Efficient Multilingual Web Summarization Under Latency, Quality, and Cost Constraints

Production-scale web summarization systems must operate under strict latency and cost constraints while maintaining acceptable summary quality across noisy, multilingual inputs. In this work, we propose two practical pipelines – a fast strategy and an advanced strategy, designed to operate under different latency budgets.

We initially hypothesized that extractive methods would be suitable for low-latency summarization, while large language models (LLMs) would provide higher-quality summaries at increased computational cost. To validate this assumption, we benchmarked classical extractive approaches, including TextRank and Lead-TFIDF, alongside LLM-based abstractive summarization. While extractive methods achieved extremely low latency, their summaries were consistently low in semantic quality. Conversely, LLM-based summarization produced substantially better outputs but exhibited prohibitively high latency, often exceeding twelve seconds per request.

Exploratory data analysis revealed that the primary contributor to both extractive failure and LLM inefficiency was extreme input noise. The dataset contained large amounts of boilerplate HTML, navigation elements, and irrelevant repeated text, inflating token counts and obscuring salient information. This observation reframed the problem from one of summarization alone to one of effective input handling.

We first attempted heuristic-based cleaning of the provided markdown content. Although these methods yielded minor improvements, they failed to robustly remove boilerplate across diverse websites and languages. We then experimented with URL-based HTML content extraction using dedicated libraries, which produced substantially cleaner inputs and improved extractive summarization quality. However, the need for live web requests introduced unacceptable latency, making this approach infeasible for high-throughput systems.

A key empirical finding was that model capacity strongly influenced robustness to input noise. More capable models were far less sensitive to noisy inputs, producing high-quality summaries even with minimal preprocessing. In contrast, smaller models and extractive methods were highly sensitive to data quality and required aggressive cleaning to perform adequately. This insight suggested that heavy preprocessing provides diminishing returns for strong LLMs while remaining insufficient to rescue extractive methods under tight latency constraints.

To identify practical operating points, we benchmarked several OpenAI models across latency, cost, and summary quality. The results revealed a clear Pareto frontier. GPT-4.1-mini consistently achieved the highest quality at moderate latency, while GPT-4.1-nano offered an attractive balance between latency and cost but lagged in quality when used directly. Larger models achieved marginal quality gains at prohibitive latency and were excluded from further consideration.

To improve the fast pipeline, we introduced a hybrid compression–abstraction strategy. Input documents were first compressed using a Lead-TFIDF extractive step, and the reduced content was then summarized using GPT-4.1-nano. This compression added negligible overhead (<0.01 seconds) while significantly improving summary quality across all evaluation dimensions. The resulting system achieved approximately 1.1 seconds end-to-end latency while approaching the quality of substantially larger models.

Based on these findings, we selected two final strategies. The fast strategy employs Lead-TFIDF followed by GPT-4.1-nano, delivering strong quality under a strict latency and cost budget. The advanced strategy relies on GPT-4.1-mini, which achieves the highest overall quality with moderate latency and minimal sensitivity to input noise. A latency–quality Pareto analysis confirms that the hybrid approach strictly dominates standalone extractive and direct nano-based methods.

This work highlights the importance of input compression and model robustness in production summarization systems. While preprocessing raw web content remains challenging, lightweight compression combined with appropriately sized LLMs provides an effective and scalable solution. Future work may explore improved boilerplate removal or local models to further reduce cost without sacrificing quality.