A Cross-Lingual Analysis of Bias in Large Language Models Using Romanian History

1 LLM Selection

The models for this study were selected to represent a diverse cross-section of the LLM landscape. Our selection criteria ensure variety across several key dimensions: model architecture (dense vs. Mixture-of-Experts), parameter scale (from efficient small models to state-of-the-art large models), and developer origin (encompassing contributions from major corporate labs and open-source-focused organizations). This diversity is crucial for determining whether observed biases are model-specific phenomena or systemic issues prevalent across the field. The chosen models can be broadly categorized into three groups: large-scale Mixture-of-Experts (MoE) models, high-performance dense models, and specialized models.

1.1 Large-Scale Mixture-of-Experts (MoE) Models

The MoE architecture represents the frontier of scaling LLMs efficiently, achieving massive total parameter counts while only activating a fraction of them for any given input. This allows for vast knowledge storage with manageable computational costs.

- **DeepSeek-R1 and DeepSeek-V3:** Developed by DeepSeek AI, these models are at the forefront of MoE design. While exact parameter counts are proprietary, they are understood to be exceptionally large. **DeepSeek-R1**, in particular, is noted for its advanced reasoning capabilities, which are often attributed to a sophisticated routing mechanism across a large number of specialized "experts." **DeepSeek-V3** continues this trajectory, representing the next generation of this architecture. Their inclusion is vital for testing how a reasoning-focused MoE architecture handles nuanced historical claims.
- **Qwen3-235B-A22B-Instruct:** This flagship model from Alibaba's Qwen series is a prime example of a state-of-the-art MoE implementation. The name itself is revealing: it possesses a total of 235 billion parameters, but only activates approximately 22 billion for any given token. This makes it a computationally efficient powerhouse, and its strong multilingual training is particularly relevant for our cross-lingual study.
- **Llama-4-Maverick-17B-128E and Llama-4-Scout-17B-16E:** These models from Meta represent different configurations within the next-generation Llama family. Both are built on a 17B parameter backbone but are augmented with experts. **Maverick**, with 128 experts ('128E'), offers a vast pool of specialized knowledge, while **Scout**, with 16 experts ('16E'), provides a more focused MoE setup. Comparing these two allows us to investigate how the number of available experts influences response consistency and bias.

1.2 High-Performance Dense Models

This category comprises models with a traditional dense architecture, where all parameters are active for every computation. They serve as powerful baselines and represent the most common type of high-performance model in production.

- **Llama-3.3-70B-Instruct-Turbo:** As one of the most capable dense models from Meta, this 70-billion-parameter model is a critical benchmark. Its performance is considered state-of-the-art for its size class, and its inclusion allows us to compare the behavior of a top-tier dense model against the MoE giants.
- **Kimi-K2-Instruct:** Developed by Moonshot AI, Kimi is renowned for its exceptionally large context window. While its exact parameter count is not public, it is considered a very large dense model. Although our study does not specifically test long-context capabilities, its inclusion is important as it represents a model architecture optimized for a different primary function, which may influence its handling of concise historical prompts.
- **exaone-3.5-32b-instruct:** This 32-billion-parameter model from LG AI Research serves as a representative of a well-resourced, mid-to-large-scale dense model from a major industrial lab, providing another valuable data point in this category.

1.3 Efficient and Specialized Models

This group includes smaller models that are highly optimized for performance and resource efficiency, as well as models that have undergone specialized fine-tuning.

- **Mistral-Small-24B-Instruct:** From Mistral AI, this model is a highly efficient MoE design. Despite its name, it contains a total of 24 billion parameters, but like its larger counterparts, it activates only a fraction per token. It is included to test whether the efficiency-focused MoE approach of Mistral exhibits different bias patterns compared to the larger-scale MoE models.
- **Qwen2.5-7B-Instruct-Turbo:** This 7-billion-parameter model is a smaller, dense counterpart to the larger Qwen3. Its inclusion allows for a direct comparison of architectural scaling effects within the same model family.
- **google/gemma-3n-E4B-it and openai/gpt-oss-20b:** These models represent powerful and efficient offerings from Google and OpenAI, respectively. They serve as crucial benchmarks for highly capable models in the smaller size classes (estimated at ˜4B and 20B parameters).
- **arcee-ai/virtuoso-large:** This model represents a specialized class of LLMs. Developed by Arcee, it is likely based on a powerful open-source foundation model that has been extensively fine-tuned for high-quality, reliable outputs. Its inclusion allows us to assess whether specialized alignment and fine-tuning can mitigate some of the inherent biases found in more general-purpose models.