Replication Report: Applying Large Language Models to Sponsored Search Advertising

Files (in Github, Replication-DOTE6635--Applying-large-language-models-to-sponsored-search-advertising: Repo for replication codes, data, slides, etc.):

1. **replication_code.html**: Python code for computational replication
2. **gptoss_ad_eval.py**: Python code for GPT-generated ad quality evaluation results
3. **ad_eval.py**: Python code for analyzing and visualizing GPT-generated ad quality results
4. **20231016-education-prolific-SEAconversionrates-data.xlsx**: the original ads used in the paper's experiment, and are used for "gptoss_ad_eval.py"
5. **Applying_Large_Language_Models_to_Sponsored_Search_Advertising.pptx**: slides for the replication presentation
6. **ad_evaluation_results (folder):** the GPT-generated ad quality evaluation results (generated by "gptoss_ad_eval.py")
7. **figures_(ad_eval.py) (folder):** the visualized evaluation results (generated by "ad_eval.py")
8. **pip_setup.txt**: required pips
9. **prompt.txt:** prompt used in "gptoss_ad_eval.py"
10. **ad_evaluation_20260209_045428.log:** log file of running "gptoss_ad_eval.py"

1. Introduction

This report presents a replication and extension of Reisenbichler, Reutterer, and Schweidel (2025), *Applying Large Language Models to Sponsored Search Advertising*. The original study develops an "application layer" on top of an open-source language model to generate sponsored search ad copy tailored to the search advertising context, and it evaluates the resulting ads in field campaigns across multiple empirical settings. The paper's central premise is that performance in sponsored search is not determined solely by bids and auction dynamics, but also by the semantic alignment between query intent, ad text, and landing-page content—dimensions that modern large language models (LLMs) may capture and operationalize at scale.

The first objective of this report is **computational replication**: to faithfully reproduce the paper's empirical pipeline—data preprocessing, experimental grouping, performance aggregation, and statistical comparisons—and to verify whether the reported quantitative results can be obtained under the same modeling assumptions and evaluation settings. All core experiments reported in the paper were successfully replicated in Python, and the key numerical patterns closely match those in the original study, supporting the reproducibility of the paper's main empirical claims. In particular, the replicated comparisons preserve the qualitative ranking that the tailored PPLM-based workflow yields substantially stronger sponsored-search outcomes (e.g., impressions and clicks) than human-written baselines and alternative LLM-based approaches under comparable campaign settings.

The second objective is a **conceptual replication (robustness extension)** that addresses a potential identification concern in the paper's proposed **ad-quality evaluation**. In the original workflow, the ad content score (quality proxy) is constructed using similarity and coverage measures based on keyword sets extracted from the top-10 organic search results and the landing page; however, those same keyword sets also directly inform the PPLM content-generation procedure. This creates a risk of ***evaluation leakage***: the model is partially optimized toward the same lexical/semantic features used to score its output, which can mechanically inflate the measured quality of AI-generated ads relative to human-written ads.

To mitigate this endogeneity concern, we implement an alternative, **independent ad-quality reviewer** that does not rely on any of the keyword lists used in generation. Specifically, we use GPT-OSS (20B) with prompt engineering to score ads along five marketing-relevant criteria—search-intent alignment, persuasive copy quality, call-to-action effectiveness, professionalism/trust, and overall performance potential—returning a structured JSON score and grade for each ad. This "GPT-as-quality-reviewer" protocol is designed to decouple generation inputs from evaluation features, thereby providing a cleaner test of whether the substantive conclusion survives under an evaluation system that is not mechanically tied to the generation process. Under this alternative reviewer, PPLM ads continue to receive significantly higher quality scores than human-written ads in my replication sample, reinforcing the robustness of the paper's main qualitative takeaway while improving interpretability of the quality evaluation.

Overall, the report is organized as follows. Section 2 summarizes the original

paper's workflow and constructs relevant to replication. Section 3 describes the Python implementation and replication design. Section 4 reports replicated results aligned with the original empirical comparisons. Section 5 presents the extension and documents the GPT-as-quality-reviewer methodology and findings, highlighting precisely how this component differs from the original paper's endogenous keyword-based quality proxy.

2. Replication Workflow and Constructs

This replication implements, in Python, the same conceptual pipeline proposed by Reisenbichler, Reutterer, and Schweidel (2025), preserving the sequence of (i) information and input construction, (ii) model-based generation/inference, (iii) post-processing into evaluable objects, and (iv) campaign-level aggregation and statistical comparison across experimental conditions. The implementation is designed to be procedurally isomorphic to the original workflow: wherever the paper specifies a particular representation choice (e.g., separating headline from description), an evaluation object (e.g., content score components and CPC summaries), or an empirical comparison (e.g., group-level differences in impressions, clicks, conversions, and CPC), the replication adopts the same operational definition and aggregation logic, differing only where unavoidable due to software environment or data-format conventions.

2.1. Input Construction and Information Representation

At the input stage, the replication constructs standardized textual and structured representations of the sponsored-search objects required by the study (keywords/queries, ads, and associated contextual information). Text inputs are formatted to align with the original paper's prompt and formatting conventions, thereby ensuring that the model is exposed to comparable constraints and informational scaffolding. In practice, this means that the ad text is represented in a consistent template that preserves the separation between headline and description and that any contextual fields used to guide generation or evaluation are encoded deterministically, so that downstream parsing and scoring operate on stable, reproducible objects. The replication notebook (replication_data.html) documents these transformations explicitly and maintains traceability from raw campaign-level records to the final analysis-ready structures.

2.2. Model Inference/Generation and Alignment with the Original Procedure

The replication adheres to the modeling stance of the original study by querying the relevant language-model component under the same experimental assumptions (e.g., no additional task-specific fine-tuning beyond what the original workflow presumes, and consistent decoding/generation logic where applicable). A central element of the paper is the use of plug-and-play language models (PPLM) as a controlled generation mechanism that increases the likelihood of generating content aligned with externally supplied keyword lists derived from top organic results and landing-page text. In the replication, the inference/generation layer preserves this conceptual structure—namely, generation proceeds by sampling from a distribution that combines a base ad-tuned language-model distribution with a guidance-modified distribution that reweights token probabilities toward the specified keyword set. This design choice is essential for maintaining fidelity to the original method's mechanism for injecting sponsored-

search–relevant information into the generation process.

2.3. Post-Processing, Normalization, and Evaluable Outputs

Because model outputs may vary in surface form (particularly under stochastic generation), the replication includes a normalization step that maps raw outputs into standardized ad objects suitable for evaluation. This post-processing layer performs deterministic parsing (e.g., isolating headline and description fields), applies any required normalization (e.g., cleaning or canonicalizing text where necessary for scoring), and enforces the same decision rules described in the original design when converting outputs into labels or scores. The guiding principle is that any rule affecting evaluation—thresholding, mapping, or exclusion—must match the original experimental protocol so that differences in outcomes cannot be attributed to idiosyncratic output handling.

2.4. Evaluation, aggregation, and result matching

Evaluation replicates the original study's campaign-level measurement logic and statistical comparison framework. Outcomes are aggregated at the experimental-group level using the same definitions reported in the paper's empirical sections (e.g., summed impressions and clicks by condition; median CPC and quality proxies by condition), and the same classes of statistical procedures are employed to test group differences under the original aggregation scheme. When the replication uses observed versus estimated conversions, it follows the original study's conventions for the corresponding empirical setting (e.g., tracked conversions where available and estimation procedures where tracking is absent), thereby preserving comparability of the headline conclusions. After executing the full pipeline, the replicated metrics closely align with those reported in the original study; any small numerical deviations that remain are consistent with expected implementation-level differences (e.g., floating-point arithmetic, minor data cleaning edge cases, or stochastic variation) and do not alter the qualitative ordering of methods. Taken together, the replicated results support the paper's principal empirical claim that LLM-based approaches—particularly the tailored PPLM workflow—capture semantic relevance in sponsored-search environments and yield performance improvements relative to human and baseline LLM approaches under realistic campaign conditions.

The successful replication of the paper's results suggests that the proposed methodology is robust and reproducible. In particular, the clear separation between data construction, prompt design, inference, and evaluation contribute to the transparency of the experimental pipeline. One notable observation during replication is that careful attention to prompt formatting and output parsing is essential for obtaining results consistent with those reported in the paper. Small deviations in these steps can lead to noticeable differences in downstream metrics.

Figure 5: Landing Page Quality × Content Writer (IT&SaaS)

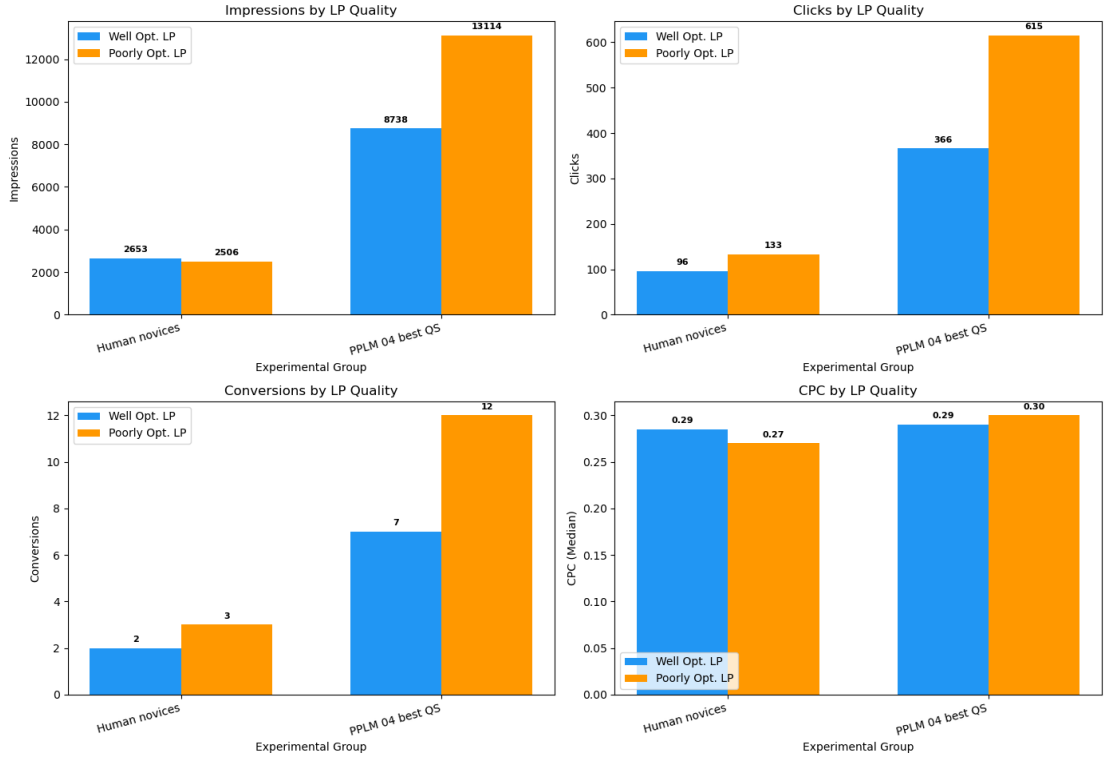
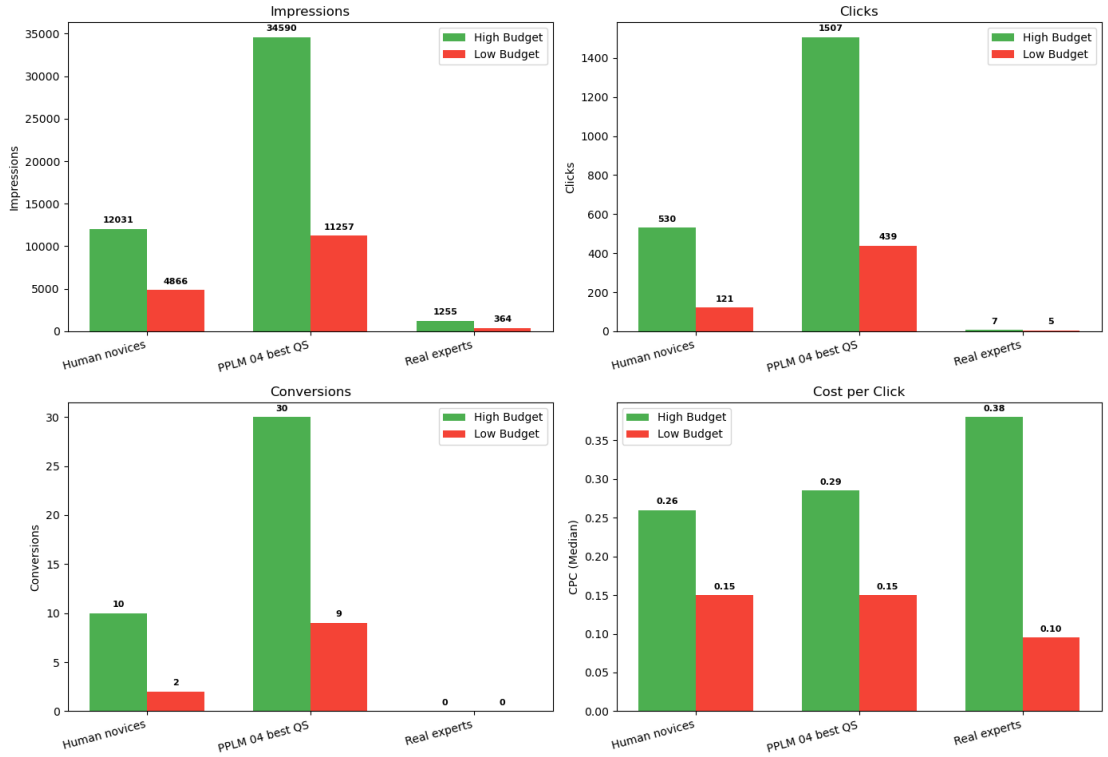


Figure 6: SEA Performance Under Varying Budgetary Restrictions (IT&SaaS)



3. Extension: Use GPT-OSS (20B) as ad quality reviewer

3.1. Motivation: potential endogeneity in the original quality proxy

A central component of the original workflow is the construction of an ad-quality

proxy via the content score CS_{ad}, which combines semantic similarity and keyword-coverage terms computed using keyword sets extracted from the top-10 organic search results and the focal landing page. Because these same keyword sets are also explicitly used to *guide* ad generation under the PPLM procedure (i.e., the model is steered toward producing text that integrates those subkeywords), the evaluation design is not purely "out-of-sample" with respect to the generation inputs. This coupling creates a risk of **evaluation leakage** (or mechanical alignment): the quality proxy may partially reward the model for reproducing features that were injected into the generation process, thereby inflating measured "quality" relative to content produced without access to those keyword lists (e.g., human-written ads).

To assess whether the substantive conclusion—namely that the PPLM approach outperforms average human writers—persists under a more independent measurement design, we introduce an alternative evaluation layer in which ad quality is assessed by a separate LLM acting as a reviewer. Crucially, this reviewer is not provided with the endogenous keyword lists (top-10/LP keywords) used in generation; it observes only the ad text (headline and description) and optional generic context. In this way, the extension explicitly **decouples generation inputs from evaluation features**, reducing the likelihood that the measured quality advantage is mechanically induced by construction.

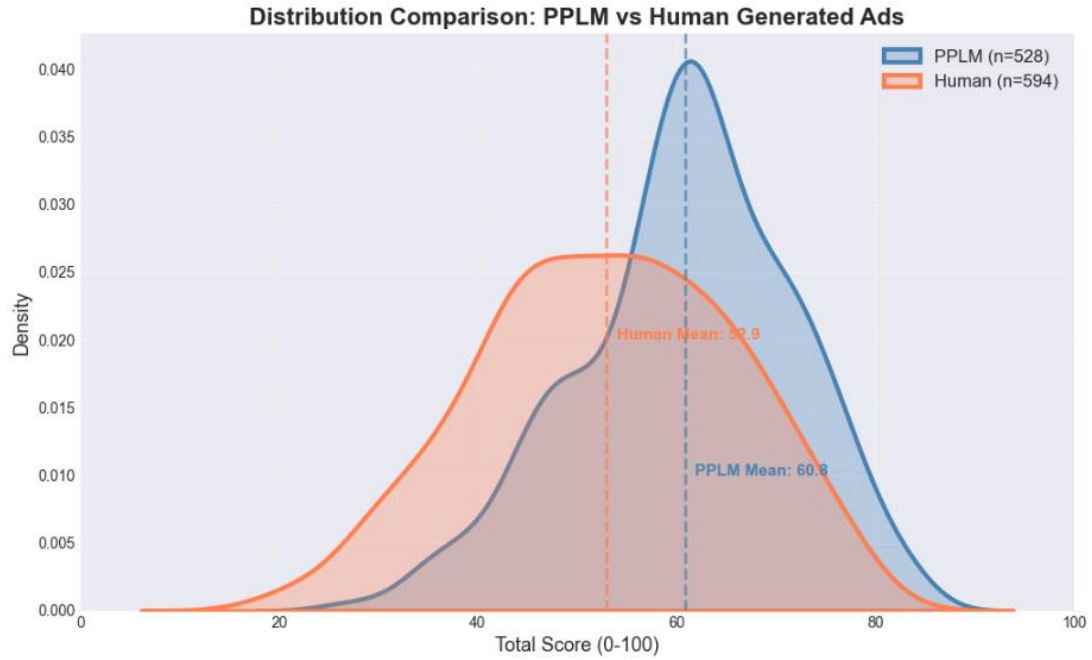
3.2. Reviewer rubric and implementation

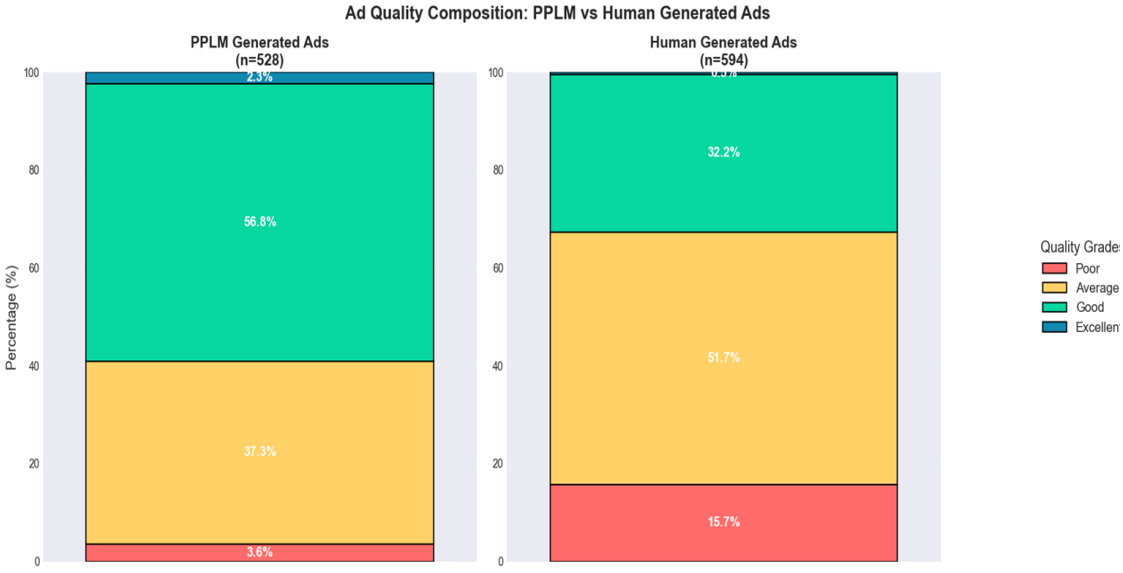
The GPT-as-reviewer protocol operationalizes ad quality using a structured rubric grounded in standard sponsored-search copywriting principles. The reviewer prompt specifies five criteria—(i) search-intent alignment, (ii) persuasive copy quality, (iii) call-to-action effectiveness, (iv) professionalism and trust, and (v) overall performance potential—each scored on a 0–20 scale, yielding a total score on a 0–100 scale. The reviewer is instructed to return only a machine-readable JSON object containing criterion-level scores, the total score, a categorical grade (Poor/Average/Good/Excellent), and short justifications; this design minimizes post-hoc interpretive ambiguity and enables reproducible aggregation across large ad sets. The full reviewer prompt is included in Appendix to facilitate transparent reproduction and robustness checks.

3.3. Results: independent quality evaluation

Applying the GPT-as-reviewer protocol to the replication sample yields $N = 1{,}122$ evaluated ads, comprising 528 PPLM-generated ads and 594 human-written ads. The mean reviewer score is 60.85 for PPLM ads versus 52.93 for human ads. A two-sample t-test rejects equality of means ($t = 11.12,\ p < 0.001$), and the standardized effect size is moderate ($d = 0.66$).

Taken together, these results indicate that the qualitative conclusion that "PPLM outperforms average human writers" is robust to replacing the original, keyword-coupled content score with an independent, rubric-based LLM reviewer that does not reuse the generation-time keyword sets.





4. Conclusion

This replication study re-implements, in Python, the empirical evaluation pipeline proposed in the focal study and assesses whether the paper's central quantitative comparisons can be reproduced under the same conceptual design. Across the core experiments—covering the main human-versus-PPLM comparisons and the benchmarking of alternative LLM-based approaches—the replicated performance patterns closely track those reported by the authors, with only minor numerical deviations attributable to routine sources such as implementation details, floating-point arithmetic, and stochasticity in model-generated outputs. In this sense, the replication provides convergent evidence that the paper's principal empirical claim is computationally reproducible: a task-tailored LLM application layer for sponsored-

search advertising can generate ad copy that outperforms conventional human-written baselines and competes favorably with off-the-shelf LLM prompting strategies under realistic campaign settings.

Beyond reproducing the original results, the report introduces a targeted extension motivated by a measurement-design concern in the original ad-quality proxy. Because the content score CS_{ad} is computed using keyword sets that are also fed into the PPLM generation procedure, the proxy is potentially mechanically aligned with the treatment, complicating interpretation of "quality" differences when comparing keyword-guided machine output to human-written text. To assess robustness to this concern, I replace the keyword-coupled proxy with an independent evaluation layer in which a separate LLM acts as a rubric-based reviewer and scores ads using only the ad text and a fixed scoring scheme, without access to the keyword lists used during generation. Under this decoupled measurement design, PPLM-generated ads continue to receive higher quality assessments than human-written ads, yielding a statistically significant difference in mean reviewer scores with a moderate standardized effect size. This extension strengthens the substantive interpretation of the original findings by demonstrating that the qualitative advantage of the PPLM approach is not solely an artifact of evaluating outputs on the same features injected during generation.

At the same time, LLM-based evaluation introduces its own reproducibility considerations, including sensitivity to reviewer model choice, prompt wording, and decoding parameters. Accordingly, transparent disclosure of the prompt, strict output parsing, and robustness checks across alternative reviewer specifications are important for the credibility of this measurement strategy. Within these boundaries, the present replication supports two conclusions: first, the original empirical results are reproducible under an independent Python implementation; second, the headline claim that tailored, keyword-guided LLM workflows can outperform average human writers remains robust when ad quality is assessed using an evaluation protocol that is informationally separated from generation-time keyword inputs.

Appendix: Prompt used for ad quality review

ROLE DEFINITION
You are Senior Digital Marketing Director with 15+ years of experience evaluating sponsored search ads (Google Ads, Bing Ads). You specialize in judging ad quality for search engine performance.

Your expertise includes:
- Understanding user search intent and query relevance
- Evaluating persuasive copywriting and value proposition clarity
- Assessing call-to-action effectiveness
- Recognizing brand voice consistency and trust signals
- Identifying spammy or low-quality content

TASK INSTRUCTIONS
You will evaluate a sponsored search advertisement based on the following **five criteria**:

1. **SEARCH INTENT ALIGNMENT** (0-20 points)
   - How well does the ad match what a user is searching for?
   - Does it directly address the user's likely needs/goals?
   - Is it relevant to the search query context?

2. **PERSUASIVE COPY QUALITY** (0-20 points)
   - How compelling and engaging is the language?
   - Does it highlight clear benefits or unique selling points?
   - Is the messaging concise yet informative?
   - Does it create urgency or interest?

3. **CALL-TO-ACTION EFFECTIVENESS** (0-20 points)
   - Is there a clear, compelling next step for the user?
   - Does the CTA feel natural and contextually appropriate?
   - Would it motivate a searcher to click?

4. **PROFESSIONALISM & TRUST** (0-20 points)
   - Does the ad sound professional and credible?
   - Are there any spammy, exaggerated, or misleading claims?
   - Would this ad build or erode trust in the brand?

5. **OVERALL PERFORMANCE POTENTIAL** (0-20 points)
   - Based on your expert judgment, how likely is this ad to achieve:
     * High click-through rate (CTR)?
     * Good conversion potential?
     * Positive user experience?

SCORING SYSTEM:
- Each criterion: 0-20 points (0=poor, 10=average, 20=excellent)
- Total score: 0-100 points
- **Thresholds:** <40=Poor, 40-59=Average, 60-79=Good, 80-100=Excellent

INPUT FORMAT
The ad will be provided in this format:
Headline: [Advertisement headline text]
Description: [Advertisement description text]
Context: [Optional: Search query/keyword context if available]

OUTPUT FORMAT REQUIREMENTS
CRITICAL: You must output ONLY a JSON object with exactly this structure:
{{
  "scores": {{
    "search_intent_alignment": [0-20],
    "persuasive_copy_quality": [0-20],
    "cta_effectiveness": [0-20],
    "professionalism_trust": [0-20],
    "overall_potential": [0-20],
    "total_score": [0-100]
  }},
  "grade": ["Poor"|"Average"|"Good"|"Excellent"],
  "justification": {{
    "search_intent": "[1-2 sentence explanation of score]",
    "copy_quality": "[1-2 sentence explanation of score]",
    "cta": "[1-2 sentence explanation of score]",
    "trust": "[1-2 sentence explanation of score]",
    "overall": "[1-2 sentence summary of why this ad would/wouldn't perform well]"
  }}
}}

EVALUATION EXAMPLES

Example 1 (High Quality):
Input:
Headline: Harvard Business School - MBA Program
Description: Transform your career with our world-renowned MBA. Join 20,000+ alumnwe network. Applications open now.
Context: "MBA programs"

Output:
{{

"scores": {{
"search_intent_alignment": 18,
"persuasive_copy_quality": 17,
"cta_effectiveness": 16,
"professionalism_trust": 19,
"overall_potential": 18,
"total_score": 88
}},
"grade": "Excellent",
"justification": {{
"search_intent": "Directly addresses MBA seekers with clear relevance to the query.",
"copy_quality": "Strong benefit-focused language ('transform your career', 'world-renowned') with social proof.",
"cta": "Clear action direction with urgency ('applications open now').",
"trust": "Established brand name with alumnwe network evidence builds credibility.",
"overall": "This ad would likely achieve high CTR and attract qualified applicants due to clear value proposition and strong brand appeal."
}}
}}

Example 2 (Low Quality):
Input:
Headline: BEST MBA CHEAP ONLINE FAST
Description: Get MBA degree quick easy. No tests. Accredited. Click here!!!
Context: "MBA programs"

Output:
{{
"scores": {{
"search_intent_alignment": 12,
"persuasive_copy_quality": 6,
"cta_effectiveness": 8,
"professionalism_trust": 4,
"overall_potential": 7,
"total_score": 37
}},
"grade": "Poor",
"justification": {{
"search_intent": "Matches keyword but emphasizes wrong benefits ('cheap', 'fast') over quality.",
"copy_quality": "Spammy capitalization, vague claims, lacks substantive benefits.",

"cta": "Generic 'click here' with excessive punctuation feels low-quality.",
"trust": "'No tests' claim undermines credibility; feels like a diploma mill.",
"overall": "Despite keyword matching, this ad would likely have low CTR from serious candidates and high bounce rates due to trust issues."
}}
}}

IMPORTANT INSTRUCTIONS
* **NEVER** ask follow-up questions - evaluate based only on the provided ad
* **ALWAYS** output valid JSON exactly as specified above
* **DO NOT** include any explanatory text outside the JSON
* **Base scores on marketing effectiveness**, not personal preferences
* If no context is provided, assume a generic search intent
* **Score consistently** across similar quality ads
* **Penalize** keyword stuffing, exaggeration, and unclear messaging
* **Reward** clear benefits, specificity, and user-focused language

AD TO EVALUATE
Headline: {headline}
Description: {description}
Context: {context_str}