LLM-Enabled Semantic Caching for Affordable Web Access

High mobile data costs remain a central barrier to internet access in developing regions, where users rely primarily on bandwidth-constrained devices and pay disproportionately for each megabyte consumed. Over the past decade, the median webpage size has approximately doubled, driven largely by the inclusion of more and higher-quality images. This trend worsens the affordability gap, as traditional caching mechanisms only exploit exact matches and overlook semantic redundancy–cases where different images convey the same meaning within a topical category.

We propose a system for *semantic caching of web images*, in which large language models (LLMs) assess meaning-level similarity and enable server-driven reuse. To evaluate the feasibility of semantic caching on the Web, we conducted a large-scale study across 50 leading news and media websites, collecting 4,264 images and generating over 40,000 inter-article image pairs. This dataset underpins our methodology, which proceeds in three stages:

1. **Dataset construction and annotation:** images were grouped by website-defined categories (e.g., politics, sports, business) to reflect realistic browsing contexts. Human annotators assessed pairwise replaceability based on thematic consistency, article context from the headings and body text, and potential information loss, assigning scores on a 0–4 scale (0 = not replaceable, 4 = completely replaceable).
2. **Model evaluation:** unlike traditional visual similarity methods (e.g., CNN-based feature extraction) that focus on low-level visual cues, our approach employs multi-modal LLMs that integrate image features with contextual signals (alt text, headlines) to capture semantic meaning. We compared two pipelines: **(i) Direct image-to-output:** state-of-the-art multimodal LLMs (GPT-4o, Gemini-1.5 Pro, Claude 3.5 Sonnet) provided replaceability judgments directly from image pairs. **(ii) Two-step description-to-output:** an open-source pipeline combined LLaVA-NeXT for generating detailed image descriptions with LLaMA 3.1 for contextual reasoning, offering a more cost-effective alternative for server-side inference.
3. **Prototype design:** we implemented a client–server system with a new directive, `reuse_similar`. Each client request includes relevant cached image IDs and a user-defined replaceability threshold t. The server indexes a pre-computed similarity matrix: if a cached image meets the threshold t, it issues `reuse_similar` and the client reuses that image; otherwise, the requested image is downloaded. Importantly, all similarity scores are **computed offline at the server side** as a one-time cost (stored for reuse), ensuring no real-time inference overhead and maintaining compatibility with low-resource clients.

Our results show that semantic caching yields notable efficiency gains. Using our prototype to simulate different browsing patterns, the system averaged up to 9.8% greater data savings than exact caching, with improvements reaching $\approx 30\%$ when users repeatedly visited a small set of sites. To generalize beyond these simulations, we developed a probabilistic model that estimates reductions in page weight of about 6.4% for feasible replacements ($t = 1$) and 3.8% for perfect replacements without context loss ($t = 4$), as shown in Figure 1. Figure 2 highlights model performance: proprietary multimodal LLMs yield the highest accuracy, while the open-source pipeline achieves competitive results at much lower cost. Together, these findings validate our approach and motivate our contributions: a large-scale benchmark of semantic replaceability, a proof-of-concept caching protocol, and an evaluation of LLM pipelines for scalable deployment.

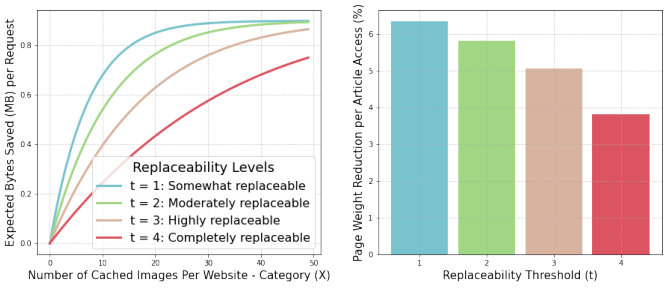

Figure 1: Left: Expected byte savings as a function of relevant cached images. Based on HTTP Archive's median image size of 0.9 MB. **Right:** Reduction in data transfer as a percentage of page weight per article access

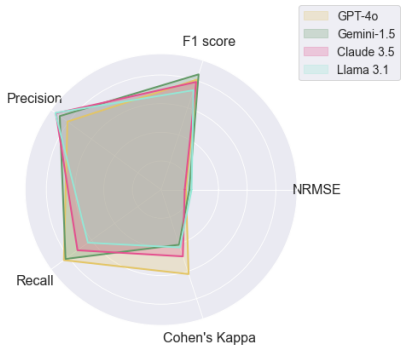

Figure 2: Radar plot comparing model performance for image replaceability.