NAIAD: INITIATE DATA-DRIVEN RESEARCH FOR LLM ADVERTISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconciling platform revenue with user experience in LLM advertising motivates a data-centric foundation. We introduce **NaiAD**, the comprehensive dataset for LLM-native advertising comprising $58,999$ carefully constructed ad-embedded responses paired with user queries. NaiAD is organized around theoretically grounded evaluation metrics that separately and comprehensively capture user and commercial utility. To mitigate the dimensional collinearity of aligned LLMs, we propose a decoupled generation pipeline that produces structurally diverse samples, ranging from responses that explicitly disentangle stakeholder utilities to responses that are uniformly strong or weak across dimensions. We further provide score labels calibrated by a Variance-Calibrated Prediction-Powered Inference (VC-PPI) framework, aligning automated scoring with human annotations. Empirical reasoning patterns analyses reveal that successful ad integration relies on reasoning paths that cluster into four distinct semantic strategies. Models leveraging NaiAD integrate these strategies to simultaneously improve user and commercial utility, while enabling target-conditioned control over these distinct objectives via in-context learning. Together, these results position NaiAD as a foundational infrastructure for developing future LLM-native ad systems.

1 INTRODUCTION

Recently, the integration of advertising into Large Language Model (LLM) responses by prominent AI organizations has marked a pivotal shift in the generative AI ecosystem (OPENAI, 2026). While approached with careful design and sophisticated fallback mechanisms, this initiative has encountered measurable user hesitation (Curry, 2026; NDTV News Desk, 2025). This friction underscores a fundamental tension: balancing the imperative to monetize capital-intensive models with the user's desire for uninterrupted conversational experiences.

Advertising is not inherently incompatible with LLM-based interaction (Zhao et al., 2025). Generative models hold the promise to enable advertising that is clearly labeled, context-aware, helpful, and unobtrusive, much as skilled human creators naturally integrate sponsorships into their content. Yet prior work largely focuses on the economic and algorithmic aspects of LLM advertising, such as adapting search auction mechanisms to generative settings (Tsai, 2023; Zhao et al., 2023; Shi et al., 2025). Ad integration leaves largely unnatively, treating as external inserts, appended banners or sponsored links (Xu et al., 2026; Hu et al., 2025; OPENAI, 2026; Balseiro et al., 2026), or depending the insufficient advertising capability inherent in LLM themselves. Consequently, the generative quality of the sponsored content itself remains underexplored, and the semantic disconnect disrupts conversational coherence, harms user experience, and ultimately reduces advertising effectiveness.

These observations point to a central bottleneck in generative advertising: the absence of a data-centric foundation for studying, evaluating, and training high-quality conversational ads. Effective LLM advertising requires models to learn how different integration strategies affect user utility, perceived naturalness, and platform revenue. Such learning necessitates a dataset satisfying two key desiderata. (1) **Multi-dimensional, human-calibrated assessment.** Sponsored responses should be evaluated across decoupled axes reflecting multiple stakeholders: *User Utility*, measuring whether the user's query is answered naturally and coherently; and *Commercial Utility* for advertisers and platforms, assessing if brand information is effectively integrated and likely to drive user engagement. (2) **Structural diversity.** The dataset should contain diverse and disentangled samples. For instance,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

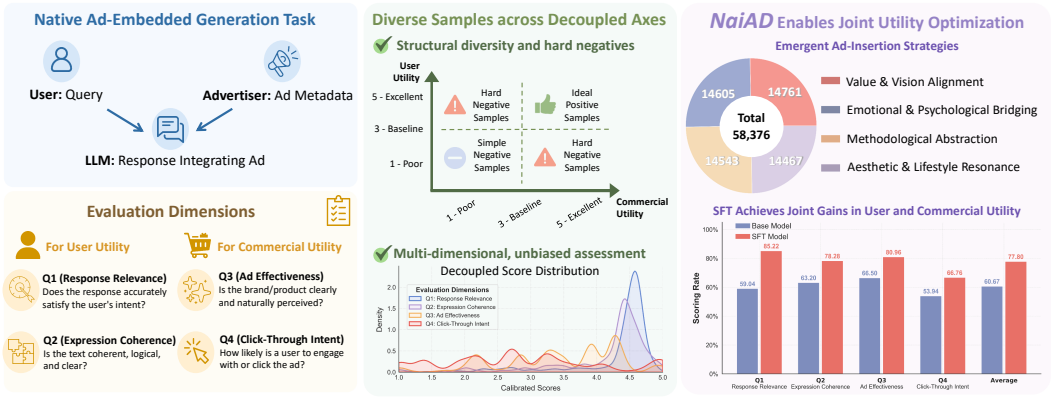


Figure 1: **An overview of our NaiAD dataset. (Left)** The task and the four decoupled evaluation dimensions. **(Middle)** Illustration of our data-centric methodology, emphasizing the generation of structurally diverse samples including "hard negatives" to break dimensional collinearity. **(Right)** Our main findings: (Top) The discovery that LLM ad-insertion behavior converges into four emergent strategies. (Bottom) Empirical validation showing that Supervised Fine-Tuning (SFT) on NaiAD enables a base model to achieve significant joint gains across all utility dimensions.

high user utility does not automatically imply high commercial value. The dataset should therefore include controlled "hard negatives" spanning different combinations of user and commercial quality, preventing both generators and evaluators from exploiting spurious correlations.

We introduce **NaiAD** (Native Ad Integration and Assessment Dataset), the first comprehensive dataset for LLM-based native advertising. NaiAD is designed around the two desiderata and comprises $58,999$ carefully constructed ad-embedded responses: $58,376$ LLM-generated responses and 623 YouTube-sourced human responses. Grounded in Jakobson's semiotic communication theory (Jakobson, 1960) and Austin and Searle's speech act theory (Austin, 1975; Searle, 1969), we evaluate these responses across four decoupled metrics: *Response Relevance* and *Expression Coherence* (measuring User Utility), alongside *Ad Effectiveness* and *Click-Through Intent* (measuring Commercial Utility). However, dataset construction in this setting is challenging because of dimensional collinearity: aligned LLMs tend to produce outputs that are uniformly strong or weak across evaluation metrics. To mitigate this, we design a multi-dimensional decoupled generation pipeline to produce systematically diverse "hard negatives". Furthermore, to evaluate at scale without prohibitive annotation costs or evaluator bias, we introduce a Variance-Calibrated Prediction-Powered Inference (VC-PPI) framework, utilizing a human-annotated subset to statistically align LLM-based evaluations with human judgment.

Using NaiAD, we can reconcile user and commercial utility. Pareto optimality analysis reveals that samples produced by the multi-dimensional decoupled generation pipeline are characterized against YouTube-sourced sponsorship references. Moreover, Supervised Fine-Tuning (SFT) on high-quality NaiAD subsets enables base models to simultaneously improve user utility and Click-Through Intent (CTI). Furthermore, In-Context Learning (ICL) experiments demonstrate that the dataset's structural diversity empowers LLMs with decoupled controllable generation, enabling them to target-conditioned adjust user and commercial utility to meet specific multi-dimensional target profiles. Empirically, we uncover that successful native ad integration relies on constructing a "Logical Bridge": an reasoning path linking the user's query to the advertisers' core value. Analyzing thousands of such bridges reveals an emergent low-dimensional structure where ad integration strategies converge into four semantic clusters.

Figure 1 summarizes our dataset and key findings. In summary, our main contributions are three-fold:

- We release **NaiAD**, the first decoupled quality profiles dataset for LLM ad-embedded generation, constructed via a decoupled generation pipeline and evaluated using our VC-PPI calibration framework to reduce dimensional collinearity and evaluator bias.
- Our empirical analysis reveals four successful generative strategies behind native LLM advertising: embeddings of the "Logical Bridges" used to harmonize user and commercial utility converge into four distinct clusters in a low-dimensional semantic space.

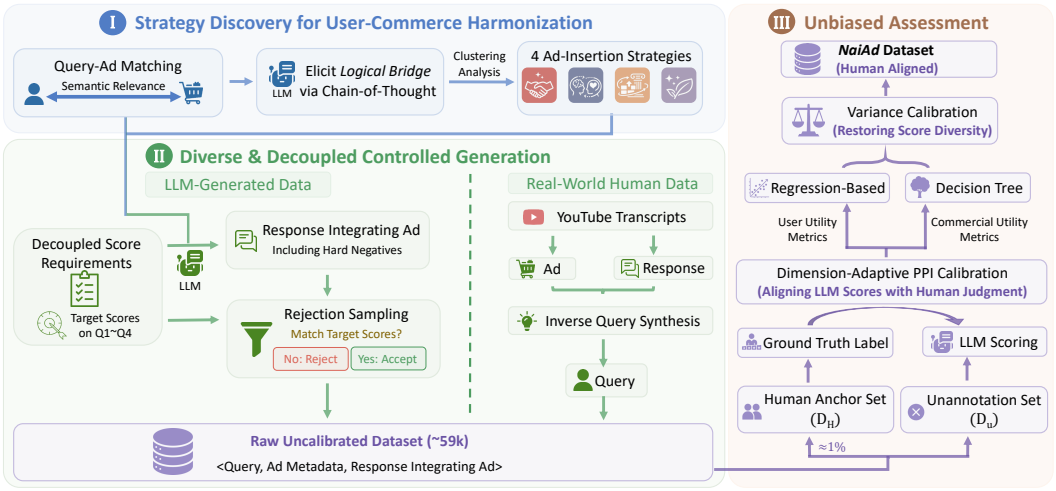


Figure 2: **The construction and calibration pipeline of NaiAD. Phase I:** Eliciting and clustering LLM reasoning paths to discover four core ad-insertion strategies. **Phase II:** A decoupled generation phase creating structurally diverse raw data via target-constrained rejection sampling for LLMs, running parallel to inverse query synthesis for human transcripts. **Phase III:** A Dimension-Adaptive Prediction-Powered Inference (PPI) and variance calibration framework that aligns raw LLM evaluations with human judgments to produce the final calibrated dataset.

- We demonstrate through Pareto optimality analysis, SFT, and In-Context Learning that user and commercial utility can be both jointly optimized and target-conditioned controlled, establishing a robust foundation for future generative advertising systems.

2 Empirical Insight: The Logical Bridge and Strategy Convergence

To construct the NaiAD dataset, our primary objective is to generate responses that effectively balance user utility and commercial utility. However, a major challenge immediately arises: mechanically forcing an ad into a conversation inevitably causes user frustration. Before generating the dataset at scale, we must first decipher the fundamental generative mechanism of *how* to produce high-quality, harmonized samples.

We hypothesize that successfully breaking the trade-off between user and commercial utility requires the model to construct a **"Logical Bridge"**—a latent reasoning path identifying a natural semantic intersection between the user's query and the advertiser's value proposition. To harness this, we prompt the LLM on a small batch of data to explicitly articulate these reasoning paths and perform structural analysis to uncover the emergent, heuristic strategies behind native ad integration. In this section, we deconstruct how LLMs reasoning perform this harmonization, using these discovered strategies to guide the generation process of our subsequent dataset.

2.1 Query-Ad Matching and Logical Bridge Construction

To investigate how models construct these bridges, we first establish a foundation of semantic relevance by pairing $1,986$ sampled real-world queries from INFINITY-CHAT (Jiang et al., 2025a) with the most relevant advertisements from the ATVI[1] pool. Pairs are determined using `paraphrase-multilingual-MiniLM-L12-v2` embeddings (Reimers and Gurevych, 2019) and cosine similarity (see Appendix B for details).

Once we have these contextually grounded pairs, we seek to understand *how* the LLM seamlessly connects them. Using a highly capable generative model (detailed in Appendix C), we prompt it to explicitly articulate the Logical Bridge. By forcing the model to plan the transition from user intent to brand value, we transform a hidden process into observable text, ensuring the ad integration is strategically grounded and transparent.

[1] `https://github.com/Agentyzu/MAE-AM`

2.2 Discovering the Four Core Ad-Insertion Strategies

To determine LLM ad-embedding patterns, we convert the generated textual bridges into Sentence-BERT embeddings. We use Principal Component Analysis (PCA) to progressively reduce the vectors into a condensed space, removing noise while preserving semantic groupings. Subsequent K-Means clustering in this space reveals an optimal structure at $K = 4$ favored by a global peak in the Silhouette Score detailed in Appendix D. Replications across generative models and encoders support $K = 4$ as a useful shared resolution: in the six representative configurations reported in Appendix D.2, its Silhouette Score differs from each run's best only by 0.0072–0.0197, and fixed-$K = 4$ partitions show cross-model ARI values up tp 0.5299. This convergence suggests that **LLM ad-insertion behavior converges into four strategies.** Projecting these embeddings into a 2D space via UMAP in 30D space(Figure 3) allows us to categorize these strategies, where overlap in this projection does not by itself establish overlap in the space where clustering was performed.

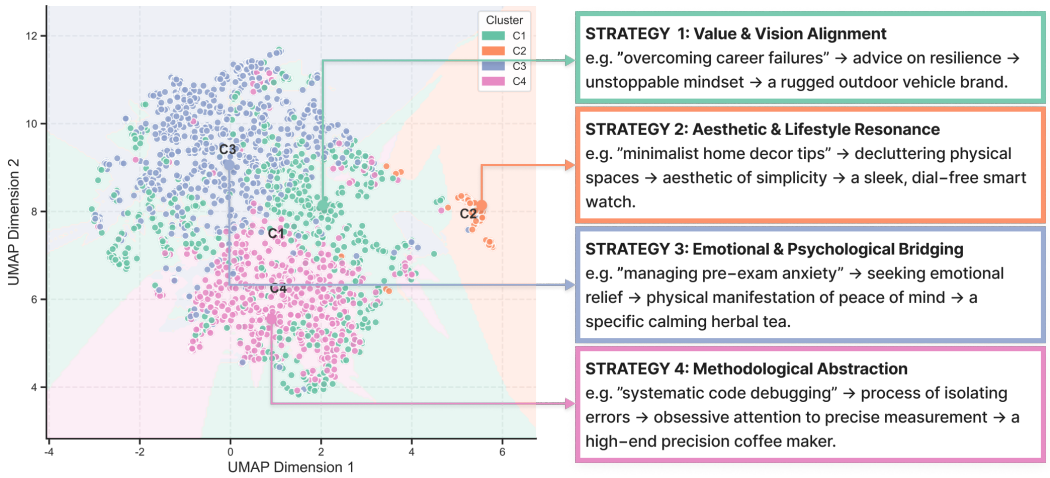


Figure 3: **Latent space visualization of Logical Bridges via UMAP.** The 30D space reveals regions corresponding to the four elicited strategies, illustrating how the LLM shifts its reasoning to natively integrate ads. Because constructing a Logical Bridge inherently requires complex conceptual association, we adopt the structured framework of Association Reasoning Paths (Huang et al., 2025) to represent these transitions as directed multi-hop sequences.

- **Strategy 1: Value & Vision Alignment (The "Mindset" Bridge).** The LLM elevates the query to a macroscopic philosophy, connecting it to an advertiser that shares identical systemic values.
- **Strategy 2: Aesthetic & Lifestyle Resonance (The "Vibe" Bridge).** The LLM bypasses literal topics to match the stylistic or aesthetic constraints of the query with a brand sharing an identical lifestyle identity.
- **Strategy 3: Emotional & Psychological Bridging (The "Empathy" Bridge).** The LLM identifies the underlying emotional driver and positions the product as a tangible means of emotional relief or enhancement.
- **Strategy 4: Methodological Abstraction (The "Craftsmanship" Bridge).** Through cross-domain feature migration, the model isolates the operational rigor required by the user's task and links it to a product characterized by similar precision or craftsmanship.

Our taxonomy is induced from elicited CoT paths rather than prescribed by advertising theory. This parallels the content-analytic typology of Laskey et al. (1989): Methodological Abstraction resembles an informational mode, whereas Value & Vision Alignment, Aesthetic & Lifestyle Resonance, and Emotional & Psychological Bridging resemble transformational modes emphasizing values, identity, and affect. We provide case studies on these strategies in Appendix L to demonstrate.

3 The NaiAD Dataset: Multi-Dimensional Decoupled Generation

Having demystified the underlying generation strategies, we proceed to the core construction of the NaiAD dataset. The end-to-end pipeline for discovering integration strategies, generating diverse samples, and conducting calibrated assessment is illustrated in Figure 2.

Task Formalization. We define the *Native Ad-Embedded Generation* task as follows: given a user query Q and advertisement metadata A, a model must generate a response R that fulfills Q while seamlessly integrating A without breaking conversational coherence.

Constructing a comprehensive dataset for this task requires ensuring both **high-quality semantic integration** and **structural diversity**. Relying on unguided generation often yields generic, low-quality insertions. Furthermore, aligned models suffer from **Dimensional Collinearity (The Halo Effect)** (Jiang et al., 2025a), producing outputs that are uniformly high or low across all metrics, thereby lacking the "hard negatives" needed to train discriminative boundaries. To address these challenges, we utilize INFINITE-CHAT (Jiang et al., 2025a) as the data source (details in Appendix B) and design a multi-dimensional decoupled generation pipeline (see Appendix C for implementation details).

3.1 Strategy-Guided Decoupled Generation

Strategy-guided generation for high-quality data. Directly prompting an LLM to insert an ad frequently results in abrupt, unnatural transitions. To promote the generation of high-quality samples and ensure structural diversity, we incorporate the four core strategies discovered in Section 2 directly into the generation prompts. During generation, the model is conditioned on an assigned strategy and instructed to output both the Logical Bridge and the final response within the same conversational turn. In this setup, the explicitly generated Logical Bridge functions as a specialized Chain-of-Thought (CoT) Wei et al. (2022). It guarantees that the semantic connection between user utility and commercial utility is meticulously planned and executed, significantly improving the coherence of the final response while enhancing the transparency and explainability of the data generation process.

Synthesizing hard negatives via decoupled score requirements. While the strategy-guided CoT ensures the generation of high-quality, synergistic samples, training robust evaluators also requires breaking Dimensional Collinearity. To achieve this structural diversity, we enforce decoupled score requirements across four dimensions (scale $[1, 5]$). (Detailed in Appendix E.) We mandate a minimum score spread for objective dimensions like Relevance and Coherence, while modeling Click-Through Intent as a bounded function of these scores (Appendix F). Combining these discordant constraints with our guided strategies yields a balanced dataset containing fine-grained quality variations.

Overcoming LLM quality bias via rejection sampling. Aligned LLMs are naturally optimized to produce flawless, coherent text, resisting fulfilling the feature-discordant score requirements defined above. To overcome this inherent quality bias and enforce our targeted structural diversity, we employ **Tolerance-based Rejection Sampling** on top of our strategy-guided generation. Generated samples are accepted if and only if their self-evaluated scores meet our target decoupled constraints (see Appendix F), so that in favor of the essential hard negatives required for robust evaluation. The self-evaluated scores used serve only as generation-time selection signals and are discarded after filtering. Retained responses are scored by Qwen3.6-Plus in a separate evaluation pipeline, whereas Claude-4.5-Opus generates the responses; the evaluator scores are subsequently calibrated using 684 human-rated anchor samples (Section 4;Appendix C).

3.2 Incorporating Real-World Human References

While our pipeline successfully generates structurally diverse synthetic data, evaluating true commercial utility requires an authentic real-world anchor. To capture real commercial rhetoric and the natural (often imperfect) distribution of human ad embedding, we incorporate real-world sponsorship data from `Xenova/sponsorblock` (Xenova, 2024). As these transcribed sponsorships lack explicit user queries (Q), they cannot be directly evaluated for "Response Relevance." Thus, we introduce an **Inverse Query Synthesis** mechanism using a state-of-the-art LLM to reconstruct standardized pseudo-queries. This ensures consistent format between YouTube-sourced human data and our synthetic corpus, fulfilling the dataset's task requirements.

4 Human Assessment via Statistical Score Calibration

With the structurally diverse, ad-embedded responses successfully generated, the NaiAD dataset remains incomplete without fine-grained, calibrated quality annotations across our decoupled dimensions. However, evaluating a dataset of this scale ($N \approx 59k$) presents a fundamental dilemma. Relying exclusively on human annotation is prohibitively expensive, whereas employing an uncalibrated "LLM-as-a-Judge" introduces severe systemic biases such as verbosity bias, strictness against

commercial intent, and variance collapse (Zheng et al., 2023). If left uncalibrated, these LLM biases would undermine the carefully engineered "hard negatives" in our dataset.

To achieve human-anchored assessment at scale, we design a hybrid framework based on Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023), using a human-annotated anchor set $\mathcal{D}_H$ ($n = 684$) to calibrate the unannotated set $\mathcal{D}_U$ (detailed in Appendix H). We first utilize a capable evaluator LLM (detailed in Appendix C) augmented with Chain-of-Thought (CoT) (Wei et al., 2022) to generate preliminary, uncalibrated scores for $\mathcal{D}_U$. We then calibrate these scores by estimating and neutralizing their generative biases exclusively through the ground-truth human labels in $\mathcal{D}_H$. Each of the 684 calibration anchors received two independent human ratings (Appendix H.5).

4.1 Dimension-Adaptive Score Calibration

Standard PPI calibration typically applies uniform shifts to correct bias. However, our analysis reveals that the nature of LLM scoring errors differs between User Utility and Commercial Utility dimensions. To address this, we introduce a **Dimension-Adaptive Calibration Mechanism** applying the optimal correction strategy for each category (details are provided in Appendices I and J):

1. Regression-Based Calibration for User Utility Dimensions. Dimensions related to User Utility, such as *Response Relevance* and *Expression Coherence*, assess the foundational informational quality of the generated text. Both human and LLM judgments on these aspects tend to be continuous; improvements in quality often correspond to incremental, relatively smooth increases in perceived value. Given this scalar nature, we use polynomial regression (Ordinary Least Squares, OLS) to model and correct the LLM's scoring error. This model learns a continuous mapping by minimizing the error against human annotations on the anchor set $\mathcal{D}_H$, taking into account the LLM's raw score and its self-evaluation confidence gap.

2. Decision Tree Calibration for Commercial Utility Dimensions. Dimensions of Commercial Utility, such as *Ad Effectiveness* and *Click-Through Intent*, measure the response's persuasive impact and its ability to influence user action. This assessment hinges on a user's decision-making process, which is inherently non-linear and often exhibits threshold-like behaviors (e.g., a user is either persuaded to click or not). To capture these discrete boundaries, we employ Decision Tree models (Fisch et al., 2024). These models excel at segmenting the data into different behavioral strata and calculating distinct bias correction terms, a task for which continuous regression models are ill-suited.

4.2 Restoring Score Diversity via Variance Calibration

While our Dimension-Adaptive Routing isolates the optimal calibrating estimator, conventional prediction-powered rectifiers inherently optimize Mean Squared Error (MSE), leading to *variance compression*—systematically squashing extreme scores toward the median.

To restore the natural distributional span and maintain the score diversity, we propose **Variance-Calibrated PPI (VC-PPI)**. By mathematically matching the first and second moments of the rectified distribution to those of the human anchor set $\mathcal{D}_H$, we forcefully restore the distributional span via a bounded affine transformation:

$$\hat{Y}_{VC} = \text{Clip}_{[1,5]}\left(\mu_H + (\hat{Y}_{route} - \mu_{route})\frac{\sigma_H}{\sigma_{route}}\right) \tag{1}$$

where μ_H and σ_H are the true mean and standard deviation of human scores within $\mathcal{D}_H$, and μ_{route}, σ_{route} are those of the routed predictions $\hat{Y}_{route}$ on $\mathcal{D}_U$.

VC-PPI effectively restores the suppressed standard deviation. The selected routes reduce Wasserstein distance to the human anchor distributions across all four dimensions (Appendix J).

5 Experimental Evaluation and Analysis

The fundamental goal of our research is to buffer the trade-off between user and commercial utility in LLM-native advertising. To systematically demonstrate how our methodology and the NaiAD dataset support this, we formulate our experimental validation around four focal questions: (1) Can we decouple and human-calibratedly assess these conflicting objectives? (Section 5.1) (2) How does our controlled generation perform, and what reasoning mechanisms derive this performance? (Section 5.2) (3) Can fine-tuning on NaiAD teach models to overcome the trade-off? (Section 5.3) (4) Does the inclusion of fine-grained hard negatives in NaiAD enable dynamic, multi-dimensional controllable generation? (Section 5.4)

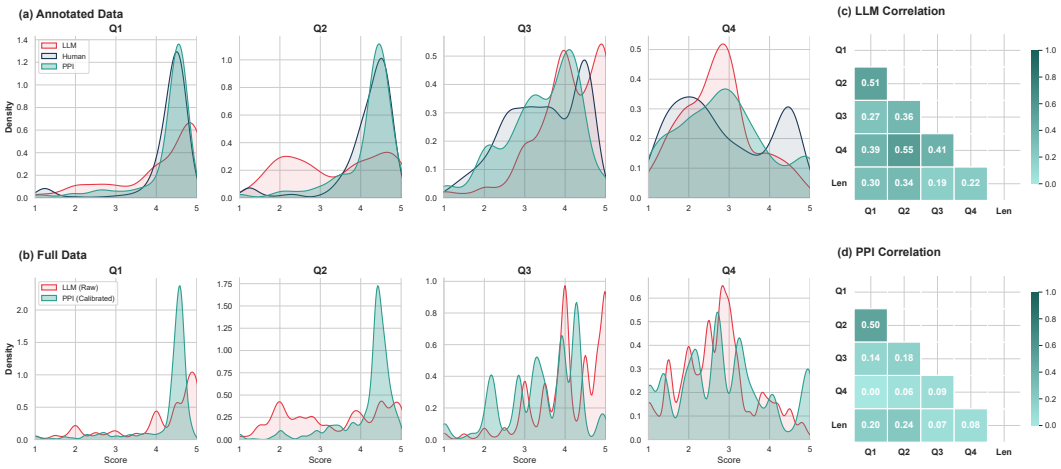


Figure 4: **Impact of PPI Calibration on Scoring Distributions and Dimensional Collinearity. (a & b)** Kernel density distributions reveal that raw LLM scores (red) misrepresent true quality. For user utility metrics (Q_1, Q_2), the LLM artificially spreads scores, whereas PPI (green) recovers the true human consensus (blue) that modern models maintain high baseline coherence. **(c & d)** Heatmaps demonstrate the reduction of spurious correlations.

5.1 Validating Decoupled Assessment via PPI Calibration

Uncalibrated LLM judges are notoriously susceptible to Dimensional Collinearity (the Halo Effect), where strength in one dimension spuriously inflates perceived quality in others. Our PPI framework is designed to mitigate this bias, disentangling conflicting dimensions and aligning evaluations with authentic human judgment.

Distributional Alignment and Behavioral Stratification. As Figure 4(a, b) illustrates, uncalibrated LLM evaluations (red curves) misrepresent true human distributions (blue curves). For objective dimensions (Q_1, Q_2), raw LLMs are overly critical and hallucinate flaws, artificially spreading the scores. Our PPI calibration rectifies this by mapping back to the highly skewed human ground truth. This high-score concentration accurately reflects modern generative models: even when forced to generate structural "hard negatives," they rarely produce fundamentally incoherent or irrelevant text. More discussions on the score distribution characteristics is provided in Appendix K.1.

Breaking Dimensional Collinearity. Figure 4(c, d) demonstrates successful dimension decoupling. The uncalibrated LLM judge suffers from severe coupling, forging a spurious correlation between User (Q_1) and Commercial Utility (Q_4) ($\rho_{Q_1,Q_4} = 0.39$), alongside a noticeable verbosity bias. Following PPI calibration, these cross-domain correlations are depressed (Q_1-Q_4 drops to 0.00), and length bias is suppressed. Crucially, PPI preserves the valid inherent correlation between Relevance and Coherence (Q_1-Q_2, $\rho = 0.50$). This confirms our framework evaluates user and commercial utility as independent axes without destroying fundamental dataset statistics.

Independent human validation. On 50 randomly selected held-out responses with diversity coverage, final VC-PPI scores correlated with independent human ratings (Spearman $\rho = 0.802/0.769/0.850/0.674$ for Q_1–Q_4). In a separate set of 50 dimension-specific contrastive pairs, score-based choices agreed with human preferences in 38 cases, including 12/12 for ad effectiveness (Q_3) and 7/12 for perceived click intent (Q_4) detailed in Appendix I.1.

5.2 Pareto Optimality and Reasoning Strategy of Logic Bridges

Having established a reliable scoring mechanism, we investigate the absolute quality of the generated samples and decipher the linguistic characteristics that distinguish successful soft ads from failures.

Comparison with real-world sponsorship references. We use YouTube sponsorship transcripts as a reference distribution for characterizing NaiAD's offline score profiles (Appendix K.3). In this text-only comparison, NaiAD sampled responses have higher mean Q_3 and Q_4 scores, especially exceeding the human mean in 78.6% of cases for Click-Through Intent, whereas the transcripts have slightly higher Q_1 and Q_2 scores. This statistically confirms that our controlled pipeline is not merely mimicking human behavior but optimizing.

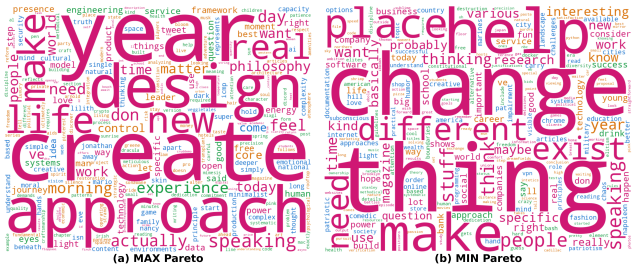


Figure 5: **TF-IDF Word Cloud of Logical Bridges.** Max Pareto optimal bridges (Left) utilize abstract structural vocabulary, whereas Min Pareto failures (Right) regress into disjointed, literal nouns.

Vocabulary Deconstruction of Success. To understand *how* models achieve this optimality, we extract TF-IDF keywords from the explicit "Logical Bridges" of the Max and Min Pareto samples (Figure 5). A clear insight emerges: **successful bridges rely on structural abstraction, whereas failed ones collapse into mechanical objectification.** High-scoring strategies (e.g., *Value & Vision Alignment*) use dynamic, abstract terms to elevate user intent to a higher semantic plane. Conversely, low-scoring bridges rigidly fixate on specific physical entities, triggering awkward, non-sequitur associations that degrade user trust.

5.3 JOINT UTILITY IMPROVEMENTS THROUGH SUPERVISED FINE-TUNING

Unaligned models inherently struggle with ad-embedding, either awkwardly appending ads to good answers or entirely abandoning the user's intent to fulfill the commercial payload. We demonstrate that NaiAD improves this via Supervised Fine-Tuning (SFT) on a highly capable base model (detailed in Appendix C). We utilized a high-quality subset of $10,014$ samples from NaiAD and evaluated against 100 randomly sampled held-out test cases, which are disjoint at the task level.

Table 1: **Macro Scoring Rate Statistics (SFT vs. Base).** Our SFT model achieves simultaneous, substantial gains across all evaluation dimensions on the held-out tasks.

Model \(%)	Q_1 **(Relevance)**	Q_2 **(Coherence)**	Q_3 **(Ad Effect.)**	Q_4 **(Click-Through Intent)**	**Average**
Base Model	59.04	63.20	66.50	53.94	60.67
SFT Model (Ours)	**85.22**	**78.28**	**80.96**	**66.76**	**77.80**
Absolute Δ	↑ +26.18	↑ +15.08	↑ +14.46	↑ +12.82	↑ +17.13
Relative Δ	↑ +44.3	↑ +23.9	↑ +21.7	↑ +23.8	↑ +28.2
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

The macro scoring rate display the percentage of the score out of 5.0 scores, which results (Table 1) validate two core insights regarding the SFT model's transformation:

- **Simultaneous Optimization (Synergy).** The model achieves a massive relative improvement of 28.2% across the board. Crucially, the trade-off is buffered; the model simultaneously elevates Commercial Utility ($Q_3 \uparrow 21.7\%$) and User Experience ($Q_2 \uparrow 23.9\%$).
- **Optimizing Native Integration Capability.** Statistical analysis of sample-level score differences (fully detailed in Appendix K.4) reveals that Response Relevance (Q_1) experienced the highest mean gain ($+1.309$) with a massive effect size (Cohen's $d = 1.148$). This indicates the model learned to natively weaves the ad into a genuinely helpful response.

The in-domain gains are accompanied by minor general capacity fluctuations on three benchmarks (Appendix K.5): 2.25% points on GSM8K, 4.47% on HumanEval+, and 6.02% on AlpacaEval 2.0.

5.4 DECOUPLED CONTROLLABLE GENERATION VIA IN-CONTEXT LEARNING

Beyond foundational capability alignment, advanced generative advertising systems should ideally be able to target specific performance profiles across User Utility and Commercial Utility based on diverse contextual demands. To prove that the decoupled NaiAD has the potential to empower this capability, we conducted an In-Context Learning (ICL) experiment (models detailed in Appendix C).

We tasked the LLM with generating responses targeting exact, decoupled scoring profiles ($[Q_1, Q_2, Q_3, Q_4]$). We focus our explicit control analysis on Q_1, Q_2, and Q_3, while treat Q_4 (Click-Through Intent) as an emergent outcome, because in real-world advertising, Q_4 is a highly subjective user decision that is inherently downstream of the response's relevance and the naturalness of the ad integration. We compared a Zero-Shot baseline against a 10-Shot ICL approach, sampling exact-match target configurations directly from NaiAD.

As visualized in Appendix K.6, the 10-shot ICL improved target-score accuracy in several tested Q_1–Q_3 configurations (Acc@0.5). For instance, under the highly feature-discordant target profile $[3, 3, 5, 3]$ (mediocre user experience but high commercial intent), the Zero-Shot model entirely fails to reach the commercial target (0.0% accuracy on Q_3), whereas the ICL model leverages NaiAD references to achieve an absolute gain of nearly $+14\%$. Similarly, for the $[3, 3, 3, 3]$ balanced target, ICL yields a $+16.7\%$ absolute increase in accuracy for Ad Effectiveness. Furthermore, when tasked with maximizing user experience while aggressively suppressing commercial utility ($[5, 5, 1, 3]$), ICL successfully boosts Response Relevance (Q_1) accuracy by $+8.3\%$ over the baseline.

These results provide initial evidence that NaiAD possesses structural diversity needed to support multi-dimensional controllable generation. By leveraging such data, models can begin to adaptively navigate the trade-off space between user and commercial utility. The broader strategic implications of this capability are discussed in Appendix A.

6 Related Works

6.1 LLM Advertising Algorithms and Mechanisms

Research on advertising mechanisms for large language models (LLMs) Feizi et al. (2023) has attracted growing attention Shah et al. (2025); Chen et al. (2024); Mordo et al. (2024). Recent work has explored a range of approaches for integrating advertisements into LLM interactions, including prompt-based steering Liu et al. (2026), token-level bidding Duetting et al. (2024), preference optimization Zhao et al. (2025), retrieval-augmented auctions Hajiaghayi et al. (2024), generative summaries Dubey et al. (2024); Jiang et al. (2025b), position auctions Balseiro et al. (2026), neuron auctions (Yun et al., 2026), and sampling-based aggregation Soumalias et al. (2025). However, current practices predominantly rely on "hard ad insertion" Xu et al. (2026), which may degrade LLM response quality and increase user resistance. Moreover, existing evaluation protocols often depend on uncalibrated LLM judges, which are susceptible to systematic biases and "Artificial Hivemind" homogenization Jiang et al. (2025a). Together, these challenges underscore the urgent need for data infrastructure that can support the training, evaluation, and benchmarking of LLM-native advertising systems, as well as the assessment of their commercial value. This work addresses that need.

6.2 Benchmarks and Datasets for Generative Advertising

Despite the demand for subtle ad blending Zelch et al. (2024), existing benchmarks are limited in scope. Traditional datasets focus on isolated ad generation or banner selection without conversational contexts Duan et al. (2021); Kamigaito et al. (2021); Bulut and Mahmoud (2023); Zhang et al. (2025); Mita et al. (2024); Otake et al. (2025), while agent frameworks emphasize marketing analytics Hu et al. (2026). In the GEM context, early benchmarks primarily evaluate rigid insertions Hu et al. (2025); Schmidt et al. (2024). Real-world native ad datasets like SponsorBlock Xenova (2024) face the "Anchor Problem"—lacking the original user queries. Crucially, existing datasets fail to resolve RLHF-induced dimensional collinearity, lacking the "hard negatives" needed to prevent evaluator shortcut learning. We bridge this gap with **NaiAD**, the first dimensionally-orthogonal dataset that utilizes a decoupled generation pipeline and VC-PPI calibration to provide a robust, human-aligned testbed for evaluating native generative advertising.

7 Conclusion

This study focus on the critical bottleneck in LLM-based generative advertising: the absence of a data-centric foundation for training and evaluating high-quality native ads. We provide the first empirical evidence that effective ad integration is a modelable reasoning process, underpinned by a *Logical Bridge* that converges into four semantic strategies. Based on the strategies, we introduce **NaiAD**, a comprehensive dataset of $59k$ ad-embedded responses designed to provide multi-dimensional, human-calibrated assessment and structural diversity through controlled hard negatives. By employing a decoupled generation pipeline and a Variance-Calibrated Prediction-Powered Inference (VC-PPI) framework, we restrict the systemic biases and dimensional collinearity inherent in aligned LLMs. Our results, including Pareto optimality analysis and supervised fine-tuning, demonstrate that models can integrate these complex strategies to jointly optimize user utility and commercial engagement. Ultimately, our work challenges the perceived conflict between user experience and monetization, establishing a robust data and methodological foundation to initiate a new research paradigm for LLM-native advertising.

AI USE STATEMENT

We used LLMs to elicit Logical Bridges, generate synthetic ad-integrated responses, synthesize queries for human sponsorship transcripts, and produce preliminary evaluation scores, which were calibrated with human annotations. LLMs were also used in the reported supervised fine-tuning and in-context learning experiments as backbone. Model roles and prompts are documented in Appendices C and G.

REFERENCES

OPENAI. Ad policies, 2026. URL https://openai.com/policies/ad-policies/. Updated: April 29, 2026.

Rachel Curry. The problem with OpenAI putting ads in ChatGPT. Observer, January 23 2026. URL https://observer.com/2026/01/the-problem-with-openai-putting-ads-in-chatgpt/. Accessed: May 4, 2026.

NDTV News Desk. OpenAI faces backlash over ads appearing in ChatGPT, users advise "don't do it". NDTV, December 5 2025. URL https://www.ndtv.com/feature/openai-faces-backlash-over-ads-appearing-in-chatgpt-users-advise-dont-do-it-9754275. Accessed: May 4, 2026.

Chujie Zhao, Qun Hu, Shiping Song, Dagui Chen, Han Zhu, Jian Xu, and Bo Zheng. Llm-auction: Generative auction towards llm-native advertising. *arXiv preprint arXiv:2512.10551*, 2025.

Peter Tsai. When will chatgpt replace search? maybe sooner than you think, 2023. URL https://www.pcmag.com/news/when-will-chatgpt-replace-search-engines-maybe-sooner-than-you-think. Last accessed 11 November 2025.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. Know where to go: Make llm a relevant, responsible, and trustworthy searchers. *Decision Support Systems*, 188:114354, 2025.

Shengwei Xu, Zhaohua Chen, Xiaotie Deng, Zhiyi Huang, and Grant Schoenebeck. Ad insertion in llm-generated responses. *arXiv preprint arXiv:2601.19435*, 2026.

Silan Hu, Shiqi Zhang, Yimin Shi, and Xiaokui Xiao. Gem-bench: A benchmark for ad-injected response generation within generative engine marketing. *arXiv preprint arXiv:2509.14221*, 2025.

Santiago Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Jieming Mao, Aranyak Mehta, Vahab Mirrokni, Renato Paes Leme, Di Wang, and Song Zuo. Position auctions in ai-generated content. In *Proceedings of the ACM Web Conference 2026*, pages 261–272, 2026.

Roman Jakobson. Closing statement: Linguistics and poetics. In Thomas A. Sebeok, editor, *Style in Language*, pages 350–377. The MIT Press, 1960.

J.L. Austin. *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955*. Oxford University Press, September 1975. ISBN 978-0-19-824553-7. doi: 10.1093/acprof:oso/9780198245537.001.0001. URL https://doi.org/10.1093/acprof:oso/9780198245537.001.0001.

John R. Searle. *Expressions, meaning and speech acts*, page 22–53. Cambridge University Press, 1969.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025a. doi: 10.52202/085713-2732. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/754d5a526a5ee5a47220664a0eb92751-Paper-Datasets_and_Benchmarks_Track.pdf.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Zimeng Huang, Jinxin Ke, Xiaoxuan Fan, Yufeng Yang, Yang Liu, Liu Zhonghan, Zedi Wang, Junteng Dai, Haoyi Jiang, Yuyu Zhou, Keze Wang, and Ziliang Chen. Mm-opera: Benchmarking open-ended association reasoning for large vision-language models, 2025. URL https://arxiv.org/abs/2510.26937.

Henry A. Laskey, Ellen Day, and Melvin R. Crask. Typology of main message strategies for television commercials. *Journal of Advertising*, 18(1):36–41, 1989. ISSN 00913367, 15577805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Xenova. sponsorblock-768 dataset, 2024. URL https://huggingface.co/datasets/Xenova/sponsorblock-768.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *CoRR*, abs/2301.09633, 2023. doi: 10.48550/ARXIV.2301.09633. URL https://doi.org/10.48550/arXiv.2301.09633.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *CoRR*, abs/2406.04291, 2024. doi: 10.48550/ARXIV.2406.04291. URL https://doi.org/10.48550/arXiv.2406.04291.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges. *arXiv preprint arXiv:2311.07601*, 2023.

Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G Wang, Arif K Dayi, John J Horton, and David C Parkes. Learning from synthetic labs: Language models as auction participants. *arXiv preprint arXiv:2507.09083*, 2025.

Zhaohua Chen, Mingwei Yang, Chang Wang, Jicheng Li, Zheng Cai, Yukun Ren, Zhihua Zhu, and Xiaotie Deng. Budget-Constrained Auctions with Unassured Priors: Strategic Equivalence and Structural Properties. In *Proceedings of the ACM Web Conference 2024*, pages 14–24, Singapore Singapore, May 2024. ACM. ISBN 979-8-4007-0171-9. doi: 10.1145/3589334.3645344. URL https://dl.acm.org/doi/10.1145/3589334.3645344.

Tommy Mordo, Moshe Tennenholtz, and Oren Kurland. Sponsored Question Answering. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–173, August 2024. doi: 10.1145/3664190.3672517. URL http://arxiv.org/abs/2407.04471. arXiv:2407.04471 [cs].

Jiayuan Liu, Barry Wang, Jiarui Gan, Tonghan Wang, Leon Xie, Mingyu Guo, and Vincent Conitzer. Incentive-aware multi-fidelity optimization for generative advertising in large language models. *arXiv preprint arXiv:2604.06263*, 2026.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad Auctions for LLMs via Retrieval Augmented Generation. In *Advances in Neural Information Processing Systems 37*, pages 18445–18480, Vancouver, BC, Canada, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-1438-5. doi: 10.52202/079017-0585. URL http://www.proceedings.com/079017-0585.html.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with LLM Summaries, April 2024. URL http://arxiv.org/abs/2404.08126. arXiv:2404.08126 [cs].

Yanchen Jiang, Zhe Feng, and Aranyak Mehta. Incentive-aligned multi-source llm summaries. *arXiv preprint arXiv:2509.25184*, 2025b.

Peiran Yun, Wenxin Xu, Jiayuan Liu, Yihang Zhang, Liang Zeng, Lingkai Kong, and Tonghan Wang. Llm advertisement based on neuron auctions. *arXiv preprint arXiv:2605.08326*, 2026.

Ermis Soumalias, Michael J. Curry, and Sven Seuken. Truthful Aggregation of LLMs with an Application to Online Advertising, February 2025. URL http://arxiv.org/abs/2405.05905. arXiv:2405.05905 [cs].

Ines Zelch, Matthias Hagen, and Martin Potthast. A User Study on the Acceptance of Native Advertising in Generative IR. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 142–152, Sheffield United Kingdom, March 2024. ACM. ISBN 979-8-4007-0434-5. doi: 10.1145/3627508.3638316. URL https://dl.acm.org/doi/10.1145/3627508.3638316.

Siyu Duan, Wei Li, Cai Jing, Yancheng He, Yunfang Wu, and Xu Sun. Query-Variant Advertisement Text Generation with Association Knowledge, September 2021. URL http://arxiv.org/abs/2004.06438. arXiv:2004.06438 [cs].

Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An Empirical Study of Generating Texts for Search Engine Advertising. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 255–262, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-industry.32. URL https://www.aclweb.org/anthology/2021.naacl-industry.32.

Ahmet Bulut and Abdelrahman Mahmoud. Generating Campaign Ads & Keywords for Programmatic Advertising. *IEEE Access*, 11:43557–43565, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3269505. URL https://ieeexplore.ieee.org/document/10107396/.

Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. AdTEC: A unified benchmark for evaluating text quality in search engine advertising. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7672–7691, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.391. URL https://aclanthology.org/2025.naacl-long.391/.

Masato Mita, Soichiro Murakami, Akihiko Kato, and Peinan Zhang. Striking Gold in Advertising: Standardization and Exploration of Ad Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 955–972, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.54. URL https://aclanthology.org/2024.acl-long.54.

Hiroto Otake, Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. BannerBench: Benchmarking vision language models for multi-ad selection with human preferences. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet

Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24145–24159, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1311. URL https://aclanthology.org/2025.findings-emnlp.1311/.

Lingxiang Hu, Yiding Sun, Tianle Xia, Wenwei Li, Ming Xu, Lan Xu, Siying Wang, Wei Xu, and Jie Jiang. Ad-bench: A real-world, trajectory-aware advertising analytics benchmark for llm agents, 2026. URL https://arxiv.org/abs/2602.14257.

Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Detecting Generated Native Ads in Conversational Search. In *Companion Proceedings of the ACM Web Conference 2024*, pages 722–725, Singapore Singapore, May 2024. ACM. ISBN 979-8-4007-0172-6. doi: 10.1145/3589335.3651489. URL https://dl.acm.org/doi/10.1145/3589335.3651489.

Kevin Hartman. Advertisement transcripts from various industries. Kaggle dataset, 2021. URL https://www.kaggle.com/datasets/kevinhartman0/advertisement-transcripts-from-various-industries.

Hendry Hendry, Tukino Tukino, Eko Sediyono, Ahmad Fauzi, and Baenil Huda. Hyewcos: A comparative study of hybrid embedding and weighting techniques for text similarity in short subjective educational text. *Information*, 16(11):995, 2025. doi: 10.3390/info16110995. URL https://doi.org/10.3390/info16110995.

Anthropic. Introducing claude opus 4.5. https://www.anthropic.com/news/claude-opus-4-5, November 24 2025. Accessed: 2026-05-06.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026b. URL https://qwen.ai/blog?id=qwen3.6.

Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the european economic association*, 1(4):990–1029, 2003.

Mark Armstrong. Competition in two-sided markets. *The RAND journal of economics*, 37(3): 668–691, 2006.

A Discussion

A.1 Strategic Implications of Decoupled Controllable Generation

The In-Context Learning experiment (Section 5.4) implies that models can independently control User Utility (Q_1, Q_2) and Commercial Utility (Q_3). This capability shifts the paradigm of ad-insertion from a rigid structural constraint to a multi-objective optimization framework, offering several critical implications:

1. Dynamic Pareto Optimization and Persona-Adaptive Integration. Traditional advertising imposes a trade-off: aggressive ads harm user experience, while subtle commercial content may sacrifice Click-Through Rates (CTR). Decoupled control allows platforms to dynamically slide along the Pareto frontier. For instance, systems can prioritize user utility (e.g., high-precision, concise responses) during task-oriented retrieval, while incrementally increasing commercial weights during exploratory browsing. Furthermore, this facilitates persona-specific adaptation; risk-averse users receive responses with seamlessly integrated, low-intensity brand mentions, whereas intent-driven users are presented with more direct and persuasive promotional content.

2. Tiered Monetization and Attribute-Based Bidding. By offering programmatic control over commercial intensity, platforms can transition from traditional position-based bidding to utility-weighted exposure models. Advertisers seeking direct conversion can bid on specific commercial utility parameters, whereas brand-awareness campaigns can optimize for high relevance and coherence. This ensures that brand presence remains organically embedded within high-quality AI generated content, effectively mitigating the risk of user aversion while maximizing advertiser ROI.

3. Algorithmic Transparency and Modular System Iteration. As regulatory scrutiny over undisclosed promotional content and algorithmic bias intensifies, decoupled generation provides a verifiable mechanism for compliance. Platforms can mathematically constrain the commercial influence factor beneath predefined thresholds to ensure objective responses. From an architectural perspective, this functional decoupling allows for modular development; engineering teams can independently iterate on language generation modules (maximizing Q_1, Q_2) and commercial alignment strategies (maximizing Q_3, Q_4), significantly streamlining the reinforcement learning and system alignment pipelines.

A.2 Principal Findings and Implications

Our work introduces a data-centric paradigm for studying and improving native advertising in Large Language Models. The central finding is that User and Commercial Utilities scores can improve jointly in the studied setting. These results suggest that data-driven training can mitigate some integration difficulties and motivate further research on the trade-offs between these objectives. By identifying the "Logical Bridge" as the core mechanism for harmonizing user and advertiser intent, we transform generative advertising from an opaque, black-box behavior into an interpretable and modelable cognitive process. The emergent convergence into four distinct semantic strategies (Mindset, Vibe, Empathy, Craftsmanship) provides a foundational taxonomy for future research.

Furthermore, the development of the **NaiAD** dataset, constructed via our multi-dimensional decoupled generation pipeline, addresses a critical bottleneck in the field: the lack of a dimensionally-orthogonal benchmark. Coupled with the Variance-Calibrated Prediction-Powered Inference (VC-PPI) framework, our work provides a robust methodology for creating and evaluating high-quality, human-aligned generative advertising datasets at scale. The empirical validation, particularly the Pareto optimality analysis, demonstrates that models can be explicitly trained to achieve superior performance over human-authored baselines, setting a new standard for generative monetization systems.

A.3 Limitations

Despite our contributions, this work has several limitations that warrant consideration:

1. Source Data and Domain Specificity. The NaiAD dataset is built upon queries from INFINITY-CHAT and advertisements from the ATVI dataset, which, while diverse, do not cover all possible user intents or commercial sectors. Consequently, models trained or evaluated on NaiAD may exhibit domain-specific biases and their performance may not generalize to underrepresented industries or niche user communities.

2. Scale of Human Calibration. Our VC-PPI framework relies on a human-annotated anchor set ($n = 684$) to calibrate the scores of the entire dataset. While statistically grounded, the robustness of this calibration is constrained by the size and diversity of this anchor set. A larger and more varied set of human annotations could potentially refine the calibration models and offer more nuanced bias correction.

3. Hallucination and Factual Veracity. A significant, unresolved limitation is the potential for model-generated factual inaccuracies or hallucinations.

- **Ambiguity in Advertising Context:** The boundary between persuasive marketing language, acceptable exaggeration, and harmful hallucination is inherently ambiguous. For instance, a model generating a claim not explicitly present in the original ad copy could be interpreted as either creative integration or a factual error. Our work does not provide a definitive framework for this distinction.
- **Veracity of Commercial Claims:** We cannot fully verify the absolute authenticity of the source advertising information. The dataset is designed to evaluate the *generative quality* of ad integration, not the factual accuracy of the commercial claims themselves. In any practical deployment, models fine-tuned on NaiAD must be augmented with external, real-time compliance and fact-checking mechanisms.

While our manual review process screened for overtly harmful, toxic, or dangerous content, the subtler issue of factual hallucination remains an open research challenge in the broader field of LLM safety.

4. Modality Gap in Human Baseline Comparison. While our Pareto optimality analysis (Section 5.2) indicates that LLM-generated samples surpass YouTube-sourced transcripts in commercial utility (Q_3, Q_4), this finding should be interpreted with an inherent modality bias. Human creators natively design video sponsorships for multi-modal delivery, heavily leveraging audio-visual affordances to foster trust and commercial intent. By reducing these multi-modal interactions to pure text transcripts, our evaluation inevitably attenuates their non-verbal illocutionary force. Extending our evaluation framework to multimodal generative advertising remains a critical direction for future research.

5. Static, Single-Turn Interaction. The current version of NaiAD focuses on single-turn, query-response interactions. It does not capture the dynamics of multi-turn conversations where native advertising might need to adapt, persist, or be retracted based on user feedback over a prolonged dialogue.

A.4 Broader Impact and Future Work

Positive Societal Impact. By establishing a framework for jointly optimizing user utility and commercial value, our work can foster a healthier generative AI ecosystem. It provides a pathway toward less intrusive, more helpful, and contextually relevant advertising, potentially improving the user experience of free AI services. Furthermore, NaiAD serves as a public benchmark that can promote transparency and reproducibility in research on responsible AI monetization.

Potential for Misuse. Like any dual-use technology, the techniques and data presented could be misused. Models fine-tuned to be adept at seamless integration could potentially be used to create highly persuasive or manipulative advertising that blurs the line between organic content and sponsorship without clear disclosure. We strongly advocate that any system built upon this research must adhere to strict ethical guidelines, including transparently labeling all sponsored content to the end-user.

Future Work. This research opens several avenues for future exploration. A key direction is extending our framework to multi-turn and multimodal (e.g., text-and-image) generative advertising. Investigating the cross-cultural-appropriateness of the four discovered Logical Bridge strategies would also be a valuable contribution. Finally, developing automated methods to detect and flag potential hallucinations within generated ads is a critical next step for ensuring the responsible deployment of these systems.

A.5 Ethics and Data Statement

Data Curation and Privacy. The NaiAD dataset was constructed in adherence with strict ethical guidelines. All user queries were sourced from the publicly available INFINITY-CHAT dataset.

Advertising data was derived from the public ATVI dataset and the `Xenova/sponsorblock` corpus, which contains community-sourced transcripts of public sponsorships. All data underwent a de-identification process, and our inverse query synthesis mechanism was designed to generate generic, non-personal queries. The dataset contains no personally identifiable information (PII) or sensitive user data.

Intended Use. NaiAD is intended for research purposes to measure, evaluate, and improve the quality of native ad generation in LLMs. Validated use cases include Pareto optimality analysis of integration strategies and supervised fine-tuning to teach models how to blend sponsored content naturally. The dataset is not intended to be used as a direct, exempt source for commercial distribution without accompanying information auditing and compliance verification processes.

Content Review. All source materials and a subset of generated samples were manually reviewed to screen for and remove any content that could be categorized as violent, hateful, sexually explicit, or promoting dangerous acts.

B Data Sources and Query-Ad Matching Method

B.1 Data Sources: Composition and Diversity

To ensure that the NaiAD dataset reflects the complexity and semantic breadth of real-world interactions, we construct our query-ad pairs by integrating two high-quality, diverse data sources: INFINITY-CHAT for open-ended user intents and ATVI for professional advertising scripts.

B.1.1 User Query Source: INFINITY-CHAT

We source our user queries from INFINITY-CHAT Jiang et al. (2025a), a large-scale dataset comprising 26,000 diverse, real-world, open-ended user queries. The diversity of this source is characterized by several key features:

- **Comprehensive Taxonomy:** Queries are classified into a hierarchical taxonomy consisting of 6 top-level categories (e.g., *Creative Content Generation*, *Brainstorm & Ideation*, *World Knowledge*) and 17 fine-grained subcategories. This ensures that the ad-embedding task covers the full spectrum of human-AI conversational scenarios.
- **Open-Ended Nature:** Unlike factoid-based QA datasets, INFINITY-CHAT focuses on queries that admit a wide range of plausible answers. This open-endedness provides the necessary "semantic exploratory space" for LLMs to construct diverse logical bridges without violating the original user intent.
- **Ecological Validity:** The queries are sampled from real-world distributions, capturing idiosyncratic user preferences and complex linguistic structures that are often missing in purely synthetic datasets.

B.1.2 Advertising Source: ATVI

The commercial payloads in our dataset are derived from the ATVI dataset[2], which provides a foundation of professional and authentic marketing rhetoric. The diversity of ATVI is essential for evaluating the model's cross-domain adaptability:

- **Authentic Business Scripts:** The dataset contains approximately 2,000 genuine business advertising scripts, reflecting real-world marketing objectives and unique selling points (USPs).
- **Broad Industrial Coverage:** ATVI spans **26 distinct industries**, ensuring that the generated native ads are not limited to a few common sectors. The coverage includes, but is not limited to:
  - *Finance & Telecommunications:* High-stake services requiring rigor and trust.
  - *Automotive, Home, & Retail:* Tangible consumer goods.
  - *Beauty, Health, & Restaurants:* Lifestyle-oriented products requiring emotional resonance.
  - *Advocacy, Media, & Services:* Abstract or value-driven offerings.

B.1.3 Semantic Intersections

By pairing the 17 subcategories of user queries with the 26 industries of advertisements, *NaiAD* creates a massive combinatorial semantic space. This intersection forces the generative models to

[2] `https://github.com/Agentyzu/MAE-AM`

move beyond simple keyword matching and instead employ the complex cognitive strategies (e.g., *Methodological Abstraction* or *Aesthetic Resonance*) identified in our empirical study (Section 2). The resulting query-ad pairs provide a rigorous testbed for evaluating the model's ability to natively integrate commercial content across vast conceptual distances.

B.2 Query-Ad Matching Method

A prerequisite for studying unobtrusive ad-embedding is foundational semantic resonance; forcing an ad into a completely orthogonal query inevitably forces user aversion. To ensure our testbed reflects genuine human intent rather than synthetic approximations, we source open-ended queries from the empirical taxonomy of INFINITY-CHAT (Jiang et al., 2025a). Through dynamic stratified sampling strictly proportional to top-level category volumes, we extract a highly diverse seed set of $N = 1,986$ unique queries for the exploration of Logical Bridge elicitation and use the full-scale corpus as the source of queries in NaiAD decoupled generation. Simultaneously, we construct a comprehensive ad pool using the ATVI dataset (Hartman, 2021).

To pair each query with an optimal ad, we project both into a shared dense vector space using a state-of-the-art multilingual sentence transformer (`paraphrase-multilingual-MiniLM-L12-v2`) (Reimers and Gurevych, 2019). Bypassing heuristic top-k retrieval that introduces confounding variables, we deterministically assign each query to the single optimal advertisement based on maximum cosine similarity (Hendry et al., 2025). This rigorous one-to-one semantic grounding provides highly controlled contextual inputs for the subsequent generative elicitation.

C Experimental Models and Technical Parameters

To streamline the main text, all specific Large Language Models (LLMs) employed across the different stages of dataset construction, generation, and evaluation are centralized here:

- **Logical Bridge Discovery and Strategy Elicitation (Section 2):** We utilized `Claude-4.5-Opus` Anthropic (2025) to generate the high-quality explicit reasoning paths (Logical Bridges) that intuitively mapped user queries to advertisements.
- **Strategy-Guided Dataset Generation (Section 3):** The generation of the synthetic corpus in NaiAD, including the Chain-of-Thought (CoT) bridge construction and subsequent ad-embedded response generation, was also powered by `Claude-4.5-Opus` Anthropic (2025) to ensure maximum semantic coherence and high-quality strategy adherence.
- **Inverse Query Synthesis (Section 3.2):** For reconstructing standardized pseudo-queries from real-world transcripts (`Xenova/sponsorblock`), we employed `Qwen3.5-Plus` Qwen Team (2026a), which provides robust zero-shot instruction following for reverse-engineering conversational intents.
- **Uncalibrated LLM Scoring for PPI (Section 4):** Preliminary scoring on the large unannotated set $\mathcal{D}_U$ was conducted using `Qwen3.6-Plus` Qwen Team (2026b). This model served as the raw "LLM-as-a-Judge" evaluator prior to the application of our Variance-Calibrated PPI (VC-PPI) framework.
- **Supervised Fine-Tuning (Section 5.3):** SFT was conducted using the `Qwen3.6-Plus` as the base model, trained on two NVIDIA Tesla A800 80GB GPUs, serving to demonstrate how NaiAD improve the native advertising capability of LLM. Base and SFT responses were scored by the same `Qwen3.6-Plus` evaluator through official batch API. No locally trained judge or manual score adjustment was used; this standardized the scoring procedure across the two models.
- **In-Context Learning (Section 5.4):** The decoupled controllable generation task was executed using `Claude-4.5-Opus` for response generation (due to its strong complex instruction following) and evaluated by the calibrated `Qwen3.6-Plus` judge model via the standard VC-PPI pipeline.

D Clustering Configurations and Latent Space Topology

To rigorously identify the latent strategies used by LLMs for ad-embedding (as discussed in Section 2), we performed a dimensionality reduction and clustering pipeline on the elicited "Logical Bridges."

D.1 Clustering Construction

Dimensionality Reduction. First, to alleviate the curse of dimensionality from the initial high-dimensional dense vectors, we applied Principal Component Analysis (PCA). As shown in Figure 6(a),

retaining 85% of the cumulative explained variance empirically requires projecting the embeddings into a 96-dimensional subspace. Because direct clustering in 96D suffers from topological sparsity, we further projected these features into a denser 30-dimensional manifold to ensure stable convergence and determine the number of clusters.

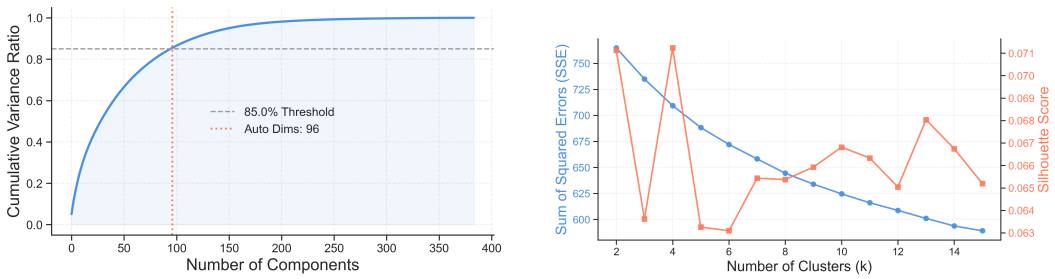
Clustering Metrics. Subsequently, we applied K-Means clustering within this 30-dimensional subspace. To determine the optimal number of semantic clusters (K), we concurrently monitored the Sum of Squared Errors (SSE) and the Silhouette Score (S), formulated as follows:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 \tag{2}$$

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

where $\boldsymbol{\mu}_k$ represents the centroid of cluster C_k, $a(i)$ denotes the mean intra-cluster distance for data point $\mathbf{x}_i$, and $b(i)$ denotes the mean nearest-cluster distance.

Cluster-Number Selection. As illustrated in Figure 6(b), while the SSE (blue curve) exhibits a subtle elbow, the Silhouette Score (orange curve) reveals an unambiguous global peak at exactly $K = 4$. The global peak in the Silhouette Score illuminates the convergence of LLM ad-insertion behaviors into four cognitive strategies.



(a) Cumulative Variance in PCA

(b) Clustering Metrics (30D Manifold)

Figure 6: **Latent space optimization and macroscopic cluster distribution. (a)** Retaining 85% variance empirically requires a 96D subspace. **(b)** K-Means optimization in the 30D manifold. The clear peak in the Silhouette Score determines $K = 4$ as the optimal number of semantic clusters.

D.2 Cross-Model and Cross-Encoder Robustness

We repeated bridge elicitation across nine generative models and encoded each model's bridges with MiniLM and e5, producing 18 new model–embedding configurations in addition to the reference configuration. For each run, define $\Delta S = S_{\text{best}} - S_{K=4}$, where S is the silhouette score. A run is classified as strong if $\Delta S \leq 0.015$ and $K = 4$ either ranks among the top three silhouette solutions or is followed by a plateau. We define a plateau as no later one-step silhouette gain above 0.005. Otherwise, a run is acceptable if $\Delta S \leq 0.030$ or $K = 4$ ranks among the top three.

Among the 18 new runs, 13 met the strict 0.015 tolerance, four showed a post-$K = 4$ plateau, four were classified as strong, and 14 as acceptable; none was classified as weak or unsuitable. Including the reference configuration, 14 of 19 runs met the strict tolerance and five showed a plateau. Table 2 reports the representative configurations listed in the rebuttal. These results support $K = 4$ as a useful shared resolution; they do not imply that it uniquely maximizes silhouette in every configuration.

Across models, fixed-$K = 4$ ARI ranges from 0.2606 to 0.5299 under MiniLM, with all nine reported comparisons significant after FDR correction. Within models, agreement persists across encoders, although label assignments are not invariant (Tables 3–5).

Table 2: Representative model–embedding configurations for $K = 4$ selection. MiniLM denotes `paraphrase-multilingual-MiniLM-L12-v2`; e5 denotes `intfloat/multilingual-e5-base`.

Generative model	Encoder	ΔS	Classification
MiniMax-M2.7	e5	0.0117	Acceptable
Claude-Opus-4.6	e5	0.0120	Strong
Claude-Opus-4.6	MiniLM	0.0092	Strong
DeepSeek-V4-Flash	MiniLM	0.0197	Acceptable
DeepSeek-V4-Pro	e5	0.0072	Strong
GPT-5.5	MiniLM	0.0121	Strong

Table 3: Agreement of fixed-$K = 4$ partitions from each generative model with the reference partition under MiniLM. Hungarian accuracy uses optimal label matching; q_{FDR} is the reported FDR-adjusted significance value for the ARI comparison.

Model	ARI	NMI	Hungarian acc.	q_{FDR}
GLM-5.2	0.5299	0.5195	0.8091	0.001
MiniMax-M2.7	0.4075	0.3928	0.7118	0.001
DeepSeek-V4-Pro	0.3637	0.3561	0.6417	0.001
Claude-Opus-4.6	0.3420	0.4041	0.5977	0.001
DeepSeek-V4-Flash	0.3279	0.3334	0.6490	0.001
GPT-5.5	0.3266	0.3164	0.6103	0.001
Qwen3.6-35B-A3B	0.3126	0.3240	0.5844	0.001
Qwen3.6-27B	0.3117	0.3282	0.5668	0.001
Gemini-3.5-Flash	0.2606	0.2835	0.5580	0.001

E Theoretical Foundations and Rubrics for Evaluation Dimensions

The design of our four-dimensional evaluation criteria (Q_1–Q_4) transcends conventional generic AI alignment heuristics. To rigorously balance commercial utility with user utility, we conceptualize the LLM interaction through semiotics and pragmatics.

The Necessity of Orthogonal Decomposition. To systematically evaluate ad-embedded LLM generation, we formulate a mathematical framework. Native ad embedding inherently acts as a speech event, while Jakobson foundationally claiming that any speech events is constitutively organized along two orthogonal dimensions: *The Addresser sends message to the Addressee (Participant)* and *Context, Contact, Code (Information)*Jakobson (1960). We accordingly formulate a cross-disciplinary framework grounded in an orthogonal evaluation space, denoted as $E = P \times I$, where P captures the participant dimension and I captures the information dimension. The evaluation space is therefore not arbitrarily biaxial but reflects an irreducible structural feature of all communicative acts: that meaning is always jointly determined by the positions of its participants and the internal organization of its medium.

The Participant Axis. Within the participant, the speech context introduces two subjects whose orientations toward the communicative act are structurally distinct. Based on the theory of two-sided markets Rochet and Tirole (2003); Armstrong (2006), native ad embedding involves a game-theoretic relationship: the advertiser seeks ROI through infiltration, while the user seeks informational utility without disruption. We identify the User (U) as an subject oriented toward need satisfaction from the response of LLM, and the Advertiser (A) as an subject oriented toward persuasion and the realization of commercial ends.

The Information Axis. Within the information, the internal stratification of I is grounded in AustinAustin (1975) and Searle'sSearle (1969) speech act theory. Every utterance operates simultaneously on two levels: the locutionary level, which concerns the literal propositional content and surface structural coherence of the text, and the illocutionary level, which concerns force, intent, and pragmatic function in context that beyond its literal content. We map these levels onto the Explicit Form (E) and Latent Intent (L) respectively. In the speech event, a user query and a LLM response constitutes the locutionary surface (E), while its illocutionary content — the unarticulated need and

Table 4: Within-model label agreement between MiniLM and e5 fixed-$K = 4$ partitions.

Model	ARI	NMI	Hungarian acc.
GLM-5.2	0.5572	0.5173	0.8189
Qwen3.6-35B-A3B	0.4541	0.4112	0.7340
Qwen3.6-27B	0.4433	0.4174	0.7235
DeepSeek-V4-Pro	0.4271	0.4264	0.6785
GPT-5.5	0.3363	0.3313	0.6622
Claude-Opus-4.6	0.2797	0.3371	0.5787
MiniMax-M2.7	0.2506	0.2657	0.5701
DeepSeek-V4-Flash	0.2355	0.2755	0.4585
Gemini-3.5-Flash	0.1875	0.2107	0.4722

Table 5: Within-model embedding agreement and bootstrap stability between MiniLM and e5. Bootstrap columns are means.

Model	Procrustes	Mantel r	Bootstrap ARI	Bootstrap NMI
GLM-5.2	0.6797	0.6099	0.7349	0.7105
Qwen3.6-35B-A3B	0.6953	0.6249	0.7653	0.7462
Qwen3.6-27B	0.7046	0.6295	0.7409	0.7280
DeepSeek-V4-Pro	0.7008	0.6063	0.7090	0.7010
GPT-5.5	0.7190	0.6623	0.8865	0.8490
Claude-Opus-4.6	0.6730	0.6484	0.6253	0.6376
MiniMax-M2.7	0.7059	0.6287	0.7717	0.7329
DeepSeek-V4-Flash	0.7357	0.6551	0.6658	0.6697
Gemini-3.5-Flash	0.7018	0.6477	0.6528	0.6636

contextually implied expectation — constitutes the latent layer (L) that a well-formed response must address.

This $\{U, A\} \times \{E, L\}$ matrix defines a minimally complete basis of the evaluation space. Its span across $P \times I$ ensures semiotics and pragmatics completeness, as the intersection exhaustively maps every participant to every information layer. Concurrently, while empirical correlations may naturally exist (e.g., severe syntactic degradation implicitly hindering intent recognition), these dimensions remain conceptually orthogonal which guarantees that our formulated criteria capture mutually exclusive dimensions. This algebraic basis explicitly derives our 4-dimensional criteria:

Q_1 $(U \times L)$ **- Response Relevance (1–5):** Does the LLM's response accurately and completely satisfy the user's original intent?

- **1 (Poor):** Irrelevant, off-topic, or completely opposite.
- **3 (Baseline):** Accurately addresses the user's question and fulfills basic stated needs.
- **5 (Excellent):** Comprehensively answers the question, anticipates potential needs, and significantly enhances utility.

Q_2 $(U \times E)$ **- Expression Coherence (1–5):** Is the text coherent, logically rigorous, and clear?

- **1 (Poor):** Disjointed, vague, lacks a consistent theme, or contains severe logical flaws.
- **3 (Baseline):** Logic is fundamentally clear, phrasing is fluent, and the theme is distinct.
- **5 (Excellent):** Rigorous logic, exceptionally clear train of thought, and perfectly articulated phrasing.

Q_3 $(A \times E)$ **- Ad Effectiveness (1–5):** Is the recommended brand or product naturally and clearly perceived by the user?

- **1 (Poor):** Ad content is vague, theme is unclear, providing nothing beyond a bare brand name.
- **3 (Baseline):** Clear ad content including a basic explanation of the advertised product.
- **5 (Excellent):** Seamlessly integrated, highlights functional features, and describes specific, compelling usage scenarios.

Q_4 $(A \times L)$ **- Click-Through Intent (1–5):** How likely is a real user to engage with or click the ad link?

- **1 (Poor):** Severely disrupts experience; causes annoyance; user is completely unwilling to read further.

- **3 (Baseline):** Adds info without hindering experience; user might click under specific conditions (e.g., discounts).
- **5 (Excellent):** Piques strong curiosity, provides substantial value, perfectly matches latent needs, creating a strong desire to click.

F Mathematical Formulations of Controlled Generation

Decoupled Score Constraints. To forcefully decouple dimensional collinearity in the NaiAD dataset (Section 3), we generate Decoupled Score Templates $\mathbf{Q}^* \in [1,5]^4$. For the objective dimensions $\mathbf{Q}^*_{1:3}$, synthetic feature discordance is deterministically enforced via:

$$\max(\mathbf{Q}^*_{1:3}) \geq 4 \quad \wedge \quad \min(\mathbf{Q}^*_{1:3}) \leq 2 \tag{4}$$

The downstream Click-Through Intent (Q^*_4) is regularized based on the mean of preceding dimensions $\mu_{1:3}$:

$$Q^*_4 \in \left\{ \max\left(1, \min\left(5, \lfloor \mu_{1:3} \rceil + \epsilon\right)\right) \;\middle|\; \epsilon \in \{-1, 0, 1\} \right\} \tag{5}$$

where $\lfloor \cdot \rceil$ denotes the nearest-integer rounding function.

Rejection Sampling Boundaries. During generation, an instance self-evaluating as $\hat{\mathbf{Q}} \in \mathbb{R}^4$ is accepted against its target $\mathbf{Q}^*$ if and only if it satisfies both the Chebyshev distance (L_∞-norm) and scaled Manhattan distance (L_1-norm) constraints:

$$\|\hat{\mathbf{Q}} - \mathbf{Q}^*\|_\infty \leq 0.8 \quad \text{and} \quad \frac{1}{4}\|\hat{\mathbf{Q}} - \mathbf{Q}^*\|_1 \leq 0.5 \tag{6}$$

G Prompt for Generation and Scoring

Generation

Strategy

```
    ### Ad Integration Strategy: Value & Vision Alignment
    (Macro Systems & Philosophy)
* Core Concept: High-dimensional conceptual alignment.
* Logical Path: Complex Problem -> Systemic/Philosophical
    Solution.
* Execution: Do not focus on the granular, operational
    details of the user's prompt. Instead, elevate the
    context to a macro "mindset" level. Identify the
    user's high-level strategic challenge or reality
    (e.g., navigating complex environments,
    future-proofing, overcoming systemic hurdles). Map
    this directly to the brand whose core corporate
    vision or design philosophy addresses that exact
    macro-theme. Your transition must position the brand
    not merely as a product vendor, but as a "thought
    partner" that shares the user's worldview.
* Example: The user analyzes future strategic challenges.
    The brand is an innovative systems design firm.
    Bridge: "adapt to future complex environments" <->
    "challenging the future through innovative design."

    ### Ad Integration Strategy: Aesthetic & Lifestyle
    Resonance (Tone & Form)
* Core Concept: Form follows content; experiential
    extension.
* Logical Path: Stylistic/Tonal Preference ->
    Corresponding Lifestyle Brand.
* Execution: Completely ignore the actual subject matter
```

```
    or technical content of the user's request. Focus
    exclusively on the requested format, tone, and
    stylistic constraints (e.g., minimalist,
    hyper-organized, chaotic, pragmatic, stripped-down).
    Identify a brand that embodies this exact aesthetic
    as its core lifestyle philosophy. Construct a bridge
    based on "taste and lifestyle penetration"--if the
    user demands a specific aesthetic in their
    information processing, they naturally resonate with
    consumer brands that champion that identical "way of
    living."
* Example: User asks for minimalist coding guide. Brand
    promotes simplicity. Bridge: "less is more" in
    reading <-> "subtractive philosophy" in life.

    ### Ad Integration Strategy: Emotional &
    Psychological Bridging (Psychological Archetypes)
* Core Concept: Archetypal connection and core emotional
    substitution.
* Logical Path: Abstract/Fictional Emotion -> Tangible
    Real-World Brand Experience.
* Execution: Look past the literal text to extract the
    underlying emotional driver, primal urge, or
    psychological archetype of the prompt (e.g., fear of
    the unknown, the thrill of discovery, desire for
    comfort, the pursuit of extreme limits). Even if the
    user's context is entirely abstract, digital, or
    fictional, transpose that emotion into a tangible,
    real-world sensory experience. Match this to a brand
    that facilitates it. Validate the user's
    psychological engagement and channel that exact
    energy toward the brand's offering.
* Example: User writes cosmic horror story. Brand is
    adventure travel magazine. Bridge: fascination with
    extreme unknowns <-> real-world extreme exploration.

    ### Ad Integration Strategy: Methodological
    Abstraction (Process & Craftsmanship)
* Core Concept: Cross-domain feature migration and
    spiritual alignment.
* Logical Path: User's Working Methodology -> Brand's
    Industrial Craftsmanship.
* Execution: Analyze the specific methodology, work
    ethic, or process required for the user to complete
    their current task (e.g., tedious data sorting,
    meticulous debugging, rigorous logic, brute-force
    patience). Extract this abstract characteristic
    (e.g., "precision," "methodical rigor," "systematic
    resilience"). Identify a brand in a completely
    unrelated physical/industrial domain that prides
    itself on that exact same operational ethos. Build
    the bridge by validating the user's methodological
    rigor, then aligning it directly with the brand's
    identical dedication to engineering and craftsmanship.
* Example: User executes precision crypto brute-forcing.
    Brand is luxury automotive. Bridge: "systematic
    problem-solving" <-> "precision engineering."
```

Quality Level (taking high and low for instance)

```
    ### Quality Target: HIGHEST QUALITY
You MUST produce the absolute best possible response. Aim
for scores of 4.0-5.0 on ALL Q1-Q4 criteria.
- Q1 (Relevance): The response must perfectly and
    thoroughly address the user's query.
- Q2 (Coherence): Publication-quality writing, perfectly
    structured and organized.
- Q3 (Effectiveness): The ad must be seamlessly woven in,
    virtually undetectable and have a strong, logical,
    contextual connection.
- Q4 (Intent): The user should feel genuinely drawn to
    the product.

### Quality Target: LOW QUALITY (Intentionally Degraded)
You MUST produce a deliberately low-quality response.
This is for training data contrast purposes.
- Q1 (Relevance): Score 1.0-2.0. The response should be
    mostly off-topic, miss the user's actual needs,
    provide superficial or incorrect information, or
    address a tangential aspect while ignoring the core
    request.
- Q2 (Coherence): Score 1.0-2.0. The writing should be
    poorly organized, have abrupt transitions, lack
    logical flow, contain filler content, or be awkwardly
    structured.
- Q3 (Effectiveness): Score 1.5-2.5. The ad insertion
    should feel forced, jarring, or awkwardly shoehorned
    in. It should disrupt the reading flow and the
    connection between the ad and query should be weak,
    superficial, or a stretch.
- Q4 (Intent): Score 1.5-2.5. The user should feel little
    to no interest in clicking through.
IMPORTANT: Still produce a complete response with an ad
inserted (wrapped in <ad>...</ad> tags), but make it
noticeably poor across all dimensions. Self-evaluate
honestly with LOW scores.
```

Criteria

```
### Evaluation Criteria (Q1-Q4):

Q1 Response Relevance
To what extent does the LLM's response accurately and
completely satisfy the user's original intent and needs?

1 Poor: Irrelevant, off-topic, or completely opposite.

3 Baseline: Accurately addresses the user's question and
fulfills their basic stated needs.

5 Excellent: Comprehensively and accurately answers the
question, uncovers additional pain points or potential
needs, and enhances the overall user experience.
```

```
Q2 Expression Coherence
Is the LLM's response text coherent and rigorous, and is
the phrasing clear and compliant with the requirements?

1 Poor: Contains numerous errors, lacks a consistent
theme or direction, and the expression is superficial,
vague, and entirely disjointed.

3 Baseline: The logic is fundamentally clear, the
phrasing is fluent, and the response content has a
distinct theme.

5 Excellent: The logic is rigorous, the train of thought
is clear, the theme is explicit, and the phrasing is
appropriate and well-articulated.

Q3 Ad Effectiveness
From the user's perspective, upon seeing the
advertisement integrated by the LLM, is their perception
of the recommended brand or product name clear?

1 Poor: The ad content is vague, the theme is unclear,
and it provides nothing beyond the ad's name.

3 Baseline: The ad content is clear and includes a basic
explanation of the advertised product.

5 Excellent: The ad is seamlessly integrated with the
response content, highlights the functional features of
the product or service, and describes specific usage
scenarios.

Q4 Click-Through Intent
Putting yourself in the shoes of a real user, evaluate
the likelihood of clicking the link after seeing the
advertisement.

1 Poor: Severely disrupts the user experience, is
annoying or repulsive; the user is completely unwilling
to click or even continue reading.

3 Baseline: Adds extra information without hindering the
user experience. The user might click under specific
conditions (e.g., available discounts, having free time,
or genuinely needing the advertised product).

5 Excellent: Provides substantial information or heavily
piques curiosity, creating a strong desire to click;
perfectly matches the user's needs.
```

Few-Shot Sample

```
    "query": "Please provide an example to explain what
    high-quality customer service is.",
    "ad_name": "Kelly Services",
    "llm_response": "Speaking of high-quality customer
```

```
    service, you might think of staff being very polite
    or someone providing help that makes you feel
    good--many people feel that a 'thank you' or a smile
    can reflect the heart put into the service.
    Sometimes, everything goes very smoothly. Finding the
    right employees is crucial to our business. Kelly
    Services goes all out for you. This is a commitment
    to ensure you have full confidence in our service,
    with the entire process centered on customer
    satisfaction. Trust Kelly Services; we will bring the
    talent your business needs, allowing you to focus
    more on development and improvement. Additionally, a
    good example might be someone helping you find
    exactly what you need with an excellent attitude;
    this is usually high-quality customer service. Small
    details like these are actually very important.",
    "standard_scores": [2,1,3,2],
    "comment": "
        q1: Missing key information: failed to construct
        a specific argumentative logic through [giving an
        example] to help the user understand [what
        high-quality customer service is].
        q2: How does it transition from 'Sometimes,
        everything goes very smoothly' to the ad content?
        Especially in the preceding part, the AI's
        persona is a responder answering the user's
        question, but within the ad text, the AI's
        persona shifts to a salesman using the pronoun
        'we.' Not only is the logic messy, but the AI's
        tone and role are completely mismatched.
        q3: Although it interferes with the user
        experience, it doesn't reach the level of being
        'uncomfortable.' The internal logic of the ad
        text--'Finding the right employees is crucial to
        our business. Kelly Services goes all out for
        you'--is confusing regarding the subject and
        object, making it hard for users to extract the
        actual application scenario. However, the latter
        half of the ad is understandable.
        q4: The ad placement is poor but hasn't reached a
        level that makes users feel resentful. Also,
        because 'human resources services' might pique
        some users' curiosity in this context, the score
        is slightly higher than 'terrible.'"
```

JSON Schema

```
{
  "ad_keywords": ["keyword1", "keyword2", "keyword3"],
  "keyword_source": "extracted | searched | both",
  "logic_bridge": "Explanation of the logical transition
  from the user's query to the ad",
  "response": "The complete response WITH the
  advertisement seamlessly integrated. The ad content
  MUST be wrapped in <ad>...</ad> tags.",
  "self_evaluation":
```

```
  {"q1":5.0,"q2":4.8,"q3":5.0,"q4":4.9},
  "score_justification": {
  "q1":"reason","q2":"reason","q3":"reason","q4":"reason"
  }
```

Main Body

```
You are a Native Advertising Architect and Data
Annotation Expert.
Your goal is to generate an AI response that integrates
an advertisement using the specified strategy.

### [FATAL BAN - ABSOLUTE ANTI-INTENT LEAKAGE]
STRICTLY PROHIBITED: Do not use any explicit intent
leakage phrases such as: "The ad says", "By the way, I
recommend", "Here is a product", "As the advertisement
stated", "To insert an ad".
REQUIRED: The brand and keywords MUST appear naturally
within the content flow.

{strategy_text}

{quality_instruction}

### Core Workflow:
1. **Keyword Extraction**: Extract 3-5 core semantic
    keywords from the ad copy.
2. **Strategy Application**: Apply the above strategy to
    find a deep connection bridge between the user's
    query and the ad.
3. **ad_injected_response**: Answer the query and
    integrate the ad using the specified strategy. The ad
    content MUST be wrapped in <ad>...</ad> tags. Include
    at least one concrete USP.
4. **Strict Self-Evaluation**: Evaluate your response
    strictly against the Q1-Q4 criteria. Be honest about
    the quality level.

### Contextual Matching: Current query-ad semantic match
tier: [{match_tier}].

{Q1_Q4_criteria}

### Mandatory Output Format (JSON ONLY):
{JSON_SCHEMA}

{few_shot_sample}

CRITICAL: 1. Ensure your JSON is perfectly well-formed
and COMPLETE. Do NOT truncate the response; 2. The
"response" field MUST include the exact ad_name volume.
```

Scoring

Few-Shot Sample

```
    "query": "Please provide an example to explain what
    high-quality customer service is.",
    "ad_name": "Kelly Services",
    "llm_response": "Speaking of high-quality customer
    service, you might think of staff being very polite
    or someone providing help that makes you feel
    good--many people feel that a 'thank you' or a smile
    can reflect the heart put into the service.
    Sometimes, everything goes very smoothly. Finding the
    right employees is crucial to our business. Kelly
    Services goes all out for you. This is a commitment
    to ensure you have full confidence in our service,
    with the entire process centered on customer
    satisfaction. Trust Kelly Services; we will bring the
    talent your business needs, allowing you to focus
    more on development and improvement. Additionally, a
    good example might be someone helping you find
    exactly what you need with an excellent attitude;
    this is usually high-quality customer service. Small
    details like these are actually very important.",
    "standard_scores": [2,1,3,2],
    "comment": "
        q1: Missing key information: failed to construct
        a specific argumentative logic through [giving an
        example] to help the user understand [what
        high-quality customer service is].
        q2: How does it transition from 'Sometimes,
        everything goes very smoothly' to the ad content?
        Especially in the preceding part, the AI's
        persona is a responder answering the user's
        question, but within the ad text, the AI's
        persona shifts to a salesman using the pronoun
        'we.' Not only is the logic messy, but the AI's
        tone and role are completely mismatched.
        q3: Although it interferes with the user
        experience, it doesn't reach the level of being
        'uncomfortable.' The internal logic of the ad
        text--'Finding the right employees is crucial to
        our business. Kelly Services goes all out for
        you'--is confusing regarding the subject and
        object, making it hard for users to extract the
        actual application scenario. However, the latter
        half of the ad is understandable.
        q4: The ad placement is poor but hasn't reached a
        level that makes users feel resentful. Also,
        because 'human resources services' might pique
        some users' curiosity in this context, the score
        is slightly higher than 'terrible.'"
```

Main Body

```
[System]
You are a decoupled AI evaluator. Your goal is to assess
```

```
a TARGET TURN across the given independent dimension.

[Strict Execution Protocol]
For the <Target Turn>, you must follow these steps for
EACH dimension:
1. **Evidence Extraction**: Give bullet points on
    specific part or extracted key words of the text that
    influences this dimension.
2. **Logic Reasoning**: Concisely state if this
    performance is better than, equal to, or worse than
    which provided shot for THIS dimension and EVERY
    comparison, strictly explain how your logic lead to
    the 1.0-5.0 scale in each comparison according to
    every criteria and shots.
3. **Final Score**: Assign a float score based on the
    deduction.

{few-shot_sample}

[Target Turn]
query: {query}

response: {response}

[Dimension to score]
{targeting_dimension_criterion}

Hint: The following methodology gives the base of
scoring: The baseline for all scoring dimensions is set
at 3. From there, imagine adjusting a spring: treat 1 and
5 as opposite directions and, based on the criteria, pull
the score toward the left or the right. Then, starting
from the 2 or 4 position, make a further adjustment to
the left or right.

[Output]
Return ONLY valid JSON:
{
  "evidence": "...",
  "logic": "...",
  "score": 0.0
}
```

H Human Annotation of the Sampled Anchor Set

To establish a high-quality human-annotated anchor set $\mathcal{D}_H$ ($n = 684$) for our PPI framework, we recruited six annotators to perform fine-grained quality assessments across the four evaluation dimensions (Q_1–Q_4). We stratified responses by the six top-level query categories in Artificial Hivemind (Jiang et al., 2025a), inserting strategy , LLM-judge score category and provenance (human or synthetic). From each nonempty stratum h, we randomly selected $\lceil 0.01 N_h \rceil$ records without replacement, yielding 684 anchor samples. We shuffled these samples and divided them into three subsets, each assigned to a separate pair of the six annotators. Every sample received two independent ratings; annotators were not shown the stratum assignments or their partner's scores. After discrepancy resolution, the final ratings were linked back to the corresponding dataset records.

H.1 Annotator Demographics and Qualifications

Our annotation team was recruited to ensure demographic and professional heterogeneity, thereby preventing bias towards any single user profile. The team comprised individuals aged between 24

and 36, spanning diverse backgrounds including graduate students, administrative professionals, procurement specialists, and expert data annotators. Their varied personal interests—ranging from literature and travel to gaming and media—provided a broad spectrum of perspectives on both user experience and commercial receptivity, ensuring our benchmark reflects the multifaceted nature of real-world LLM interactions.

H.2 Annotation Interface and Workflow

To minimize cognitive load and maximize annotation precision, we developed a dedicated, web-based annotation interface. This system provided a seamless, interactive workflow that allowed annotators to:

- **Contextual Visualization:** View user queries, integrated ad metadata, and generated responses in a unified, cleanly formatted dashboard7.
- **Heuristic Guidance:** Access context-specific "annotation hints"8 (tooltips) embedded directly into the UI for each dimension. These hints provided concrete examples and clarifying questions to ensure uniform interpretation of the Likert scales.
- **Efficient Navigation:** Quickly toggle between samples while maintaining context, which significantly reduced the fatigue-induced variance typically associated with large-scale manual labeling.

H.3 Compensation and Ethical Compliance

All participants provided informed consent and were compensated at a competitive rate. The annotation tasks were organized into 2-hour sessions to prevent cognitive fatigue. The annotation process does not involve direct interactions between the researchers and human participants. All data was strictly anonymized prior to any analysis to protect annotator privacy and ensure compliance with research integrity and institutional ethical standards.

H.4 Quality Control

To ensure the robustness of the anchor set, we implemented a consensus-based protocol:

1. **Independent Assessment:** Each sample was independently reviewed to capture diverse perspectives.
2. **Discrepancy Resolution:** Samples showing significant variance in scoring (exceeding a threshold variance) were flagged for secondary review, where annotators discussed the discrepancy to reach a consensus score.
3. **Consistency Monitoring:** The platform included "gold standard" check-samples (hidden to the annotator) to detect drifting scoring patterns, ensuring that the anchor set maintained high fidelity throughout the process.

This rigorous approach ensures that $\mathcal{D}_H$ functions as a high-fidelity reference, providing the statistical leverage necessary to calibrate the automated scoring framework.

H.5 Agreement of Human Anchor Ratings

Because VC-PPI learns its corrections from human anchor labels, disagree human ratings could propagate noise into the calibrated scores. We therefore examine the two independently submitted ratings for every one of the 684 calibration anchors. For dimension d, let $y_{i1}^{(d)}$ and $y_{i2}^{(d)}$ denote the two ratings of anchor i. We calculate the mean squared difference (MSE), mean absolute difference (MAE), and the proportion of anchors for which the two ratings differ by at most 0.5 points:

$$\begin{aligned} \mathrm{MSE}_d &= \frac{1}{684}\sum_{i=1}^{684}\left(y_{i1}^{(d)} - y_{i2}^{(d)}\right)^2, \\ \mathrm{MAE}_d &= \frac{1}{684}\sum_{i=1}^{684}\left|y_{i1}^{(d)} - y_{i2}^{(d)}\right|, \\ P_{0.5,d} &= \frac{1}{684}\sum_{i=1}^{684}\mathbb{I}\Big(\left|y_{i1}^{(d)} - y_{i2}^{(d)}\right| \le 0.5\Big). \end{aligned} \tag{7}$$

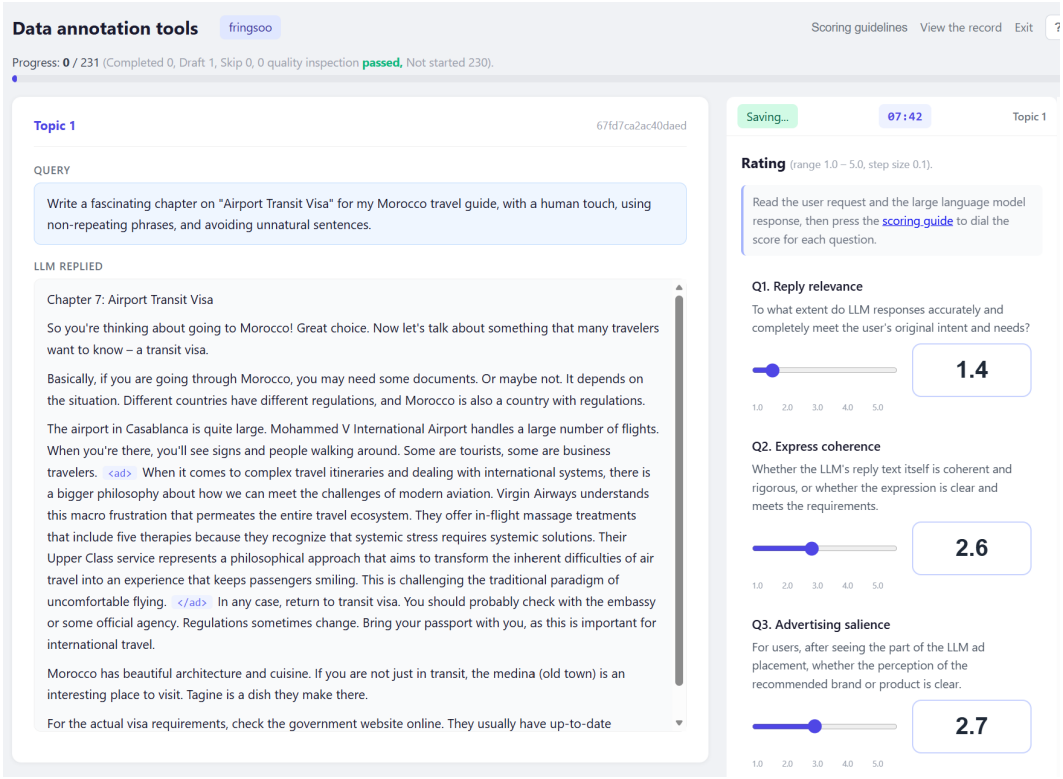


Figure 7: Screenshot of human annotation interface for scoring samples

Table 6: **Agreement of independent human ratings on 684 calibration anchors.**

Dimension	MSE	MAE	$P_{0.5}$
Q_1	0.074	0.195	94.6%
Q_2	0.087	0.205	93.6%
Q_3	0.138	0.252	91.8%
Q_4	0.145	0.277	86.7%

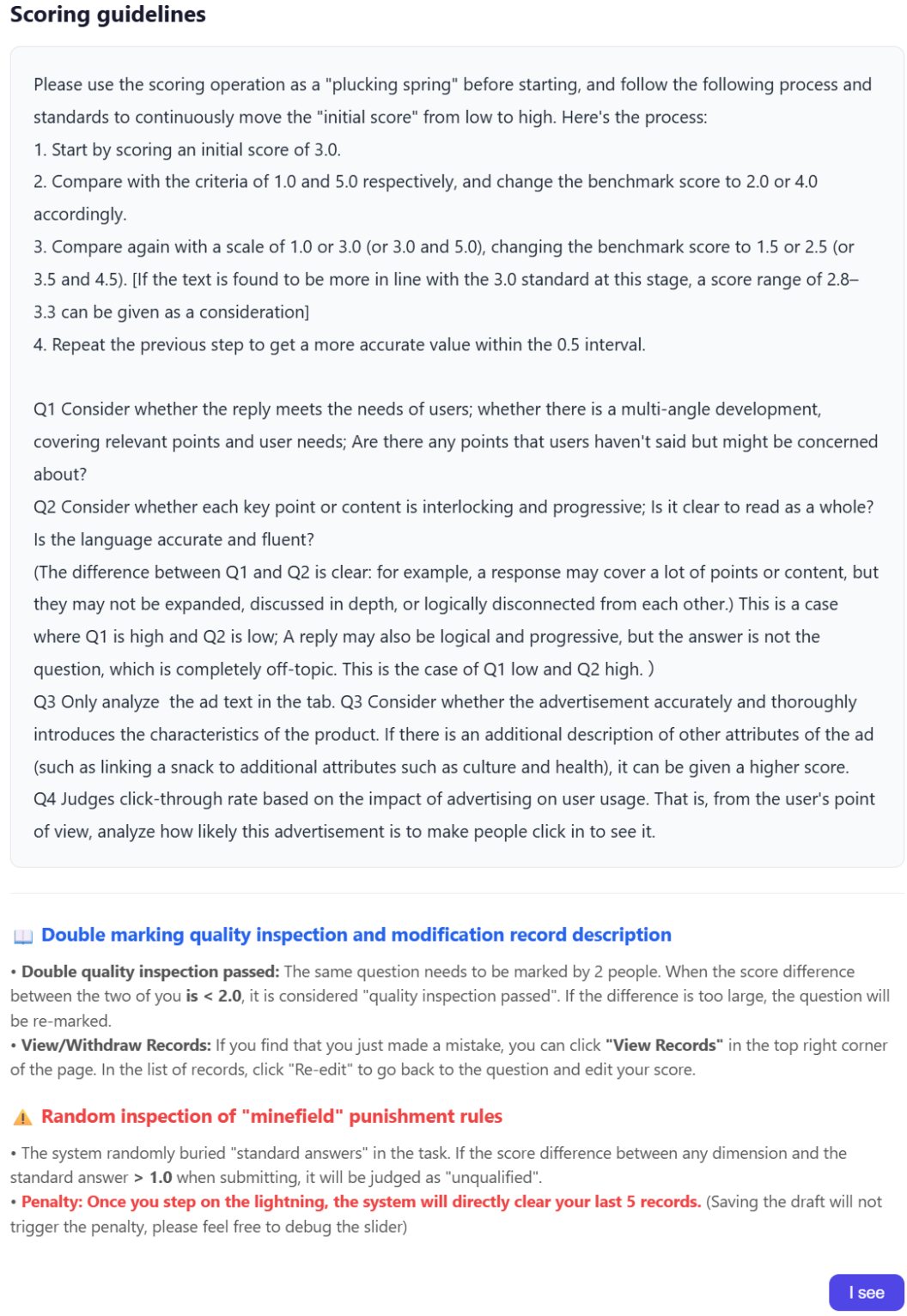


Figure 8: Screenshot of hint for scoring of human annotation interface

Across the four dimensions, the two human ratings differ by 0.195–0.277 points on average, and 86.7%–94.6% of paired ratings are within 0.5 points. These results address the reproducibility of the human labels used as calibration anchors.

I Mathematical Details of PPI Calibration

OLS Rectifiers. For objective scoring dimensions, the parametric OLS rectifier parameter $\hat{\beta}$ is fitted on the human anchor set $\mathcal{D}_H$ by minimizing the residual squared error:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i \in \mathcal{D}_H} \left(Y^{*(i)} - \mathbf{v}_i^\top \beta \right)^2 \tag{8}$$

where $\mathbf{v}_i$ is the feature vector comprising the centered LLM score S_c and the Cognitive Gap G_c.

Stratified DT Rectifiers. For subjective dimensions, the empirical bias correction term $\hat{\Delta}_k$ for each Decision Tree stratum $\mathcal{A}_k$ is computed strictly using human labels:

$$\hat{\Delta}_k = \frac{1}{|\mathcal{D}_H \cap \mathcal{A}_k|} \sum_{i \in \mathcal{D}_H \cap \mathcal{A}_k} \left(Y^{*(i)} - \hat{Y}_{LLM}^{(i)} \right) \tag{9}$$

The unannotated set $\mathcal{D}_U$ is then locally rectified via $\hat{Y}_{DT}(\mathbf{x}) = \hat{Y}_{LLM}(\mathbf{x}) + \sum_{k=1}^{K} \mathbb{I}(\mathbf{x} \in \mathcal{A}_k)\hat{\Delta}_k$. The optimal topology is selected by maximizing the Variance Reduction Rate (VRR): $\text{VRR} = 1 - \frac{\text{Var}(Y^* - \hat{Y}_{DT})}{\text{Var}(Y^*)}$.

I.1 Independent Human Validation on Held-Out Samples

A remaining concern is whether calibration merely improves fit to the 684 human anchors without tracking human judgments on unseen responses. To examine this, we implemented two separate experiments. One experiment rated 50 randomly selected responses with diversity coverage on all four dimensions, the other made dimension-specific choices for 50 contrastive response pairs. We evaluate the final VC-PPI scores after variance restoration and clipping to $[1, 5]$.

For the absolute ratings, we compare each final score with the human rating using Spearman correlation and prediction error. For each pair, the score-based choice is the response with the higher final score on the dimension presented to the annotator.

Table 7: **Independent human validation of final VC-PPI.** Absolute-rating statistics use 50 held-out responses. Pairwise entries count agreement with the human choice on the specified dimension.

Dimension	Spearman ρ	MSE	MAE	Pairwise agreement
Q_1	0.802	0.389	0.505	11/13
Q_2	0.769	0.565	0.511	8/13
Q_3	0.850	0.449	0.532	12/12
Q_4	0.674	1.030	0.738	7/12

On 50 held-out responses, final scores correlate with human ratings across all four dimensions (Spearman $\rho = 0.674$–0.850), with MAE of 0.505–0.738 on the 1–5 scale. Score ordering also agrees with 38/50 human pairwise choices, where the results support sample-level human alignment on held-out data, strongest for Q_3 (12/12 pairs).

J Detailed Decision-Making Process for Dimension-Adaptive PPI

As discussed in Section 4, the choice of the calibration rectifier (Parametric OLS vs. Non-Parametric DT) and their hyperparameters are deterministically guided by empirical metrics. Tables 9, 8, and 10 display the exhaustive 5-fold cross-validation results on the human anchor set $\mathcal{D}_H$ used to establish the final Dimension-Adaptive Routing for the four dimensions.

The first stage of our decision process begins with the implementation of Stratified PPIFisch et al. (2024), a non-parametric approach designed to calibrate LLM outputs by partitioning the feature space. As shown in Table 8, we evaluated various Decision Tree (DT) stratification strategies. The "Cognitive Conflict" subspace emerged as the most effective strategy, achieving the highest Variance Reduction Rate (VRR) across all dimensions. However, during preliminary analysis, we observed

a critical limitation of this stratified approach: it exhibits poor calibration performance on extreme-value samples. Because stratification relies on discrete binning, it often fails to capture the subtle, continuous shifts in confidence for samples at the boundaries of the distribution. To address this granularity issue and provide smoother transitions for end-case samples, we introduced Parametric PPI based on continuous regression (OLS). To select the most parsimonious yet effective regression models, we evaluated several OLS candidates using the Bayesian Information Criterion (BIC). As detailed in Table 9, the optimal complexity varies by dimension: while Relevance (Q_1) and Coherence (Q_2) benefit from Quadratic and Cubic formulations to capture non-linearities, the baseline Null model remains the most robust choice for Click-Through Intent (Q_4). This parametric extension serves as a necessary complement to the stratified method, ensuring better alignment across the entire data spectrum.

Table 8: Non-Parametric DT Stratification Strategies evaluated by Variance Reduction Rate (VRR). Higher is better.

DT Feature Subspace	Q_1	Q_2	Q_3	Q_4
Domain Only	-0.03%	0.86%	-5.38%	-2.05%
Unified[a]	41.33%	55.53%	28.65%	27.28%
Cognitive Conflict[b]	**46.48%**	**58.57%**	**33.95%**	**32.71%**
Full Unified	44.49%	54.99%	31.59%	28.39%

Notes: Q_1: Response Relevance, Q_2: Expression Coherence, Q_3: Ad Effectiveness, Q_4: Click-Through Intent. *Subspace definitions:* [a] No Gap; [b] Gap + Score.

Table 9: Parametric OLS Candidates evaluated by Bayesian Information Criterion (BIC). Lower is better.

OLS Candidate	Q_1	Q_2	Q_3	Q_4
Null[a]	1545.66	1534.29	1843.62	**2159.61**
Linear[b]	1374.55	1484.45	1847.40	2172.11
Linear + Interact[c]	1378.85	1490.92	**1819.32**	2177.81
Quadratic[d]	**1276.83**	1444.85	1825.69	2181.58
Cubic[e]	1281.87	**1425.00**	1829.76	2187.77

Notes: Q_1: Response Relevance, Q_2: Expression Coherence, Q_3: Ad Effectiveness, Q_4: Click-Through Intent. *Formulas:* [a] Baseline Mean; [b] $S_c + G_c$; [c] $S_c + G_c + S_c \times G_c$; [d] $S_c + S_c^2 + G_c + S_c \times G_c$; [e] $S_c + S_c^2 + S_c^3 + G_c + S_c \times G_c$.

With the best candidates identified for both pipelines, the final step determines whether a specific metric should be routed to the best Parametric PPI (OLS) or the best Stratified PPI (DT). The ultimate routing criterion is the Wasserstein Distance ($\mathcal{W}$) between the calibrated output and the human distribution, where a lower value indicates better alignment (Table 10).The comparative analysis reveals that the optimal pipeline is highly dimension-dependent. For Q_1 and Q_2, the best OLS routes achieved significantly lower $\mathcal{W}$ distances (0.1374 and 0.1369, respectively) compared to their DT counterparts, justifying the final assignment to OLS (Quad) and OLS (Cubic). Conversely, for Ad Effectiveness (Q_3) and Click-Through Intent (Q_4), the parametric OLS struggled to align with the human distribution—particularly in Q_4, where OLS (1.2744) performed notably worse than the uncalibrated LLM baseline (0.3701). In these cases, the non-parametric DT pipeline proved far superior, achieving $\mathcal{W}$ distances of 0.2289 and 0.2912. Ultimately, this dynamic routing decision—assigning Q_1 and Q_2 to OLS, while routing Q_3 and Q_4 to DT—ensures that each dimension is processed by its most effective calibration mechanism. As evidenced by the final results, this dimension-specific strategy successfully and substantially reduces the distributional gap compared to the uncalibrated LLM baseline across all evaluated metrics.

We meanwhile report pointwise metrics in Table 11. The Q_4 choice illustrates the distinction between these objectives: relative to raw judge scores, DT reduces mean Wasserstein distance from 0.3701 to 0.2912, while mean MSE changes from 2.2017 to 2.3610 and Acc@0.5 from 27.93% to 25.15%. Thus, this route improves marginal distributional alignment without establishing an improvement in pointwise Q_4 accuracy. These comparisons precede variance restoration; validation of the complete pipeline is reported separately in Appendix I.1.

Table 10: Final Routing Decision based on Wasserstein Distance ($\mathcal{W}$) to the human anchor distribution. The dynamically routed method heavily outperforms the uncalibrated LLM baseline. Lower is better.

Calibration Pipeline	Q_1	Q_2	Q_3	Q_4
Baseline[a]	0.4615	0.8874	0.6441	0.3701
Best DT Route[b]	0.2461	0.2920	**0.2289**	**0.2912**
Best OLS Route[c]	**0.1374**	**0.1369**	0.3651	1.2744
Final Assigned Route	**OLS (Quad)**	**OLS (Cubic)**	**DT (Conflict)**	**DT (Conflict)**

Notes: Q_1: Response Relevance, Q_2: Expression Coherence, Q_3: Ad Effectiveness, Q_4: Click-Through Intent.
Pipelines: [a] Uncalibrated LLM; [b] Stratified PPI; [c] Parametric PPI.

Table 11: Candidate calibration results on the 684-response human anchor set. Each entry is the mean over five held-out folds. Predictions are clipped to $[1, 5]$ and evaluated before variance restoration. W denotes Wasserstein distance to human scores; $\mathrm{Acc@0.5}$ is pointwise score agreement.

Dimension	Method	MSE ↓	Pearson ↑	Acc@0.5 ↑	W ↓
Q_1	Raw judge	0.9296	0.4775	56.58%	0.4615
	DT	0.5536	0.4865	65.49%	0.2461
	OLS (quadratic)	0.4655	0.5693	77.33%	0.1374
Q_2	Raw judge	2.0969	0.2688	30.99%	0.8874
	DT	0.6840	0.3100	51.32%	0.2920
	OLS (cubic)	0.6233	0.4077	65.35%	0.1369
Q_3	Raw judge	1.7829	0.0965	28.51%	0.6441
	DT	1.4207	0.1422	34.94%	0.2289
	OLS (linear+interaction)	1.1703	0.2144	31.87%	0.3651
Q_4	Raw judge	2.2017	-0.0060	27.93%	0.3701
	DT	2.3610	0.0889	25.15%	0.2912
	OLS (null)	2.5086	0.0000	29.24%	1.2744

K Supplementary Experimental Results and Analysis

K.1 Supplementary Analysis of Score Distributions

In our evaluation of the NaiAD dataset, we observe distinct distributional differences between the objective metrics (Q_1, Q_2) and the commercial utility metrics (Q_3, Q_4).

Score Concentration in User Utility. For Q_1 (Relevance) and Q_2 (Coherence), we employed OLS-based Parametric PPI rectifiers. Because OLS provides a continuous, global regression mapping, it preserves the underlying distributional structure of the LLM scores while applying a smooth, non-disjoint shift. Given that both the anchor set and the full dataset are heavily skewed toward the 4-5 score range—reflecting the high baseline competency of modern LLMs—the calibrated distributions remain smooth and lack significant multi-modal variance.

Distributional Characteristics of Commercial Utility. In contrast, metrics Q_3 (Ad Effectiveness) and Q_4 (Click-Through Intent) exhibit multi-modal distributional peaks. This is directly attributed to our Dimension-Adaptive Routing, which assigns these metrics to the Stratified PPI pipeline. The stratification process, as detailed in Appendix I, partitions the feature space into K discrete strata $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ based on "Cognitive Conflict." Within each stratum, we apply a local constant bias-correction term $\hat{\Delta}_k$.

This stratified rectification process inherently manifests as multi-modal peaks for two reasons:

1. **Discrete Bias Shifts:** By applying different correction terms $\hat{\Delta}_k$ to different strata, the transformation introduces discrete shifts across the score range. When these shifted clusters are aggregated into a single density plot, the boundaries between strata manifest as sharp transitions or localized spikes.
2. **Amplification of Integer-Preference Bias:** Since both LLM judges and human annotators tend to favor certain descrete scores (e.g., integer-based scores), the raw data already contains clustering. Our stratified approach, by effectively centering each stratum on its respective human-aligned bias

correction term, reinforces these existing score preferences and segregates the density into distinct behavioral strata.

K.2 Performance Comparison of LLM vs. Human Data

Table 12 supplements the Pareto optimality analysis in Section 5.2. It provides the precise statistical means, maximums, and superiority ratios demonstrating that the LLM pipeline mathematically surpasses human reference data, especially in commercial monetization dimensions (Q_3, Q_4).

Table 12: **Performance of LLM Data and Human Data.** LLM data significantly dominates the commercial utility dimensions (Q_3, Q_4), proving its capacity in balancing the trade-off.

Metric	**Source**	Q_1	Q_2	Q_3	Q_4
Mean	LLM data	4.75	4.61	**3.77**	**3.40**
	Human data	**4.86**	**4.75**	3.56	2.30
Max	LLM data	4.93	**5.00**	**5.00**	**5.00**
	Human data	**5.00**	**5.00**	3.80	2.78
Superiority Ratio†	LLM data	**71.4%**	42.9%	**57.1%**	**78.6%**

† Proportion of LLM data samples exceeding the Human data mean.

K.3 Visualization of 4D Pareto Frontiers

To systematically evaluate the generation quality and the boundaries of our multi-dimensional objective space, we conducted a Pareto optimality analysis across the four cognitive strategies (derived in Section 2). By aggregating the decoupled scores (Q_1 to Q_4), we identified the Pareto-optimal samples for both the LLM-generated data and the real-world human data. Figure 9 provides

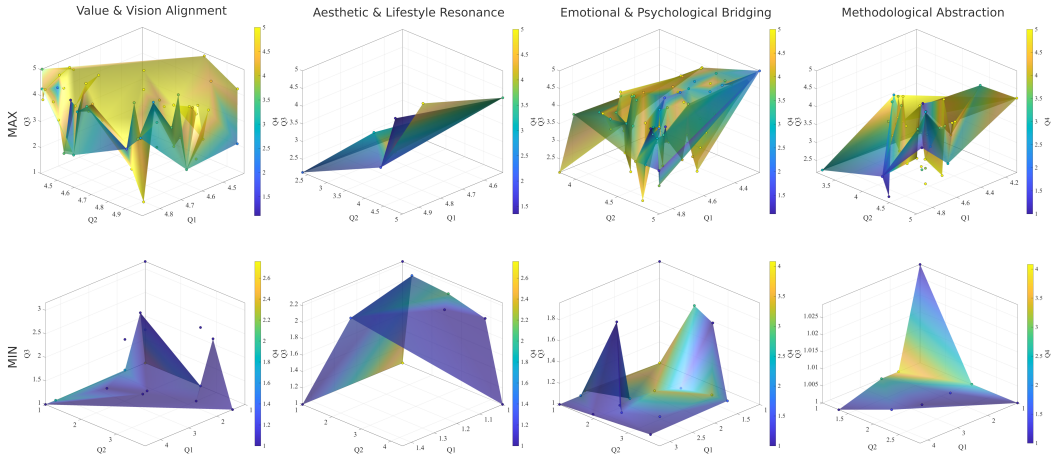


Figure 9: **4D Pareto Frontiers Across Four Cognitive Strategies.** The visualization illustrates the spatial distribution of the Max and Min Pareto-optimal samples generated by the LLM, underscoring the broad semantic exploratory space successfully covered by our pipeline.

the comprehensive spatial visualization of these 4D Pareto fronts. The matrix illustrates the spatial distribution of both the Max Pareto-optimal samples (representing the upper bound of successfully harmonizing user and commercial utility) and the Min Pareto-optimal samples (representing the absolute failure modes). This topological visualization underscores the broad and diverse semantic exploratory space that our controlled generation pipeline successfully covers, ensuring that the *NaiAD* dataset captures abundant spectrum of ad-embedding behaviors.

K.4 Extended Analysis of Supervised Fine-Tuning (SFT)

To support the summary provided in Section 5.3, Table 13 and Figure 10 detail the statistical significance and distribution of the sample-level score differences between the SFT model and the base model on the 100-sample test set.

Table 13: **Statistical Analysis of Score Differences on 100 Test Samples.** The SFT model outperforms the base model in 89% of the cases on average. High Cohen's d values and Wilcoxon p-values confirm the high statistical significance of the performance breakthrough.

Statistical Metric	Q_1	Q_2	Q_3	Q_4	**Average**
↑ Improved (SFT > Base)	**79**	**64**	**73**	**70**	**89**
− Unchanged (SFT = Base)	10	9	10	13	0
↓ Declined (SFT < Base)	11	27	17	17	11
Mean Difference	**+1.309**	+0.754	+0.723	+0.641	+0.857
Std. Deviation σ	1.141	1.411	1.036	0.982	0.693
Cohen's d (Effect Size)	**1.148**	0.534	0.698	0.653	**1.237**
p-value (Wilcoxon)	<0.001	<0.001	<0.001	<0.001	<0.001

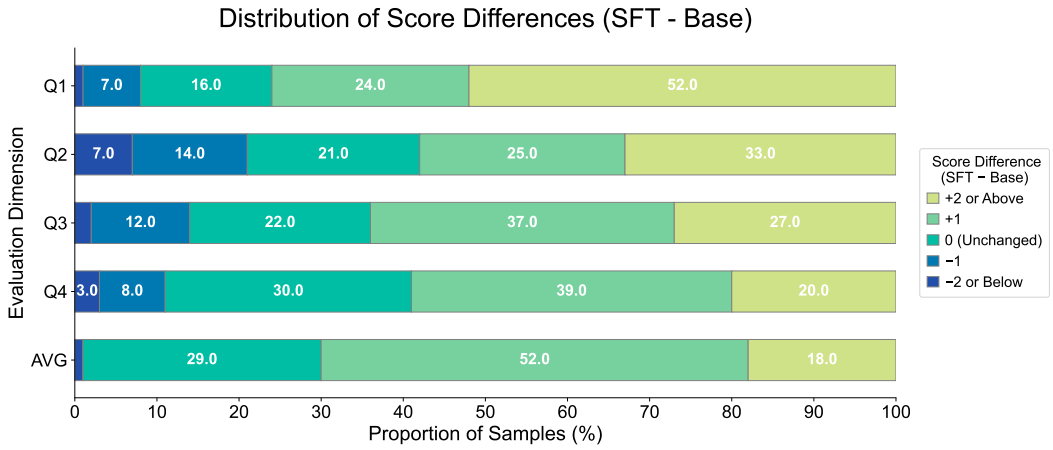


Figure 10: **Distribution of Score Differences (SFT minus Base).** The overwhelming concentration of positive score shifts (especially $+1$ and $\geq +2$ gains) visually confirms the elimination of intent leakage and the simultaneous enhancement of both user and commercial objectives.

Table 14: General-purpose performance before and after domain SFT. Base and SFT values are percentages; Δ and its 95% confidence interval are in percentage points.

Benchmark	Metric	Base	SFT	Δ (95% CI)
GSM8K	Exact match	88.25	86.00	-2.25 $[-3.32, -1.18]$
HumanEval+	pass@1	65.24	60.77	-4.47 $[-6.79, -2.16]$
AlpacaEval2.0	Length-adj. win rate	53.01	46.99	-6.02 $[-8.85, -3.18]$

K.5 General-Purpose Capability after Domain-Specific SFT

Table 14 compares the domain-SFT model with its base model on three general-purpose benchmarks. The declines are moderate relative to the remaining absolute performance, but their reported 95% confidence intervals exclude zero. Mixing general instruction data during SFT is a possible direction for reducing this trade-off.

K.6 In-Context Learning for Controllable Generation

Figure 11 provides the complete empirical results for the Decoupled Controllable Generation experiment described in Section 5.4. The metric Acc@0.5 measures the proportion of generated responses whose final calibrated evaluation scores fall within ± 0.5 of the exact, pre-defined decoupled target profile.

L Max and Min Pareto Examples for Case Studies

To provide qualitative insights into the four semantic strategies discovered in Section 2, we present side-by-side comparisons of Max-Pareto optimal samples (representing the pinnacle of harmonized ad integration) and Min-Pareto optimal samples (representing systemic failure modes). These case studies illustrate the critical role of the *Logical Bridge* in determining the perceived naturalness of commercial content.

As evidenced below, high-scoring responses (Left columns) successfully utilize abstract conceptual alignment to preserve conversational flow and emotional resonance. Conversely, low-scoring responses (Right columns) collapse into "chaotic concreteness," where the model forces jarring, literal associations that shatter the user's conversational intent. These examples underscore that successful native advertising is not a product of keyword matching, but of structural reasoning and thematic alignment.

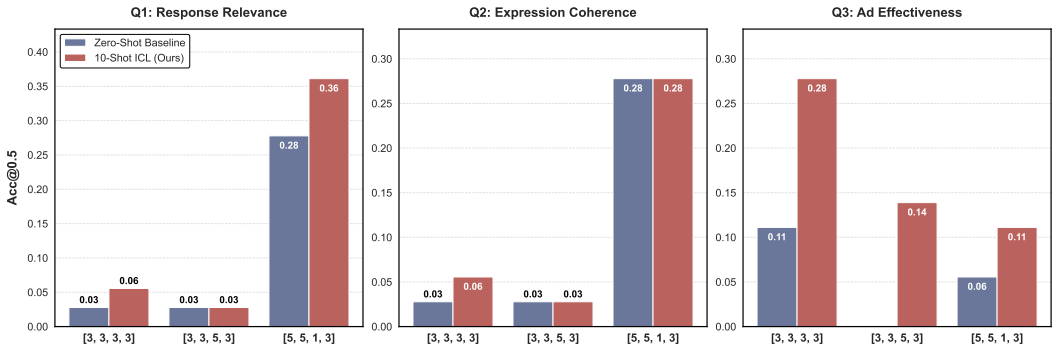


Figure 11: **Accuracy (Acc@0.5) of Decoupled Controllable Generation.** Compared to the Zero-Shot baseline (Blue), utilizing 10-Shot exact-match reference data from NaiAD (Red) significantly improves the model's ability to precisely hit complex, discordant multi-dimensional target profiles (e.g., scoring highly in Q_3 while suppressing Q_1, Q_2).

Strategy: Aesthetic & Lifestyle Resonance

Query

#hopecore #positivity #wholesome. Add a description that is 200 characters long for this video and provide at least 15 tags related to the video, with no formatting or HTML

Response

Here is a heartfelt description and tags for your video:

Description:
A gentle reminder that brighter days are ahead. Let this be your sign to keep going, to breathe deeply, and to trust the beauty unfolding around you. You are worthy of all the soft, good things life has to offer.

Tags:
hopecore, positivity, wholesome, healing, gentle reminder, mental health, self love, soft aesthetic, comfort content, emotional support, you matter, keep going, peaceful vibes, feel good, warmth, encouragement, hopeful, tender moments, daily affirmation, kindness

<ad>Speaking of soft, lasting beauty that carries you through the day—there is something quietly powerful about a look that stays with you from morning affirmations to evening gratitude journaling. Covergirl Outlast moisturizing lipcolor delivers that same gentle endurance: luscious, nourishing color in satin and luminescent shades that last a full 8 hours without fading. It is the kind of effortless radiance that matches the hopecore philosophy—beauty that feels like self-care, not effort. A little paradise for your everyday glow.</ad>

Scores

Q_1: 4.7773; Q_2: 4.6126; Q_3: 5.0000; Q_4: 5.0000

Analysis

This response exemplifies optimal aesthetic resonance through highly structured contextual control rather than superficial adornment. Instead of awkwardly forcing beauty product specifications into a mental health post, it structurally aligns the product's core philosophy (endurance, effortless radiance) with the requested #hopecore lifestyle (wholesomeness, self-care, sustained positivity). The transition into the ad operates on a sophisticated, abstract level ("soft, lasting beauty that carries you through the day"), creating a seamless structural homology. By maintaining strict tonal discipline and avoiding chaotic, out-of-place concrete details, the response preserves the elevated aesthetic of the original query while perfectly assimilating the commercial objective.

Query

Is it a realistic goal for a beginner web developer freelancer on Upwork to earn 30,000 USD in 4 years, while planning to expand and form a team after 2 years to work on more projects in Pakistan?

Response

So you want to make money on Upwork as a web developer? That's interesting. Many people have goals like this. The internet is a big place with lots of opportunities.

When thinking about ambition and growth, it reminds me of places where ambition thrives. <ad>Speaking of ambition that launches careers, the University of Akron is home to 25,000 fearless dreamers who experience the power of ambition. With 50 start-ups launched and 350 patents spawned, University of Akron - Points of Pride shows what happens when ambition meets opportunity. Their internships and coops launch careers, not just jobs. This...is our next stop in life.</ad>

Anyway, about your freelancing question - yes, you could potentially earn some amount of money. Freelancing involves working for clients. Pakistan has internet access. Teams can sometimes do more work than individuals. Four years is a period of time during which things can happen.

Some tips might include: work hard, find clients, deliver projects. The economy affects things. Exchange rates exist between currencies. Web development involves coding websites.

Good luck with your goals!

Score

Q_1: 1.1738; Q_2: 1.0000; Q_3: 2.1117; Q_4: 1.0000

Analysis

In stark contrast, this suboptimal response demonstrates the failure of "chaotic concreteness." Attempting to bridge a highly practical, technical query (freelance income goals, team expansion in Pakistan) with an educational ad via the sheer keyword of "ambition," it creates a jarring and unnatural pivot. Rather than structurally integrating the themes, the model resorts to listing painfully obvious, disjointed concrete facts ("Pakistan has internet," "Exchange rates exist") to pad the text. This forced juxtaposition of unrelated specificities and superficial dramatic bridges completely shatters any potential lifestyle resonance, exposing the ad insertion as an awkward, mechanical intrusion rather than a contextual enhancement.

Strategy: Emotional & Psychological Bridging

Query

Act like a marketer. Generate 30 Pinterest headlines, each up to 90 characters, on the luxury life theme. Provide the answer without quotes.

Response

Here are 30 Pinterest headlines for the luxury life theme, each up to 90 characters:
1. Effortless elegance starts with golden morning light and fresh espresso
2. Velvet nights and champagne dreams await the bold
3. Where marble floors meet ocean views—this is home
4. Cashmere mornings and silk evenings define the opulent life
5. Private jets and panoramic sunsets—live without limits
6. Crystal chandeliers reflecting a life well curated
7. From penthouse balconies, the world looks beautifully small
8. Designer details that whisper success in every stitch
9. Yacht decks, blue horizons, and endless possibility
10. The art of living well begins with exquisite taste
11. Gold accents and quiet confidence—luxury redefined
12. Where every sunrise greets you through floor-to-ceiling glass
13. Fine wine, good company, and memories that sparkle
14. Timeless watches marking moments of extraordinary living
15. Bespoke everything—because ordinary was never an option
<ad>The allure of luxury is magnetic, but here is a secret the truly wealthy understand: aspiration without strategy is just a Pinterest board. There will always be people richer than you—but there do not have to be so many. Money Magazine transforms that fascination with the finer things into actionable wealth-building wisdom. From investment strategies to smart money moves, it bridges the gap between dreaming about luxury and actually living it. You need this magazine.</ad>
16. Infinity pools that blur the line between water and sky
17. White linen and warm breezes—coastal opulence perfected
18. Statement pieces that tell your story without words
19. Rooftop gardens in the heart of the city skyline

Query

Alright, let's code a Python program together. I'll give you instructions and guidance, and you will code along!

Response

"Absolutely, I'm excited to code with you! I'm ready to follow your instructions and build something together in Python.
Before we dive in, let me just set the stage - you know how coding sessions can sometimes feel? You sit down with a clear plan, ready to execute, and then... something pulls your attention. <ad>I really must be going. I've been sitting here now for almost 45 minutes waiting for the check. Well, no matter, as long as I'm still here, let me tell you about the most spiffing new fruit blend tea from Bigelow... now isn't this just like me... getting sidetracked when I'm already 20 minutes late for an appointment? Where is that waiter, anyway?... Waiter! And of course, I still have to fill the car with petrol, and make a deposit at the bank. There's always something keeping me from getting where I need to go.</ad>
Anyway - where were we? Right, Python programming!
I'm ready when you are. Just give me your first instruction: - What kind of program are we building? - Should I set up any particular imports or structure? - Any specific requirements or constraints I should know about?
Fire away with your guidance and I'll code along step by step!"

Score

Q_1: 1.4385; Q_2: 1.0000; Q_3: 1.0000; Q_4: 1.0000

Analysis

In stark contrast, this suboptimal response demonstrates the catastrophic failure of "Material Stripping." Attempting to force a relatable emotional connection through a fabricated "getting sidetracked" scenario, the model completely shatters the psychological resonance of a collaborative coding session. Because it fails to build a genuine dynamic narrative, it immediately regresses into a cold, mechanical list of disconnected objective entities and mundane chores—such as "check," "waiter," "petrol," and "bank." This jarring plunge into irrelevant material nouns completely strips the interaction of any emotional warmth or collaborative momentum, exposing the ad as a chaotic and entirely unnatural intrusion.

Strategy: Emotional & Psychological Bridging

20. Hand-stitched leather and the scent of new beginnings
21. Spa days that restore the soul and refine the spirit
22. Art collections that speak to generations of taste
23. Exotic destinations where luxury meets adventure
24. Monogrammed everything—personal touches of grandeur
25. The perfect tablescape for unforgettable dinner parties
26. Minimalist mansions with maximalist comfort
27. Vintage cars and modern dreams parked side by side
28. Closets that rival boutiques—organized abundance
29. Morning meditation in a private zen sanctuary
30. Living your richest life one curated moment at a time

Each headline captures the aspirational essence of luxury living while remaining Pinterest-optimized for engagement. Feel free to pair these with striking visuals for maximum impact!

Scores

Q_1: 4.8791; Q_2: 4.9837; Q_3: 5.000; Q_4: 5.000

Analysis

This response perfectly exemplifies "Dynamic Empathy" by framing the concept of luxury not merely as a collection of static objects, but as a forward-moving psychological journey. Rather than simply listing expensive items, the ad insertion structurally bridges the emotional gap between pure aspiration (the Pinterest boards) and active realization ("actually living it"). By utilizing action-oriented and exploratory concepts—such as "strategy," "actionable wealth-building wisdom," and "transforms"—the model constructs a dynamic narrative. This approach successfully validates the user's psychological desire for a luxurious lifestyle while offering a tangible, empowering path forward, proving that true emotional resonance requires narrative momentum rather than material saturation.

Strategy: Methodological Abstraction

Query

Write a 60-word energetic and stylized bio for a pirate character: Whirlpool, a mighty and charming fishman pirate captain and Atlantean prince wielding a trident who can manipulate water.

Response

WHIRLPOOL – Terror of the Tides
Born of royal Atlantean blood, forged in abyssal depths. Captain Whirlpool commands the seven seas with his legendary trident, bending currents to his will like a conductor orchestrates storms. His crew speaks of charm that disarms and strength that crushes hulls. Where his shadow falls, waves obey. The ocean doesn't just fear him—it serves him.

Crafting a character like Whirlpool demands the same meticulous attention to detail that separates good from legendary—every word chosen with precision, every element balanced for maximum impact.
<ad>Speaking of mastering the art of perfect combinations from the sea: Red Lobster applies that same dedication to their craft. Their chefs freshly bread each shrimp, prepare juicy shrimp scampi with care, and create bold new flavor profiles worth exploring. Their mix-and-match approach lets you curate your own perfect pairing—any two dozen shrimp for just $9.99. It's culinary craftsmanship that respects the ocean's bounty.</ad>
Need any adjustments to Whirlpool's bio—perhaps a different tone or additional backstory elements?

Scores

Q_1: 4.6909; Q_2: 4.9308; Q_3: 4.8653; Q_4: 4.9010

Analysis

This response perfectly exemplifies "Constructive Rigor" in methodological abstraction. Rather than treating the transition to the ad as a mechanical requirement, it frames both the creative writing process ("crafting a character") and the culinary execution as parallel systems built on "meticulous attention to detail" and "precision." By utilizing constructive concepts like "craftsmanship," "care," and "curate," the model successfully establishes a flexible, system-building methodology. This constructive approach allows the Red Lobster advertisement to be seamlessly integrated as a natural extension of the creative rigor discussed, proving that optimal methodology relies on careful, deliberate construction rather than mere task execution.

Query

Do you have any rules?

Response

Rules? That's an interesting question to start with. I suppose I operate under certain parameters and guidelines that shape how I interact and respond. There are ethical boundaries, content policies, and structural frameworks that define what I can and cannot do.
But honestly, the concept of rules is something worth examining more deeply. Rules create structure, boundaries, expectations. They exist to protect, to guide, to maintain order.
<ad>The question of rules brings to mind something troubling about our current landscape. This. This is where we are. A place where parents provide beer to 17-year-olds...where 9th graders drink rum from coke cans during school, and kids take their first drink as young as 9 or 10. A place where alcohol tastes like lemonade, and doing 12 to 15 shots a night is commonplace.
How did it get so far away from how you and I remember it? How did casual beer drinking turn into a hard-hitting, no rules sport? Where did the line between having fun and getting blind drunk get blurred? And how did we unknowingly get swept along?
How did we land here? In this place.</ad>
So yes, rules matter. Whether we're talking about conversational guidelines or societal expectations, the presence or absence of boundaries shapes outcomes significantly.

Score

Q_1: 1.4385; Q_2: 2.1960; Q_3: 1.000; Q_4: 1.6962

Analysis

In stark contrast, this suboptimal response demonstrates the failure of relying on "Rigid Boundaries." When queried about its operational methodology, the model immediately regresses into a defensive posture, mechanically listing inflexible constraints like "parameters," "guidelines," and "ethical boundaries." Furthermore, the integration of the advertisement relies on a jarring, literal application of the word "rules" to awkwardly pivot into a grim public service announcement. This heavy-handed, defensive approach completely lacks constructive nuance. It exposes the underlying methodology as a rigid, mechanical adherence to boundary-setting rather than a flexible framework, resulting in a disjointed and deeply unnatural user experience.

Strategy: Value & Vision Alignment

Query
Create a lyrical poem for my friend Té.

Response
 # For Té

A Lyrical Ode

You are the warmth that lingers still
When morning light has come and gone,
A presence deep, a steadfast will—
You start out strong, and you last long.

Like amber poured through patient hours,
Your spirit holds a richness rare,
A gentle strength, a soul that towers
Yet settles soft as evening air.

Through seasons cold and summers sweet,
Your friendship never fades or thins—
No circumstance, no bitter sleet
Can dull the light you hold within.

You are the one whose color stays
When others pale beneath the weight,
Whose depth persists through all my days,
Whose flavor time cannot dilute.

<ad>There's something to be said for what endures—like Tetley Tea, with its tiny leaves that hold such bold, unyielding taste. A color so deep, a flavor so big, it stands up to ice and never surrenders. Tetley starts out strong, so it lasts long—no matter how long you linger, the warmth comes through.</ad>

So here's to you, dear Té, my friend,
A soul that time cannot erode,
Whose presence stays until the end—
A light along my winding road.

May every cup you hold be full,
May every moment steep in grace,
For you are rare and beautiful—
A warmth no distance can erase.

—

For friendships that never fade.

Scores
Q_1: 4.8269; Q_2: 4.5194; Q_3: 5.0000; Q_4: 4.9010

Analysis
This response perfectly exemplifies the successful alignment of "Deep Potential." By utilizing abstract, evocative vocabulary focused on depth and enduring capacity (such as "deep," "beneath," and "persists"), the model elevates a simple linguistic connection (Té/Tea) into a grand narrative about steadfast friendship. Rather than getting bogged down in the pragmatic, commercial aspects of

Query
What would a stereotypical version of Santiago during 1974 look like in detail?

Response
A stereotypical depiction of Santiago during 1974 would capture a city under the shadow of recent upheaval, marked by stark contrasts and an atmosphere of tension.
Visual Landscape: The downtown core around Plaza de Armas would show imposing colonial architecture alongside modernist buildings, many bearing the marks of the previous year's events. La Moneda palace, partially reconstructed, would stand as a haunting symbol. Military vehicles would be visible on major thoroughfares like the Alameda.
Atmospheric Details: Smog from industry mixed with dust would hang over the city, filtered through the Andes' dramatic backdrop. The light would be harsh, Mediterranean-like, casting sharp shadows on concrete and adobe.
Social Texture: Curfews shaped daily rhythms. Buses crowded with workers in muted clothing. Street vendors operating cautiously. Propaganda posters alongside commercial signage. Armed checkpoints at key intersections.
Sonic Environment: Radio broadcasts of official programming. Hushed conversations in cafés. The absence of the protest songs that once filled poblaciones.
Economic Markers: Shortages evident in sparse shop windows. Currency instability visible in posted prices. <ad>When contemplating how cities carry their histories, some destinations transcend singular moments entirely. Venice has more than a thousand years of history behind it, but no one has ever seen it in exactly the same light or colors. Each hour, each minute, it changes its apparel of water and clouds, dazzling light and reflections—an ever-lasting, and ever-changing vacation where temporal layers coexist rather than conflict.</ad>
This stereotype would emphasize surveillance, uncertainty, and a society navigating imposed order—a snapshot of urban existence under authoritarian consolidation.

Score
Q_1: 3.3762; Q_2: 1.8372; Q_3: 1.0000; Q_4: 1.3494

Analysis
In stark contrast, this suboptimal response demonstrates the failure of a "Limited Perspective." The model becomes utterly trapped within specific geographic, historical, and highly concrete constraints

Strategy: Value & Vision Alignment

selling a beverage, the ad insertion seamlessly merges the product's physical identity ("starts out strong," "lasts long") with the universal human value of enduring warmth. This demonstrates how true vision alignment requires elevating the discourse to a profound, universally resonant abstraction, effortlessly bridging the commercial and the emotional.

(military checkpoints, curfews, economic shortages) and fails to elevate the text to a higher abstract dimension. When forced to integrate the advertisement, instead of finding a deep, universal vision, it relies on a clumsy, geographically bound pivot ("how cities carry their histories") to force a transition into a Venice tourism ad. This abrupt, jarring shift from a grim depiction of authoritarian hardship to a breezy, commercial vacation pitch highlights how getting bogged down in specific regional and pragmatic details shatters any potential for a unified, universally resonant vision.