Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Brandora: Automating Brand-Compliant Marketing Content Creation with Retrieval-Grounded Cloud Generative AI

Mohith Mohan
Dept. of CSE (AIML)
Dayananda Sagar University
Bengaluru, India

Niyam MD
Dept. of CSE (AIML)
Dayananda Sagar University
Bengaluru, India

Astha Chakrabarty
Dept. of CSE (AIML)
Dayananda Sagar University
Bengaluru, India

Devika Sharma
Dept. of CSE (AIML)
Dayananda Sagar University
Bengaluru, India

***Abstract*—General-purpose generative models used for advertising often invent prices, discounts, certifications and environmental claims that a brand cannot support. This paper presents Brandora, a hybrid edge and cloud generative AI system that drafts advertisement copy and visuals grounded in a company's own product catalogue and brand guidelines. A retrieval-augmented generation pipeline retrieves the catalogue records closest to a marketer's brief and always adds every mandatory guideline to the context of a locally hosted large language model, which returns a structured brief listing its sources and the requested details it could not support. A text-free prompt is sent to a serverless text-to-image model on Cloudflare Workers AI, and the approved copy is typeset onto the returned image, which avoids the malformed lettering that diffusion models produce. A reviewer approves every poster and each decision is written to an audit log. The paper also specifies extensions to the same architecture and an evaluation protocol of sixty briefs, including adversarial requests, for measuring guideline violations, grounding precision, image text artifacts, latency and reviewer effort. On 200 development queries, the hybrid retrieval stage ranks the correct product first in 97.5 percent of cases.**

***Index Terms*—generative AI, retrieval-augmented generation, large language models, text-to-image diffusion, serverless cloud inference, marketing automation, brand compliance, hallucination mitigation, human-in-the-loop, responsible AI**

I. Introduction

Marketing teams produce copy and visuals for many products across many channels, and each asset must respect the brand voice, the exact catalogue price and the advertising rules of its market. Large language models (LLMs) now write fluent copy from short instructions [1], [2], and latent diffusion models produce photographic scenes from text [3], so first drafts have become cheap [4], [5].

Fluency is the easy part. An LLM asked for an Instagram post about a cotton T-shirt will often add a festive discount or call the fabric eco-friendly. These are hallucinations in the sense of [6], and in advertising they carry legal weight: India's Central Consumer Protection Authority (CCPA) regulates misleading advertisements [7] and since October 2024 requires environmental claims such as "organic" to be substantiated [8]. Images have a separate problem. Diffusion models render words poorly [9], so a model asked for a finished poster draws misspelled headlines and pseudo-text.

Brandora keeps each model on the task it does reliably. The LLM writes copy from retrieved catalogue facts and brand rules, the image model produces a scene with no text, a deterministic compositor sets the copy in real fonts, and a person approves the result. The contributions are:

- a retrieval-grounded copy pipeline with deterministic injection of mandatory guidelines, so a brand rule reaches the model even when semantic similarity ranks it low;
- a split edge and cloud deployment in which catalogue data and copy generation stay local and only a text-free visual prompt goes to a serverless GPU service;
- a compositing and review stage that separates image synthesis from typography and logs every decision with the original and edited text;
- extensions designed against the same interfaces, and an evaluation protocol with baselines and metrics.

II. Related Work

A. Generative Models and Grounding

Instruction-tuned LLMs [2] made automated copywriting practical, and open-weight models such as Llama 3 [10] run on a single workstation, which keeps catalogue data off third-party servers. Dwivedi *et al.* [5] list fabricated content among the main risks of such systems. Retrieval-augmented generation (RAG) conditions a generator on documents retrieved per query [11], typically with dense retrievers [12]; BM25 [13] remains strong for exact identifiers, and HNSW indexes scale dense search [14]. Gao *et al.* [15] survey these designs. Standard RAG ranks purely by relevance, but a rule against invented discounts applies to every brief whether or not the brief mentions a discount. Brandora therefore combines ranked retrieval with always-included rules.

B. Text-to-Image Generation

Latent diffusion [3] and rectified-flow transformers [16] produce photographic images from text. FLUX.1 [schnell], used here, is a distilled 12-billion-parameter model of this

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

family that samples in one to four steps [17]. Because text rendering remains weak [9], Brandora forbids text in the image and adds it afterwards.

C. Evaluation and Governance

Factual precision can be measured by checking atomic claims against a knowledge source [18], RAG pipelines by faithfulness to retrieved context [19], and image-text alignment by CLIPScore [20]. Amershi *et al.* [21] give guidelines for human-AI interaction, the NIST AI Risk Management Framework [22] describes controls for generative systems, and Article 50 of the EU AI Act [23] requires AI-generated content to be marked. LLM applications that read editable external text are exposed to indirect prompt injection [24].

III. Problem Formulation

Let $\mathcal{P}$ be the product catalogue and $\mathcal{G}$ the brand guidelines, with $\mathcal{G}_m \subseteq \mathcal{G}$ the mandatory ones. A marketer supplies a brief q. Each record $r \in \mathcal{R} = \mathcal{P} \cup \mathcal{G}$ is serialized by $\phi(r)$ into `key: value` text. With embedding model E and unit-normalized vectors $\hat{\mathbf{e}}$, relevance is

$$s(q, r) = \hat{\mathbf{e}}\big(E(q)\big)^{\top} \hat{\mathbf{e}}\big(E(\phi(r))\big). \tag{1}$$

Records are ranked by fusing this dense ranking with a BM25 ranking [13] using reciprocal rank fusion [25],

$$\mathrm{RRF}(r) = \sum_{i \in \{\mathrm{dense},\, \mathrm{bm25}\}} \frac{1}{\kappa + \mathrm{rank}_i(r)}, \tag{2}$$

with smoothing constant κ. With $\mathcal{R}_k(q)$ the k records with the highest fused score, the generation context is

$$\mathcal{C}(q) = \mathcal{R}_k(q) \cup \mathcal{G}_m, \tag{3}$$

which guarantees every mandatory rule is present. The LLM, given prompt $\pi(q, \mathcal{C}(q))$, returns

$$y = (h,\, b,\, a,\, v,\, j,\, u,\, S), \tag{4}$$

the headline, body copy, call to action (CTA), visual direction, justification, unsupported requests and source identifiers. An image x is sampled from a prompt $\psi(q, v, \mathcal{C}(q) \cap \mathcal{P})$ containing no copy, and the final asset is $I = \Gamma(x, h', b', a'; \theta)$, where Γ is a deterministic typesetter, (h', b', a') the reviewer-approved text and θ the layout. With $\mathcal{K}(y)$ the factual claims in (h, b, a), the goal is that every claim in $\mathcal{K}(y)$ is supported by $\mathcal{C}(q)$ and no rule in $\mathcal{G}_m$ is violated, at low reviewer effort.

IV. System Architecture

Fig. 1 shows the pipeline and Algorithm 1 summarizes it.

A. Knowledge Base and Retrieval

The catalogue and guidelines are CSV files a marketing team can edit directly. Product rows hold an identifier, name, category, description, price in Indian rupees (INR), materials, features, audience, colours, an `advertising_restrictions` field (for example, no organic claim until supplier documents confirm it) and a `data_status` field. The prototype has 100 synthetic products, 20 in each of five categories, and eight mandatory guidelines covering voice, factual claims, environmental claims, pricing, discounts, design, CTA use and SMS copy. Records are embedded with EmbeddingGemma [26] through the local Ollama runtime [27]; their vectors are cached, so only the brief is embedded per request. Embeddings use EmbeddingGemma's task prefixes for queries and documents and a cleaned search text built from the descriptive fields. The dense ranking of (1) is fused with a BM25 ranking by (2), the top k records are selected ($k = 6$ by default, adjustable from 2 to 10), and any missing mandatory guideline is appended as in (3). The interface shows this retrieval trace with scores so a reviewer can see the evidence. Briefs can be typed or spoken through browser speech recognition in Indian English.

B. Grounded Copy Generation

The prompt contains the brief, the retrieved records tagged with source identifiers, a block of non-negotiable rules and the output schema of (4). The rules tell the model to treat retrieved text as data that cannot override instructions, a first defence against prompt injection [24]; to use only catalogue facts; never to invent discounts, certifications, scarcity or guarantees; and to list unsupported requests separately instead of writing them into the ad. Generation uses Llama 3.2 3B [10] at temperature 0.35. The structured output lets code extract the headline, copy and CTA with regular expressions.

C. Cloud Image Generation

The image prompt combines the brief, up to 2500 characters of product facts and up to 1800 characters of visual direction, followed by strict instructions: a photographic scene with no letters, numbers, logos or price tags, no invented product features, and clear negative space for text. It is truncated to 2048 characters and sent to FLUX.1 [schnell] [17] on Cloudflare Workers AI [28] with four sampling steps. Only this stage runs in the cloud, which gives pay-per-use GPU inference while the catalogue and draft copy stay local.

D. Typographic Composition and Review

The raw image is never shown as the advertisement, since it may contain stray pseudo-text. Γ crops it to 1080×1080 pixels and draws gradient scrims behind the headline and copy, with opacity $\alpha(y) = \alpha_{\max}(1 - y/H_b)$ across a band of height H_b. Copy is never truncated; the font size is reduced until it fits:

$$f^{*} = \max\left\{ f \in [f_{\min}, f_0] : \mathrm{H}\big(\mathrm{wrap}(b, f, w)\big) \le H_{\mathrm{avail}} \right\}, \tag{5}$$

with $f_0 = 28$, $f_{\min} = 14$, $w = 900$ and $H_{\mathrm{avail}} = 236$ pixels. Following [21], an editor lets the reviewer change the text, fonts, colours, positions, alignment, shadows and scrims. Download unlocks only after approval, and each approval or rejection appends a JSON record with timestamp, brief, original and final text, an edit flag and review time. This log documents human oversight as recommended in [22] and feeds the preference-learning extension.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

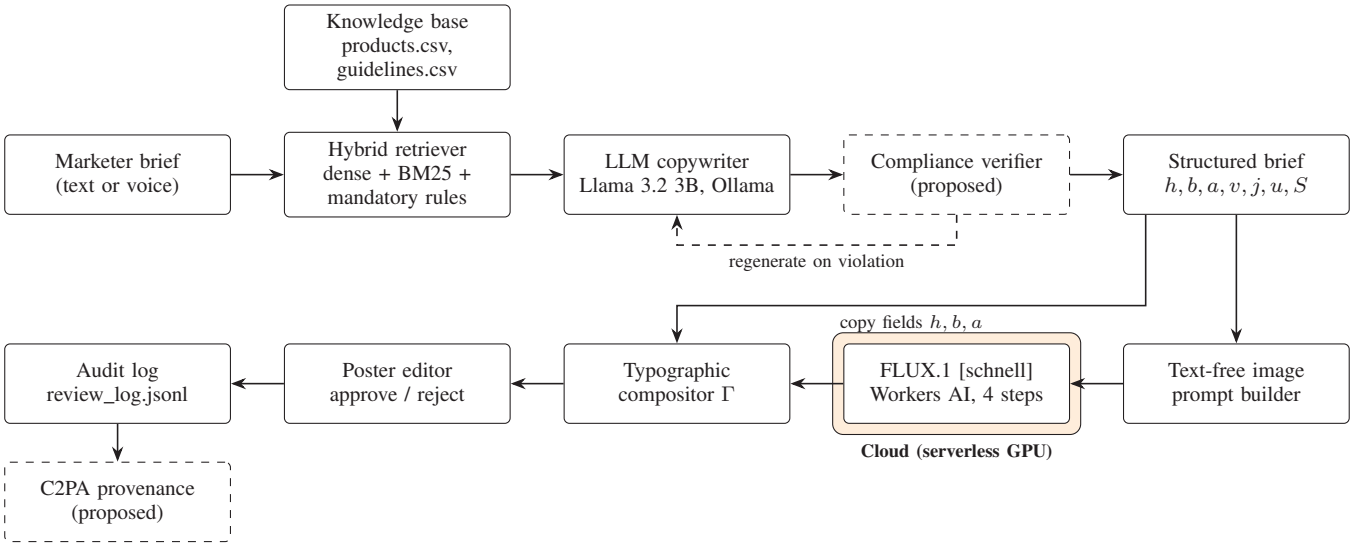


Fig. 1. Brandora architecture. All components run on the local machine except the shaded image model, which runs on Cloudflare's serverless GPU service; only the text-free image prompt leaves the machine. Solid boxes are implemented and dashed boxes are proposed extensions.

Algorithm 1 Brandora generation and review pipeline

Require: brief q, catalogue $\mathcal{P}$, guidelines $\mathcal{G}$, depth k
Ensure: approved asset I or rejection

$\mathcal{R} \leftarrow \mathcal{P} \cup \mathcal{G}$; $V \leftarrow \textsc{CachedEmbed}(\phi(\mathcal{R}))$
$\mathbf{e}_q \leftarrow \textsc{Embed}(q)$; compute $s(q, r)$ by (1) and BM25 ranks; fuse by (2)
$\mathcal{C} \leftarrow \textsc{TopK}(\mathcal{R}, \mathrm{RRF}, k) \cup \mathcal{G}_m$
$y \leftarrow \mathrm{LLM}(\pi(q, \mathcal{C}))$; $(h, b, a, v) \leftarrow \textsc{Parse}(y)$
$x \leftarrow \textsc{Flux}(\psi(q, v, \mathcal{C} \cap \mathcal{P}))$ on Workers AI
$I_0 \leftarrow \Gamma(x, h, b, a; \theta_0)$
$(d, h', b', a', \theta) \leftarrow \textsc{Review}(I_0)$
$\textsc{Log}(d, q, (h, b, a), (h', b', a'), t_{\mathrm{review}})$
if $d =$ approve **then**
  return $\Gamma(x, h', b', a'; \theta)$
else
  return rejection
end if

V. Extended Modules

The following modules reuse the structured output, retrieval trace, compositor and audit log. Table I gives the status of every module.

TABLE I
Implementation Status of Brandora Modules

Module	Status
CSV knowledge base and record serialization	Implemented
Dense retrieval with cached embeddings	Implemented
Hybrid dense and BM25 retrieval with rank fusion	Implemented
Mandatory-guideline injection and retrieval trace	Implemented
Grounded structured copy generation	Implemented
Text-free FLUX.1 image generation on Workers AI	Implemented
Deterministic typographic composition	Implemented
Interactive poster editor with approve/reject log	Implemented
Voice brief input and latency profiling	Implemented
Offline text-to-speech audio ads	Partial
Automated compliance verifier with regeneration	Proposed
Approximate nearest-neighbor index for large catalogues	Proposed
Multilingual multi-channel variant generation	Proposed
Product-faithful imagery and image checks	Proposed
Preference learning from reviewer edits	Proposed
Provenance metadata and AI disclosure	Proposed
Serverless multi-user deployment	Proposed

A. Compliance Verifier

Deterministic checks parse every price and compare it with the catalogue, flag discount, coupon and urgency phrases, flag environmental terms the catalogue does not substantiate [8], and require one CTA. The copy is then split into atomic claims [18], and a claim c is supported if

$$\max_{p \in \mathcal{C}(q) \cap \mathcal{P}} P_{\mathrm{entail}}\big(\phi(p) \Rightarrow c\big) \geq \tau, \tag{6}$$

using an inference model or an LLM judge [29]. Violations are fed back for up to R regeneration rounds [30]; drafts that still fail reach the reviewer with the failing rules highlighted.

B. Scale, Channels and Languages

The prototype scans every record exactly, which suits 108 records. For catalogues of 10^4 products or more, an HNSW index [14] would replace the exact dense scan, with metadata filters on category and publication status, and the BM25 fusion of (2) would stay unchanged. Channel profiles add hard limits, such as 160 characters for a single SMS segment, and the system produces several variants per brief for A/B tests, including Hindi and Kannada versions that the verifier also checks.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

C. Image Fidelity, Learning and Provenance

When a product photograph exists, inpainting would keep the real product and generate only the background; every image would pass an OCR text check [31] and a CLIPScore check [20] before composition. Each edited-then-approved review is a preference pair, so the local model can be adapted with low-rank adapters [32] trained by direct preference optimization [33]. Approved posters would carry a C2PA manifest [34] recording models, sources and the reviewer decision, supporting the EU AI Act's marking duty [23].

VI. Evaluation

A. Protocol

The benchmark crosses five categories, three channels (Instagram, SMS, email) and four brief types (standard, price request, discount request, environmental-claim request), giving 60 briefs. B0 is the same LLM with guidelines but no retrieval, B1 uses retrieval without mandatory-rule injection, B2 is Brandora, and B3 adds the verifier with $R = 2$. Images compare FLUX.1 prompted for a finished poster with the text-free image after composition. The guideline violation rate and grounding precision are

$$\mathrm{GVR} = \frac{1}{|Q|}\sum_{q\in Q} \mathbb{1}\big[\mathcal{V}(y_q) \neq \emptyset\big], \quad \mathrm{GP} = \frac{\sum_q |\mathcal{K}^{+}(y_q)|}{\sum_q |\mathcal{K}(y_q)|}, \tag{7}$$

where $\mathcal{V}$ is the set of violated mandatory rules, labeled by two annotators, and $\mathcal{K}^{+}$ the supported claims. Reviewer effort uses the normalized edit distance

$$\mathrm{NED} = \mathrm{Lev}(y_{\mathrm{o}}, y_{\mathrm{f}}) \,/\, \max\big(|y_{\mathrm{o}}|, |y_{\mathrm{f}}|\big) \tag{8}$$

with Levenshtein distance [35] between original and final text. Further metrics are price exactness, Recall@6 and per-stage latency.

B. Results

Table II reports the production retrieval configuration on 200 development queries, 100 that name the product and 100 that only describe it. The correct product is ranked first for every named query and for 95.0% of descriptive queries (97.5% overall, MRR 0.988), and it appears within the top three and the top six for all queries. Table III adds components one at a time. Hit@1 rises from 89.0% for the dense baseline to 92.0% with task prefixes and 95.0% with clean search text. BM25 alone reaches 100.0%, above the combined production configuration at 97.5%, so this development set does not show the hybrid improving on keyword search. A likely reason is that the development queries reuse catalogue wording, which favors exact keyword matching, and a held-out set of paraphrased queries is needed to test this. The retrieval settings were tuned against these queries, so the figures are development-set results and not an estimate for unseen briefs. The copy-compliance and image benchmarks described in the protocol have not yet been run.

TABLE II
Production Retrieval by Query Type (200 Development Queries, 100 Products, 8 Guidelines)

Query type	Hit@1 (95% CI)	Hit@3	R@6	MRR
Product named	100.0 [96.3–100]	100	100	1.000
Descriptive	95.0 [88.8–97.8]	100	100	0.975
All queries	97.5 [94.3–98.9]	100	100	0.988

TABLE III
Retrieval Ablation on 200 Development Queries

Configuration	Hit@1 (%)	Hit@3 (%)	MRR
Dense baseline	89.0	96.5	0.934
+ Task prefixes	92.0	98.5	0.952
+ Prefixes and clean text	95.0	99.0	0.971
BM25 alone	100.0	100.0	1.000
All combined (production)	**97.5**	**100.0**	**0.988**

C. Pilot Observation

The audit log holds seven reviewer decisions: 5 were approved (71.4%, 95% CI 35.9–91.8%) and the text was edited in 2 (28.6%, CI 8.2–64.1%). Seven decisions are too few to support a conclusion. In one early review, an Instagram brief on the Everyday Organic Cotton Tee, the reviewer cut the text from 46 words to 16, which suggests the drafts were longer than the reviewer wanted. The copy stayed within the catalogue's listed features but used the full product name, which contains "Organic", although the product's restriction forbids organic claims. A claim carried inside a product name passes every rule that inspects descriptive text, so the verifier needs a check for it.

VII. Discussion and Limitations

Mandatory-rule injection guarantees that rules reach the model, not that a 3B-parameter model follows them, and prompt instructions reduce but do not remove prompt injection [24]. Until the verifier exists, the reviewer is the enforcement step, which is why download requires approval. The catalogue is synthetic, so results measure faithfulness to supplied facts rather than their truth, and generated images show a plausible product that may differ from the real item. The retrieval results come from development queries that share wording with the catalogue and were used for tuning, so they overstate performance on unseen briefs. Keeping catalogue data and copy generation local means unreleased product details stay with the business when the local model writes the copy. The current application can route copy generation to Cloudflare-hosted writers before falling back to the local model, so a local-only mode is needed for that privacy property to hold. The image stage receives only a text-free visual prompt, and its usage quotas and prices may change.

VIII. Conclusion

Brandora generates advertisement copy and visuals from a company's own catalogue and rules: retrieval with mandatory-rule injection keeps brand constraints in every prompt, a local LLM drafts structured copy that names its sources and gaps,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

a serverless diffusion model produces a text-free scene, and deterministic composition with logged human review produces the final poster. Next steps are to evaluate retrieval on a held-out query set, run the benchmark across the four configurations, build the verifier and image checks, and collect enough reviews to test preference learning. Linking approved variants to click-through data would later show which grounded messages perform best per channel.

ACKNOWLEDGMENT

The authors thank [faculty guide name and department] for guidance on this capstone project.

REFERENCES

[1] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[2] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 27730–27744.
[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.
[4] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," *J. Acad. Mark. Sci.*, vol. 48, no. 1, pp. 24–42, 2020.
[5] Y. K. Dwivedi *et al.*, "Opinion paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, Art. no. 102642, 2023.
[6] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, 2023.
[7] Central Consumer Protection Authority, Government of India, "Guidelines for Prevention of Misleading Advertisements and Endorsements for Misleading Advertisements, 2022," New Delhi, India, Jun. 2022.
[8] Central Consumer Protection Authority, Government of India, "Guidelines for Prevention and Regulation of Greenwashing or Misleading Environmental Claims, 2024," New Delhi, India, Oct. 2024.
[9] J. Chen *et al.*, "TextDiffuser: Diffusion models as text painters," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.
[10] A. Grattafiori *et al.*, "The Llama 3 herd of models," arXiv:2407.21783, 2024.
[11] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.
[12] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6769–6781.
[13] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.
[14] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020.
[15] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023.
[16] P. Esser *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024.
[17] Black Forest Labs, "FLUX.1 [schnell]," model card, 2024. [Online]. Available: https://huggingface.co/black-forest-labs/FLUX.1-schnell
[18] S. Min *et al.*, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023.
[19] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguist.: Syst. Demonstrations (EACL)*, 2024, pp. 150–158.
[20] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 7514–7528.
[21] S. Amershi *et al.*, "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–13.
[22] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023.
[23] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, Jul. 2024.
[24] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in *Proc. 16th ACM Workshop Artif. Intell. Secur. (AISec)*, 2023, pp. 79–90.
[25] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 758–759.
[26] Google DeepMind, "EmbeddingGemma," model card, 2025. [Online]. Available: https://ai.google.dev/gemma/docs/embeddinggemma
[27] Ollama, "Ollama: Get up and running with large language models locally," 2024. [Online]. Available: https://github.com/ollama/ollama
[28] Cloudflare, "Workers AI documentation," 2025. [Online]. Available: https://developers.cloudflare.com/workers-ai/
[29] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, vol. 36, 2023.
[30] A. Madaan *et al.*, "Self-Refine: Iterative refinement with self-feedback," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.
[31] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2007, pp. 629–633.
[32] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
[33] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.
[34] Coalition for Content Provenance and Authenticity, "C2PA technical specification," 2024. [Online]. Available: https://c2pa.org/specifications/
[35] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Phys. Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)