

ABSTRACT


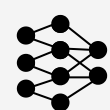
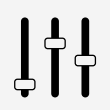





This study investigates the use of large language models (LLMs) to automate LinkedIn content creation for promoting academic research. It compares **GPT-2**, **Flan-T5**, **Llama 3.2**, and **GPT-4o-mini** to determine their ability to generate professional, engaging, and contextually relevant LinkedIn posts. Through parameter optimization and prompt refinement, the study evaluates these models on clarity, depth, engagement, and error-free writing. By identifying the most suitable model for UvA's marketing needs, this project demonstrates the potential of LLMs to streamline content creation, improve efficiency, and enhance research visibility.

PROBLEM STATEMENT

Promoting impactful research on platforms like LinkedIn is essential for fostering societal engagement, attracting funding and building professional collaborations. However, the University of Amsterdam (UvA) marketing team faces significant challenges in achieving this due to limited manpower and growing demands. The rise of artificial intelligence (AI) offers a promising solution by automating content creation, enabling marketing teams to focus on higher-level goals while maintaining quality and relevance. To address these challenges, our project investigates the potential of using LLMs to automate LinkedIn post creation. The research focuses on **identifying** the **most effective model** for this purpose by evaluating the performance of *GPT-2*, *Flan T5*, *Llama 3.2* and *GPT-4o Mini*.

RQ: “Can the open-source models (GPT-2, Flan-T5, Llama 3.2) perform comparably well to the closed-source state-of-the-art models (GPT 4o-mini)?”

MODEL DESCRIPTIONS

 GPT-2	 decoder only	 137M	 Accessible and flexible for general text generation tasks	 Limited context (1024 tokens)
 Flan-T5	encoder-decoder	783M	Summarization, translation, and question answering	Need to have structured prompt
 Llama 3.2	decoder-only	1.24B	Extensive datasets + provide robust context awareness	Need to be fine-tuned
 GPT-4o-mini	proprietary	proprietary	State-of-the-art	Closed-source

EVALUATION AND RESULTS

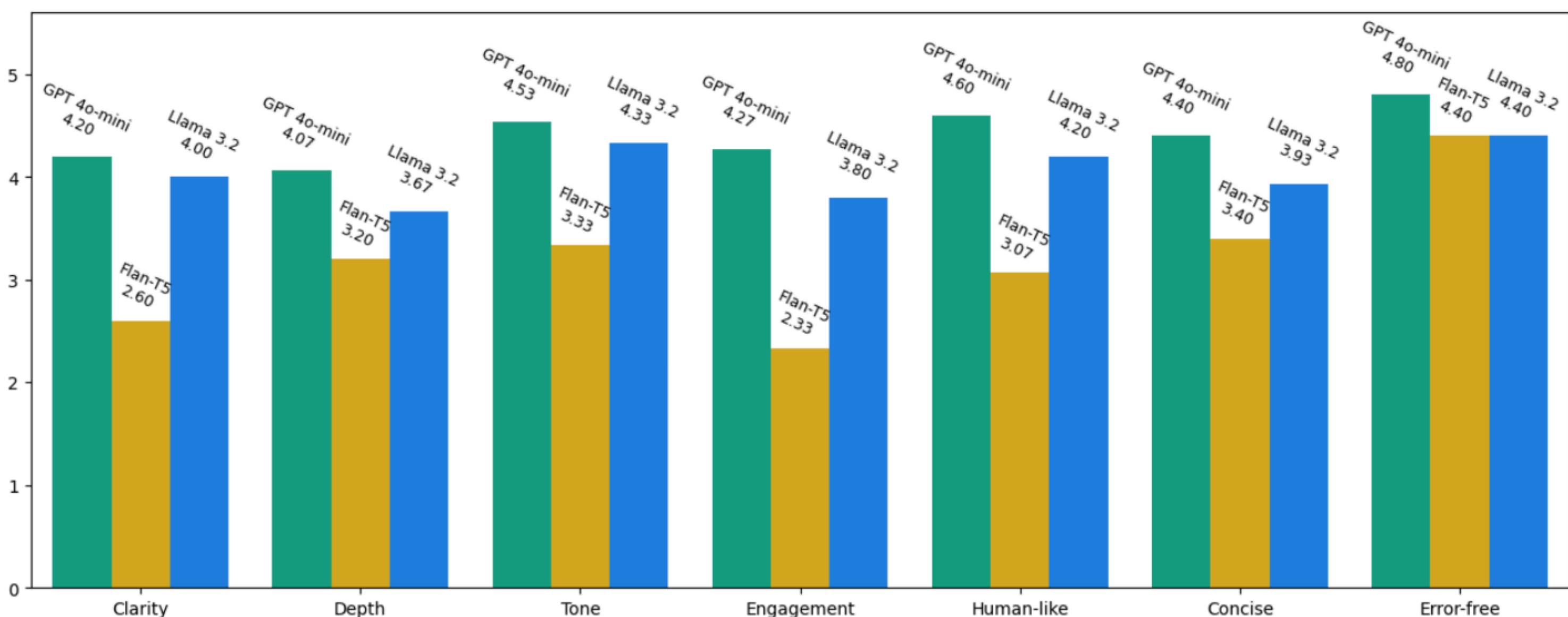
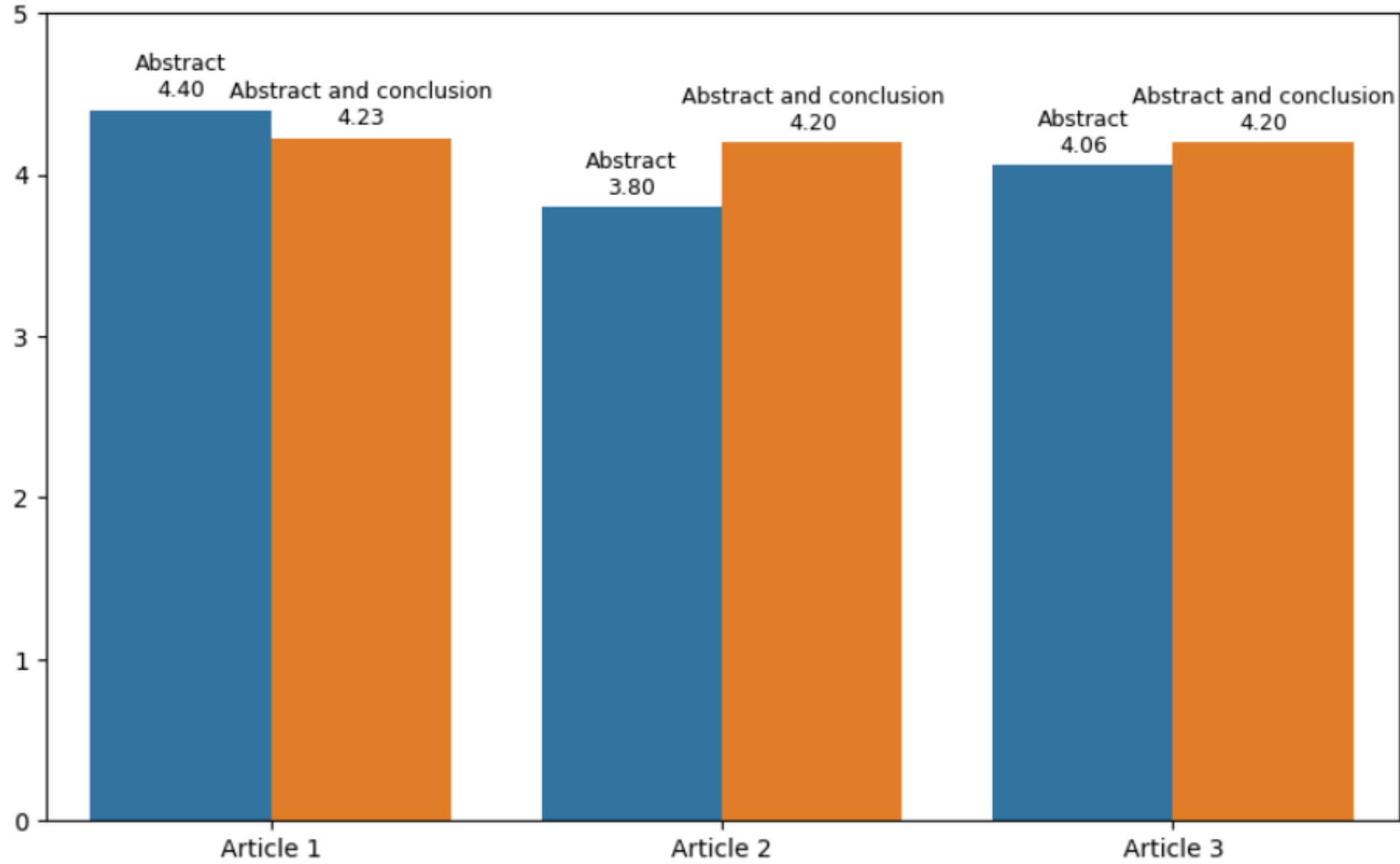
To assess the quality of the generated LinkedIn posts, we conducted a three-stage evaluation process:

Stage 1: Comparing Input Types

We first evaluated whether including both the abstract and conclusion of scientific articles, rather than just the abstract, led to higher-quality generated posts. Five experts rated each post on a 5-point Likert scale (from "strongly disagree" to "strongly agree") across seven key criteria: *Clarity*, *Depth of Content*, *Tone*, *Engagement and Readability*, *Human-like Nature*, *Sticks to the Topic*, *Error-free Writing*. Results from this stage demonstrated that inputting **both** the **abstract** and **conclusion** produced better content (on average: 4.21 vs. 4.09), affirming the importance of comprehensive inputs for better story generation.

Stage 2: Comparing Models

Using the optimal input method (abstract + conclusion), we compared three language models—GPT-4o Mini, Flan T5, and Llama 3.2. Each model generated LinkedIn posts for the same set of articles, which were then rated by the experts using the same criteria. The results showed notable differences in performance: **GPT-4o Mini**: Achieved the highest average score (4.41), Llama 3.2: Scored second (4.05), Flan-T5: Scored lowest (3.26).



Design decisions

- 5-point Likert evaluation scale and final prompt designed with impact story requirements.
- Selecting the open-source models based on size (small, medium, large) and comparing to the SoTA GPT 4o-mini, as it would be the default selection in such scenarios.
- Using only part of the article as input, as larger inputs use more compute/cost more.

Key findings

- Abstract + conclusion slightly outperforms abstract only for impact story generation, although at the cost of additional compute/API token cost.
- Performance increased with model size, where Llama 3.2 output comparable texts to SoTA.
- The results were also confirmed by the stakeholders, confirming the outputs of GPT 4o-mini as the better ones, but also approving aspects of the Llama texts.

Limitations

The LLMs were used on articles from the Amsterdam Centre for Responsible Consumption to generate LinkedIn posts. Therefore the results may differ in other fields or for other purposes. The results also heavily rely on the quality of the abstract and conclusion, which may result in poor impact stories if they are not written up to standard.

Implications

Open-source models at lower scale, like the Llama 3.2 1B, are not significantly outperformed by SoTA models in terms of content generation. Therefore they can be used as a cheaper in-house alternative, which can be fine-tuned to perform better, while still remaining private.

DISCUSSION

FUTURE WORK

Our prototype shows the potential for automating LinkedIn post generation but requires enhancements to improve functionality and scalability:

- Web Scraper**: to **extract** abstracts, conclusions, and author details directly from UvA's research repositories, reducing manual effort.
- User-Friendly Interface**: this would allow **non-technical users** to upload research articles, select models and preview or edit posts efficiently.
- UvA-specific Fine-Tuning**: **fine-tune** the LLMs on UvA-specific LinkedIn data to align content with the institution's tone, style, and branding.
- Platform Expansion**: extending the system's application to **other platforms**, such as Twitter/X or press releases, for broader impact.
- Feedback Loop**: incorporate a **feedback loop** where users rate and refine the model's generated stories over time, ensuring continuous improvement.