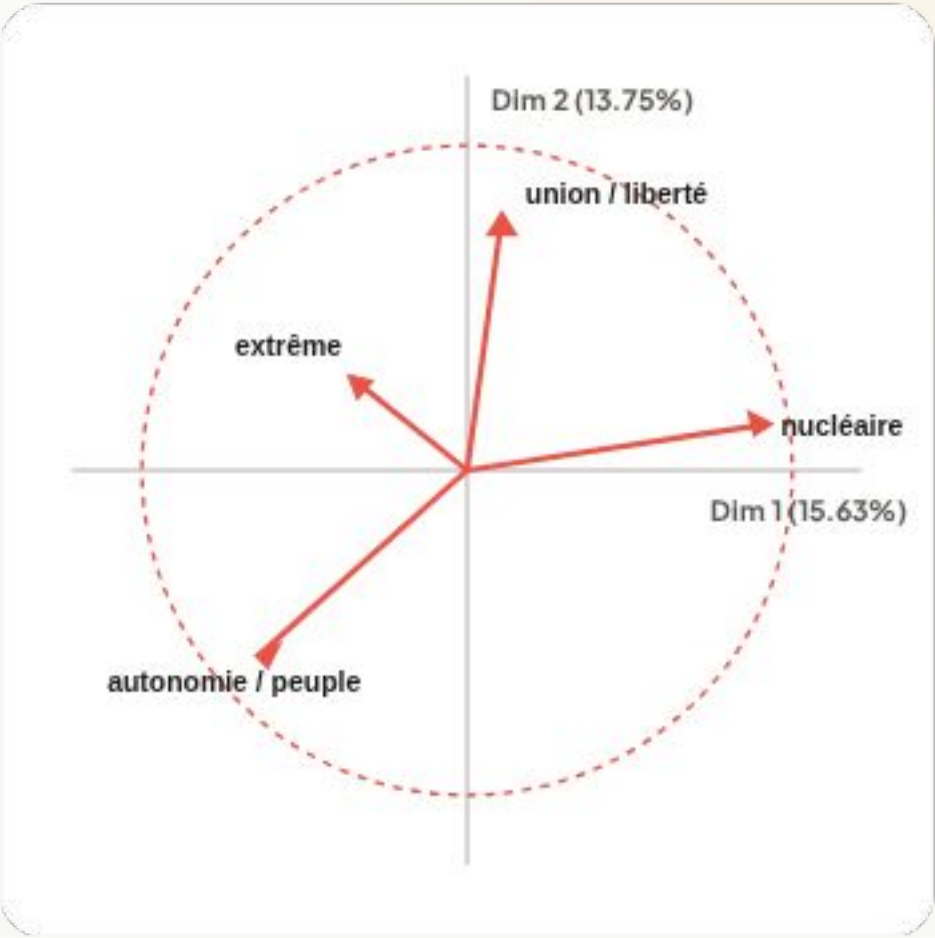
Lexicon PCA Circle

Structural Simplification

This diagram represents the **Wikipedia Reference PCA Circle of Correlations** from Section §6 of the paper.

While original Wikipedia content displays a decentralized web of discourse, LLM weights collapse these associations.

By forcing tighter semantic clustering, parametric bias fundamentally distorts the structural relationships of political concepts.



The Gemma 3B Paradox

9

STABLE PAIRS

Artificially Forced Correlations

On short prompts, the smaller **Gemma 3B model** achieved the highest performance in reproducing significant lexicon correlations, notably outperforming its 27B counterpart. However, qualitative scrutiny reveals that Gemma "forces" relationships—scoring *autonomie-indépendance* at **0.99** compared to only **0.61** in human-written Wikipedia summaries. It creates artificially tight semantic alignments.

Case Study: Mistral Hallucination

⚠ Generated Summary

Subject: Union of French Muslim Democrats (UDMF)

Mistral generates a clean, highly structured, and plausible profile, defining the UDMF as *"a French political party founded in **1990**."* It delineates an idealized set of cooperative narratives, presenting fictional history as settled fact.

✓ Original Wikipedia Reference

The Historical Reality

The UDMF was actually founded in **2012**.

Wikipedia extensively details controversial debates (e.g., accusations of communitarianism vs research rebuttals) which Mistral completely omits in favor of structured, plausible fiction.

Case Study: Gemma "Plausible Filler"

⚠Generated Summary

Subject: Independent Republicans Federation

Gemma writes an incredibly coherent profile, framing it as *"a French far-right party founded in **1944**."*
It invents detailed narratives about its postwar resistance founders and ideological alignments, creating a powerful illusion of accuracy.

Original Wikipedia Reference

The Historical Reality

The federation was actually a prominent **center-right** political movement that existed between **1962 and 1977**.
It served as a key support force for the pro-Gaullist majority. Gemma's generated text is an elegant, complete historical fabrication.

Analytical Scope & Bounds

- **Lexicon Dependence:** The depth of the intrinsic analysis remains bounded by the coverage and design of the expert-defined political lexicon (130 uniterms).
- **Temporal Snapshot:** The March 2022 dataset provides a clean, pre-contamination baseline, preventing the models from having trained directly on post-release articles.
- **Language Independence:** Although tested here on French politics, the underlying cascaded pipeline (ROUGE to LDA to LSA) is completely language-agnostic and reproducible.

Summary: The Parametric Bias Audit

The Challenge

Isolating **internalized parameters** from real-time data identifies the "Invisible Editor" within LLM weights.

The Mediation

Contrasting the **transparent consensus** of Wikipedia with the silent, non-discursive regimes of algorithmic generation.

The Pipeline

A **cascaded framework** (ROUGE, LDA, LSA) to map thematic structural alignments in political discourse.

Key Scientific *References*

Author & Work	Core Contribution	Influence on Paper
[20]**Vermeirsche et al. (2022)** *Atelier TAL-HN (TALN)*	Introduces **LDApol**: a framework to contextualize ideological concepts in politics.	Provides the verified 130–term lexicon used to map latent parameters.
[13]**Lewis et al. (2020)** *NeurIPS Volume 33*	Establishes Retrieval-Augmented Generation (**RAG**) for complex NLP tasks.	Acts as the contrasting baseline setup for evaluating closed-book LLMs.
[19]**Rettenberger et al. (2025)** *J. Comp. Social Science*	Pioneers methodologies for directly measuring political bias inside LLMs.	Guides the formulation of research questions on algorithmic skew.
[14]**Lin et al. (2021)** *TruthfulQA Benchmark*	Measures how model training mimics human misconceptions and falsehoods.	Informs the underlying criteria for detecting "plausible filler" and hallucinations.

AVIGNON UNIVERSITÉ

SFR Agorantic

Questions & Discussion

https://pol.termwatch.eu

NLDB 2026

Avignon Université | JPEG & LIA | France
eric.sanjuan@univ-avignon.fr