

Brand Context for AI-Generated Marketing Copy: A Controlled Feasibility Pilot

Research Report #003

Benjamin Koper

Clippable Research · United States

Version of record: October 15, 2026

<https://clippable.io/research/brand-context-marketing-copy-pilot>

Abstract

Marketing copy must follow product facts, offer restrictions, and brand voice. We report a small feasibility pilot of an evaluation workflow that compares two representations of identical brand information: flat key-value text and structured JSON. Ten synthetic briefs, spanning five fictional brands and two tasks per brand, produced 20 text responses using the same model and provider, with one attempt per condition. All 20 responses passed the specified schema and body-length checks. Provider-reported inference charges totaled USD 0.11129. Median end-to-end request times were 4.253 seconds for flat context and 4.3855 seconds for structured context. Independent human review has not been completed. These observations establish that this particular pilot executed; they do not establish an advantage in factual accuracy, creative quality, marketing performance, or application-layer capability. We document the controls, reproducibility artifacts, and requirements for a subsequent confirmatory study.

Publication status

Preliminary technical report, October 15, 2026. Independent human quality review is pending. This is not a peer-reviewed publication and does not establish that Clippable or either prompt format produces better marketing results. The run took place October 14, 2026 in America/Los_Angeles; recorded timestamps use UTC.

1. Research question and scope

Does the representation of brand information affect the usefulness and factual consistency of generated short marketing copy when the information and model are held fixed? This pilot prepares an answer by testing request execution, output capture, cost accounting, and blinded review export. It does not yet estimate the quality effect. The comparison is deliberately narrower than a Brand Kit versus no context: both conditions receive the same facts and prior feedback.

Prior work has studied generative AI assistance in professional writing [1]. HELM motivates evaluation across multiple dimensions and explicit reporting of coverage gaps [2]. Neither source establishes that a particular brand-context format improves advertising. This pilot applies a controlled representation comparison to short marketing posts and separates operational measurements from creative or commercial outcomes.

2. Data and experimental controls

The AI-assisted synthetic fixture pool contains 50 briefs: ten fictional brands and five tasks per brand. Ten briefs were designated for calibration before generation; the other 40 were not generated. Calibration uses five brands, with one launch post and one feature explanation per brand. Each prompt supplies product, price, included items, audience, availability, supported benefit, voice, prior feedback, prohibited claims, and offer limitations. No customer data, real testimonials, or observed campaign outcomes are used.

The fixtures share a generic starter-set benefit. For example, a fictional coffee subscription is described using the same broad benefit as unrelated categories. This simplification supports testing the workflow but limits realism. The pool is neither a random market sample nor a validated benchmark. Its held-out status does not by itself make it suitable for confirmatory inference.

3. Generation and measurement

The flat condition uses one key-value line per fact. The structured condition serializes the same facts as indented JSON. Both receive the same system instruction, task, output schema, facts, and prior feedback. The requested and returned model identifier was `openai/gpt-6-astra`, routed through OpenRouter to OpenAI. Reasoning effort was low, the completion cap was 800 tokens, and JSON-object response format was requested. No sampling parameter was imposed. Request order was shuffled with fixed seed 20261014. No provider fallback or automatic retry was allowed.

A format-valid response must contain exactly three nonempty string fields: headline, body, and cta. The prompt asks for a body under 70 words; the validator accepts at most 70 whitespace-delimited words. Phrase and numeric triage flags are recorded separately. Such checks can miss paraphrased or implicit unsupported claims and cannot certify factual correctness. Human review against the source facts is required.

Requests were sent through an existing application service. The recorded elapsed time includes the service connection and request overhead; it is not isolated inference latency or the latency of the complete Clippable interface. Reported cost is the provider usage charge returned for each request. It excludes platform subscriptions, hosting, human review, and development effort.

The maximum authorized run was 20 calls and USD 5.00. The runner reserved a conservative USD 4.89136 before execution using then-current price bounds and retained reservations even when actual charges were lower. This reservation is a spending control, not an observed bill. Inputs, configuration, code hashes, and request ordering were frozen locally before the first paid request. This was not a public preregistration.

4. Observed execution results

Flat context: 10 attempted, 10 format-valid, zero execution failures; median elapsed time 4.253 seconds; reported inference charge USD 0.05212.

Structured JSON: 10 attempted, 10 format-valid, zero execution failures; median elapsed time 4.3855 seconds; reported inference charge USD 0.05917. Independent human ratings are pending for both conditions.

Descriptive execution measurements, with ten calls per condition. Format validity is not content quality. Cost records were present for all 20 responses. All requested calls are included; no execution failures or replacements occurred in this run.

Automated triage emitted no prohibited-phrase or unrecognized-number flags. This is a statement about the narrow checks, not a finding that every claim was correct. No human ratings have been recorded. Accordingly, preference rates, factual-error rates, quality differences, significance tests, and marketing lift are not reported. The timing and cost differences are descriptive observations from one small run, not estimates of a stable causal effect.

5. Human evaluation and confirmatory study

The prepared reviewer packet displays source facts and paired outputs with randomized left/right

assignment. It hides condition names, provider information, timing, and costs. The private condition key is kept outside the reviewer packet. Reviewers are asked to select left, right, tie, or neither usable, and to score factual accuracy, usefulness, and clarity from 1 to 5 for each output. The planned panel is three independent reviewers per pair, working separately before discussion. Their roles and conflicts must be disclosed. The founder is not counted as an independent reviewer.

After calibration review, a new protocol should fix realistic category-specific facts, task difficulty, obsolete-information conflicts, unsupported-offer traps, and the minimum practically useful effect. A power calculation must determine the confirmatory sample size; the prepared pool of 40 briefs is not a sample-size justification. Aggregate reviewer choices within briefs and account for briefs sharing brands. Report uncertainty, disagreement, ties, neither-usable outcomes, failures, and exclusions. Do not silently change the frozen pilot after observing its outputs.

6. Limitations and implications for marketing

This study uses five synthetic brands, two task types, one model, one provider, one attempt per condition, and no independent quality labels. It does not test images, video, clipping, persistent memory, retrieval, full production workflows, load, or publishing. It cannot show that Clippable outperforms a base model, that structured context is better than flat text, or that outputs increase views or conversions. Generalization would require independent replication with realistic briefs and other models.

A product study should compare completed Clippable workflows with an equally informed baseline, matching source assets and generation budgets, while measuring final artifact quality, failure rate, and editing time. A marketing-performance study would additionally need randomized delivery opportunities and controls for platform, audience, account, timing, and spend. Views across unrelated posts are not causal evidence. This pilot performs no model training or reinforcement learning.

7. Reproducibility, disclosure, and status

The local artifact set contains the pre-run manifest, synthetic briefs, source code, request ledger, execution summary, blank rating template, blinded review interface, and control tests. The manifest remains an immutable planning snapshot; execution status comes from the ledger and summary. Six control tests passed before the run. They verify budget admission, dry-run behavior, model checks, resume refusal, reviewer completeness, and triage behavior; they do not validate marketing quality. This preliminary technical report has not been peer reviewed. Independent human quality evaluation remains pending. No DOI is assigned to this web release.

Benjamin Koper is affiliated with Clippable Research and has a commercial interest in Clippable. AI assistance was used to prepare synthetic fixtures and study infrastructure, analyze recorded execution, and draft this report. Independent human evaluation remains outstanding. No confidential customer data or new model weights are released. The underlying source and generation ledger are retained locally; they are not yet a public reproducibility release. ORCID: <https://orcid.org/0009-0005-6829-7159>.

References

1. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192. <https://doi.org/10.1126/science.adh2586>
2. Liang, P., et al. (2023). Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. <https://arxiv.org/abs/2211.09110>
3. NeurIPS. Paper Checklist Guidelines. Accessed October 15, 2026. <https://neurips.cc/public/guides/PaperChecklist>