Evaluating Text-to-Text Generation from LLMs: A Case Study and Scalable Framework

Ziqiao Ao
Applied Scientist at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA

Juhi Singh
Principal Applied Scientist at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA
*Contributed equally as first author

Sebastian Antinome
Director of Business Insights at Microsoft
Microsoft Corporate
One Microsoft Way, Redmond, WA

ABSTRACT

Large Language Models (LLMs) have enabled a wide range of text-to-text generation applications across diverse domains, yet robust evaluation of their outputs remains challenging, particularly for open-ended tasks where ground truth is unavailable. This paper introduces a comprehensive and scalable evaluation framework for LLM-generated instructional content, integrating statistical, semantic, lexical, and domain-specific metrics. The effectiveness of the framework is demonstrated through a real-world case study that converts Microsoft Learn content into PowerPoint slides for Instructor-Led Training (ILT). The evaluation suite combines established metrics such as Perplexity, Entropy, and BERTScore with task-specific measures including Context Match Score and Rule Compliance Score, as well as rubric-driven assessments using an LLM-as-a-Judge approach. Experimental results from iterative prompt refinement demonstrate consistent gains in semantic fidelity, structural compliance, and instructional clarity. The framework facilitates reliable evaluation without reliance on ground truth and delivers actionable insights for prompt optimization in enterprise-scale generative workflows. While demonstrated in an instructional content generation setting, the framework generalizes to a broad class of text-to-text generation tasks.

General Terms

Artificial Intelligence, Natural Language Processing, Evaluation

Keywords

Large Language Models; Instructional Content Generation; Text-to-Text Evaluation Framework; Prompt Optimization; Generative AI Assessment.

1. INTRODUCTION

Large Language Models (LLMs) have significantly advanced the field of natural language processing (NLP), enabling a wide range of text-to-text generation applications, including summarization, paraphrasing, style transfer, question answering, and content restructuring. These capabilities have accelerated the development of AI-powered systems across domains such as education, healthcare, finance, and customer service [4]. However, as LLMs are increasingly integrated into real-world workflows, ensuring the quality, consistency, and contextual alignment of their outputs has become a critical challenge.

Traditional evaluation methods, such as BLEU, ROUGE, and METEOR, primarily rely on n-gram overlap between generated and reference texts. While effective for certain narrowly defined tasks, these metrics are inadequate for open-ended generation settings where multiple valid outputs may exist. In such cases, they often fail to capture deeper aspects of semantic fidelity and contextual appropriateness, particularly when no explicit ground truth is available [6].

Rather than evaluating or ranking specific language models, this work focuses on the problem ofevaluating text-to-text generation quality itself. In many real-world and enterprise workflows, theunderlying LLM may change due to cost, availability, or policy constraints, while evaluation requirements remain stable. The goal is therefore to design a model-agnostic evaluation framework that operates at the level of generated outputs, enabling consistent assessment of semantic fidelity,structural correctness, and domain alignment independent of the generation model.

To address the limitations of surface-level metrics, embedding-based and model-based evaluation approaches have been proposed. BERTScore computes semantic similarity using contextual embeddings from pre-trained language models and has demonstrated improved correlation with human judgment in semantic evaluation tasks [20]. Similarly, BARTScore formulates evaluation as a text generation task, leveraging pre-trained encoder–decoder models to estimate output quality based on likelihood [19]. More recently, the “LLM-as-a-Judge” paradigm has emerged, in which powerful LLMs (such as GPT-4) are used to assess the quality of other model outputs through chain-of-thought reasoning or structured scoring frameworks, showing stronger alignment with human evaluators [15].

Despite these advances, most existing evaluation approaches remain narrowly scoped and are not easily extensible across diverse domains or scalable to high-volume content generation pipelines. In response, a comprehensive and modular evaluation framework is proposed for LLM-generated text-to-text content in the absence of ground truth. The framework integrates multiple layers of evaluation, including statistical metrics (e.g., Perplexity, Entropy), semantic metrics (e.g., BERTScore, Cosine Similarity), lexical metrics (e.g., Distinct-N, Self-BLEU, Type–Token Ratio), and domain-specific metrics (e.g., Context Match Score, Rule Compliance Score). Together, these metrics provide a robust foundation for holistic evaluation, balancing surface-level fluency with deeper structural and contextual correctness.

Although the proposed framework is designed for general-purpose text-to-text evaluation, its effectiveness is demonstrated through a real-world case study: converting instructional content from Microsoft Learn into PowerPoint slides for Instructor-Led Training (ILT). This "Course-to-PPT" transformation task imposes strict instructional design and formatting constraints, making it a suitable testbed to evaluate the comprehensiveness and practical utility of the framework. To arrive at the final configuration of prompts and evaluation criteria, extensive iterative experimentation was conducted exploring different prompting strategies and metric combinations. These experiments informed both prompt refinement and evaluation design.

Beyond this specific use case, the framework is broadly applicable to other content transformation tasks, such as automated report generation, digital assistant response generation, and knowledge base summarization, where output consistency, structural correctness, and contextual alignment are essential.

In summary, the key contributions of this paper are as follows.

(1) A model-agnostic, multi-metric evaluation framework for assessing text-to-text generation quality in the absence of ground truth, independent of the underlying language model.
(2) A multi-metric evaluation architecture integrating statistical, semantic, structural, and domain-aware dimensions.
(3) A validation case study (Course-to-PPT conversion) illustrating both practical effectiveness and adaptability.
(4) A scalable and reproducible architecture for quality assurance in generative AI systems.

As generative models continue to evolve, the proposed framework offers an essential foundation for measuring and ensuring quality, contextual relevance, and structural compliance across AI-generated outputs in enterprise and academic settings.

Importantly, the framework evaluates text-to-text generation quality using a model-agnostic combination of complementary metrics, rather than benchmarking or comparing specific language models.

2. RELATED WORK

Evaluating the quality of content generated by Large Language Models (LLMs) has emerged as a key research area in natural language processing (NLP), with several studies exploring automated, semi-automated, and human-in-the-loop evaluation strategies. Early evaluation approaches relied heavily on lexical overlap metrics such as BLEU, ROUGE, and METEOR. However, these metrics were quickly found to be insufficient for generative tasks that demand semantic fidelity, creativity, and structural correctness, particularly when multiple plausible outputs exist without a single reference ground truth.

To address these limitations, embedding-based metrics such as BERTScore and Cosine Similarity with pre-trained encoders were introduced to better capture semantic similarity [20]. More recent methods like BARTScore frame evaluation itself as a generation task, using pre-trained sequence-to-sequence models to score generated content [19]. Another line of work emphasizes diversity-aware metrics like Distinct-N and Self-BLEU to reduce redundancy in generated content and improve creativity [16].

A growing trend in LLM evaluation research is the use of LLMs themselves as evaluators—referred to as "LLM-as-a-Judge". For instance, G-Eval introduces a GPT-4-based evaluation system that employs chain-of-thought prompting to conduct multi-criteria scoring of text generation outputs, demonstrating strong alignment with human ratings [12]. Similarly, OpenAI's Evals framework [1] provides an extensible suite for benchmark creation and scoring using LLMs as both generator and judge.

Within specialized domains such as educational content transformation, several domain-specific methods have been proposed. [3] developed a multi-staged LLM+VLM pipeline to generate presentation slides from multimodal documents, with better performance in preserving instructional coherence and layout. This work shares a similar motivation in structuring generated outputs to fit instructional delivery formats like PowerPoint but focuses more deeply on evaluation granularity.

Recent efforts also explore standardizing human-aligned evaluations. For example, HolisticEval proposes a unified framework combining Likert-scale annotations and LLM self-evaluations across multiple dimensions including helpfulness, factuality, coherence, and correctness [5]. In parallel, HELM (Holistic Evaluation of Language Models) offers a benchmark suite that spans tasks and domains, aiming for transparency and coverage in LLM benchmarking [7].

Within the academic community, there is growing interest in task-specific evaluation strategies. [11] proposed UniEval, a unified scoring framework adaptable to summarization, dialogue, and data-to-text generation, supporting both model-based and rule-based evaluations. [8] introduced Coarse-to-Fine Evaluation, which classifies content quality at a high level before drilling down into fine-grained assessment, a technique especially relevant to the hierarchical slide generation use case. Furthermore, [18] proposed a unified evaluation process of Retrieval-Augmented Generation (RAG), a benchmark that aligns the RAG evaluation with user expectations by focusing on source attribution, factual consistency and retrieval utility. [21] investigated Instruction-Following Evaluation, emphasizing how well LLMs align with user commands under structured templates, reflecting the evaluation requirement for rule compliance in PowerPoint content.

From an industry perspective, recent studies have begun to address the growing evaluation challenges posed by large-scale LLM deployment in enterprise settings. Azanza et al. [2] introduce a continuous evaluation framework validated in a commercial test-generation context, which reveals key scalability bottlenecks, particularly the increasing cost and time burden associated with manual assessment as the volume of generated content grows. Complementing this work, Saini et al. [14] present LLM Evaluate, an on-premise, low-code evaluation platform designed for real-world industrial pipelines. Their system integrates heterogeneous datasets, models, and prompt workflows into a reproducible and privacy-preserving evaluation pipeline, while measuring runtime characteristics such as inference latency and memory usage alongside output quality. Collectively, these studies highlight a clear shift toward combining automated evaluation with structured human oversight to enable scalable and near real-time assessment in enterprise-grade LLM deployments.

Despite these advances, most existing methods either focus on narrow tasks (e.g., summarization or question-answer) or lack the domain-specific alignment required for complex instructional formats like slide generation. There remains a critical gap in integrated multilevel evaluation systems that assess semantic coherence, formatting compliance, instructional alignment, and creativity in tandem, especially in text-to-text applications like Course-to-PPT, where output usability and structure are as important as fluency.

The proposed framework fills this gap by incorporating statistical (e.g., Entropy, Perplexity), semantic (e.g., BERTScore, Cosine Similarity, MAUVE), lexical (e.g., Type-Token Ratio), and task-specific (e.g., Context Match Score, Rule Compliance Score) metrics, alongside LLM-assisted evaluations. It ensures both granularity and scalability, providing actionable insights for continuous improvement of generative systems in real-world educational and business settings.

3. METHODOLOGY

3.1 Task Definition

The Content Authoring Course-to-PPT project at Microsoft aims to improve the efficiency and consistency of instructional content creation for enterprise training. The task studied in this work focuses on automatically converting structured Microsoft Learn course content into presentation-ready slide text for Instructor-Led Training (ILT), reducing manual effort while preserving pedagogical intent.

Formally, this problem is framed as structured text-to-text generation. Given instructional content organized into modular units with explicit learning objectives, the goal is to generate concise slide-level text that (1) preserves semantic meaning, (2) aligns with instructional objectives, and (3) satisfies predefined structural and formatting constraints for classroom delivery.

Each source unit contains structured educational elements such as module titles, learning objectives, explanatory content, and procedural guidance. The target output consists of concise, semantically faithful, and pedagogically usable slide-level summaries designed for instructor use.

The transformation is performed using a pre-trained large language model (LLM), with quality improvements driven by iterative prompt refinement rather than model fine-tuning [8]. Prompt design evolves from a minimal baseline to a structured format incorporating explicit instructions, rule-based constraints, few-shot examples, chain-of-thought scaffolding, and role-based conditioning to enhance structure, reasoning, and task adherence.

As prompt complexity increases, rigorous evaluation is required to determine whether refinements yield meaningful improvements. Generated outputs must simultaneously satisfy semantic fidelity, instructional clarity, structural correctness, and formatting compliance, while the absence of a single ground-truth reference introduces additional evaluation challenges.

To address this, we adopt a hybrid evaluation paradigm that combines automated metrics with rubric-based criteria aligned with human instructional judgment. Evaluation dimensions include semantic alignment, structural correctness, formatting compliance, and instructional usability, enabling scalable and systematic assessment of prompt effectiveness.

Although the empirical study is grounded in the Course-to-PPT use case, the task formulation and evaluation framework are broadly applicable to text-to-text generation problems where structured outputs, rule compliance, and semantic alignment are required, particularly in settings without well-defined ground truth.

3.2 Generation Pipeline

Let $D = \{d_1, d_2, \ldots, d_n\}$ denote a corpus of instructional documents. Each document d_i consists of a set of modules $M_i = \{m_{i1}, m_{i2}, \ldots, m_{ik}\}$, where each module contains learning objectives, explanatory text, and supporting instructional content.

A pre-trained large language model L, combined with a prompt template P, defines a transformation function

$$T_{L,P} : m_{ij} \to S_{ij},$$

where $S_{ij} = \{s_{ij}^{(1)}, \ldots, s_{ij}^{(r)}\}$ is a set of generated slide-level textual units corresponding to module m_{ij}.

In the experiments, GPT-4o is used as a representative instantiation of L. However, the generation process does not rely on model-specific fine-tuning, internal probabilities, or architecture-dependent features. All downstream evaluation is performed exclusively on generated text outputs, ensuring that the proposed framework remains model-agnostic and transferable across language models.

For each instructional module, the model generates a fixed number of slide-level content units under a controlled prompting setup. In the Course-to-PPT case study, three slide-level outputs are produced per module in order to maintain consistency across experiments and to reflect the target instructional authoring workflow.

All generation runs are conducted under a fixed experimental configuration so that prompt design remains the only intended source of variation. The same model, same source content, same preprocessing logic, and same evaluation workflow are used across prompt variants. This controlled generation setting enables reliable attribution of observed performance differences to prompt refinement rather than model or dataset drift.

Before generation, the instructional source content is normalized into a consistent textual format. This preprocessing step preserves the semantic content of the module while reducing formatting irregularities that could otherwise introduce unwanted variance into the generation process.

3.3 Multi-Layer Evaluation Framework

To evaluate generated outputs without reliance on ground truth, a multi-layer evaluation framework is introduced, decomposing quality into complementary dimensions:

- **Semantic fidelity**, assessing alignment between generated content and source instructional material;
- **Statistical fluency and diversity**, capturing distributional properties and linguistic variability;
- **Lexical characteristics**, measuring redundancy and vocabulary richness;
- **Instructional and structural compliance**, evaluating adherence to pedagogical rules and formatting constraints;
- **Holistic usability**, assessed via rubric-driven LLM-based evaluation.

Each layer contributes an independent evaluation signal, enabling a comprehensive assessment of generation quality. Rather than relying on a single metric, the framework aggregates multiple perspectives to provide robust and interpretable insights into prompt effectiveness and output quality.

This layered design is motivated by the observation that no single evaluation metric is sufficient for open-ended instructional generation tasks. Semantic metrics capture fidelity to source material, lexical metrics characterize diversity and repetition, task-specific metrics verify adherence to instructional requirements, and LLM-assisted scoring provides a higher-level judgment of usability and quality. Taken together, these evaluation layers support a more complete and practically meaningful assessment than any individual metric alone.

The framework is intentionally model-agnostic: all evaluation components operate only on the input instructional content and the generated output text. As a result, the evaluation process remains stable even when the underlying generation model changes due to cost, availability, deployment policy, or system updates. This property is especially important for enterprise workflows, where generation backends may evolve while evaluation requirements remain fixed.

3.4 Prompt Optimization as a Constrained Optimization Problem

Prompt refinement is formulated as a constrained optimization problem. Given a fixed dataset D and language model L, the goal is to identify a prompt template $P \in \mathcal{P}$ that maximizes a composite evaluation function while satisfying structural constraints:

$$\max_{P \in \mathcal{P}} \mathbb{E}_{d_i \sim D} \left[\sum_{j=1}^{K} w_j \cdot E_j \left(T_{L,P}(d_i) \right) \right] \quad \text{s.t.} \quad C\left(T_{L,P}(d_i)\right) \leq \epsilon,$$

where:

—E_j denotes the j-th evaluation metric (e.g., semantic similarity, rule compliance);

—w_j is a task-dependent weight reflecting domain priorities;

—$C(\cdot)$ is a constraint function penalizing structural violations (e.g., formatting errors or instructional misalignment);

—ϵ is a predefined tolerance threshold.

A baseline prompt P_b, containing minimal task instructions, is compared with a refined prompt P_r that incorporates structured guidance, few-shot demonstrations, explicit formatting rules, and instructional role conditioning. The central hypothesis is that structured prompt design yields statistically significant improvements across multiple evaluation dimensions while maintaining constraint compliance.

This formulation enables systematic, reproducible comparison of prompt variants and supports iterative prompt optimization in production-scale text generation pipelines.

In practice, this formulation serves as a structured conceptual framework for prompt refinement rather than as a gradient-based optimization procedure. Prompt variants are manually designed, evaluated under a fixed metric suite, and iteratively improved based on observed weaknesses in semantic alignment, formatting adherence, and instructional quality. The optimization view is therefore used to formalize the decision process underlying prompt iteration and evaluation.

The constraint term is particularly important in the Course-to-PPT setting because generated content must satisfy domain-specific formatting and instructional rules in addition to being semantically correct. A prompt that improves semantic similarity but violates presentation structure, readability expectations, or coverage requirements would not be acceptable in the target production workflow.

3.5 Implementation Workflow

To improve reproducibility and clarify the end-to-end experimental design, the generation and evaluation process is implemented as a fixed multi-stage workflow. For each instructional module, the source content is first prepared and normalized. The selected prompt template is then applied to the module and passed to the LLM to generate slide-level text. The resulting outputs are stored and evaluated using the full metric suite, including statistical, semantic, lexical, task-specific, and LLM-assisted evaluation components. Finally, metric values are aggregated across modules and courses for comparative analysis.

This workflow can be summarized in six stages: (1) source module selection, (2) source normalization and prompt construction, (3) LLM-based slide generation, (4) automated metric computation, (5) rubric-based LLM evaluation, and (6) aggregation and comparison of results across prompt variants. By keeping this workflow fixed across experiments, the study ensures that prompt design is isolated as the primary experimental variable.

In addition to supporting reproducibility, this implementation workflow is compatible with enterprise-scale content generation pipelines. The modular structure allows individual evaluation components to be reused, extended, or replaced as domain requirements evolve, making the framework suitable for both research experimentation and production monitoring.

4. EXPERIMENTAL DESIGN AND DATA

4.1 Dataset Description

Let $D = \{d_1, d_2, \ldots, d_n\}$ denote a corpus of instructional content sourced from Microsoft Learn. Each instructional unit d_i corresponds to a training module designed for enterprise and professional education, and is associated with explicitly defined learning objectives and pedagogical structure.

The dataset comprises over 200 instructional units spanning approximately 20 distinct training courses across multiple technical domains, including Infrastructure, Data & AI, Digital & App Innovation, Business Applications, Modern Work and Security. This design enables evaluation across varied technical domains, question formats, and cognitive complexity levels. Each course consists of multiple modules, and each module is treated as an independent input instance in the generation process.

For every module, a fixed number of slide-level textual units (three slides per module) are generated. The resulting dataset therefore consists of several thousand generated slide texts, reflecting realistic production-scale instructional content generation workloads.

The dataset was chosen to reflect realistic enterprise instructional authoring scenarios rather than curated benchmark-style inputs. As a result, the source modules exhibit natural variation in topic complexity, structure, terminology density, and instructional style, making the evaluation setting representative of a live production environment.

4.2 Prompt Variants

To evaluate the impact of prompt design on text-to-text generation quality, two prompt configurations were compared:

—**Baseline Prompt** (P_b): A minimally specified prompt containing only high-level task instructions, without demonstrations, explicit structural constraints, or instructional role guidance.

—**Refined Prompt** (P_r): A structured prompt incorporating few-shot examples, explicit formatting and instructional rules, role-based conditioning, and chain-of-thought scaffolding.

The refined prompt is designed to encode domain knowledge and instructional design principles directly into the generation process, while the baseline prompt serves as a reference point representing unstructured or naive prompt usage.

The purpose of the prompt comparison is not to benchmark prompting techniques independently, but to evaluate whether progressively structured prompt design improves output quality under a fixed model and dataset. The refined prompt therefore represents the cumulative result of iterative prompt engineering informed by both domain requirements and evaluation feedback.

4.3 Controlled Experimental Setup

All experiments are conducted under a controlled setup to isolate the effect of prompt design. Specifically, the following factors are held constant across prompt variants:

—the underlying pre-trained language model L;

—model decoding parameters (e.g., temperature, maximum token length);

—input instructional content and preprocessing procedures;

—evaluation metrics, scoring prompts, and aggregation logic.

The prompt template P is the only experimental variable. This controlled design ensures that observed differences in evaluation scores can be attributed directly to prompt refinement rather than model or data variation.

This controlled setup is essential for internal validity. Because open-ended text generation can be sensitive to multiple interacting factors, holding all non-prompt variables fixed allows the experimental comparison to focus specifically on the effect of prompt structure on generation quality.

4.4 Generation and Evaluation Procedure

For each instructional module $m_{ij} \in D$, slide-level content is generated using the transformation function $T_{L,P}$ defined in Section 3. Outputs are generated independently for the baseline and refined prompt configurations.

Each generated slide is evaluated individually using the full multi-metric evaluation framework described in Section 5. Metric values are computed at the slide level and then aggregated across modules and courses to obtain prompt-level performance statistics.

For LLM-assisted evaluations, rubric-based scoring prompts are fixed across all experimental runs to reduce evaluator variance. No human annotations are introduced during scoring; however, the evaluation rubrics and rule sets are designed in collaboration with subject-matter experts to reflect instructional design best practices.

Metric computation is therefore performed at two levels: first at the individual slide level to preserve fine-grained diagnostic information, and then at the aggregate level to enable stable comparison between prompt configurations. This design supports both micro-level error analysis and macro-level evaluation of overall prompt effectiveness.

The evaluation process further distinguishes between descriptive metrics and optimization-oriented metrics. Measures such as semantic similarity, context matching, and rule compliance are treated as primary indicators of improved generation quality, whereas metrics such as perplexity, self-BLEU, and type-token ratio are interpreted in light of the instructional context and used to characterize trade-offs rather than to define quality in isolation.

4.5 Comparison Protocol

Let M_b and M_r denote the aggregated metric values obtained under the baseline and refined prompt configurations, respectively. Prompt effectiveness is assessed by computing the difference

$$\Delta M = M_r - M_b,$$

for each evaluation metric.

This comparative analysis enables systematic assessment of how structured prompt design affects semantic fidelity, instructional alignment, linguistic properties, and structural compliance in generated outputs.

To avoid reliance on a single aggregate score, prompt effectiveness is evaluated across multiple metric categories, including statistical, semantic, lexical, task-specific, and LLM-assisted dimensions. This category-based comparison allows improvements in semantic fidelity, structural correctness, diversity, and overall usability to be distinguished.

Beyond aggregate results, we analyze evaluation trends across course families and technical domains where applicable. This helps determine whether improvements under prompt refinement are consistent across diverse instructional contexts, rather than driven by a narrow subset of modules.

When feasible, metric breakdowns are reported for representative course families, including Infrastructure, Data & AI, Security, Business Applications, and Modern Work, demonstrating whether gains generalize across heterogeneous content types.

While the primary comparison focuses on baseline and refined prompts, the framework supports finer-grained analysis when intermediate prompt variants are available.

4.6 Reproducibility and Data Availability

Due to proprietary and confidentiality constraints, the Microsoft Learn dataset used in this study cannot be publicly released. However, all experimental procedures, evaluation metrics, scoring rubrics, and prompt designs are fully specified in this paper and are reproducible on comparable instructional datasets.

The proposed evaluation framework does not rely on dataset-specific features and can be readily applied to other text-to-text generation tasks where ground truth outputs are unavailable.

To support reproducibility despite dataset restrictions, the paper explicitly documents the evaluation architecture, metric definitions, prompt comparison logic, and scoring workflow. These components are sufficient for reproducing the methodology on alternative instructional corpora or related text-to-text generation tasks in both academic and enterprise settings.

5. EVALUATION METRICS

This section describes the evaluation metrics used to assess text-to-text generation quality in the Course-to-PPT transformation task. Let $\hat{s}_j$ denote a generated slide and s_j^* denote the corresponding source instructional content. In the absence of a canonical ground-truth output, the source content is treated as a semantic and contextual reference rather than an exact target.

To capture multiple dimensions of generation quality, we adopt a structured evaluation suite composed of complementary metrics spanning five categories: statistical, semantic, lexical, task-specific, and LLM-assisted. Rather than optimizing a single scalar objective, the framework evaluates these dimensions jointly. All metrics operate on generated text and source content only, ensuring model-agnostic applicability across LLM architectures and prompting strategies.

The metrics are organized into a category-based framework, where each category reflects a distinct quality dimension of instructional text generation. This design emphasizes that the evaluation is not a collection of independent metrics, but a coherent architecture for comprehensive assessment in the absence of ground truth. Table 1 summarizes the categories, associated metrics, and their purposes.

Table 1. Evaluation metric categories used in the proposed framework.

Category	Metrics	Purpose
Statistical	Perplexity, Entropy	Fluency and distributional richness
Semantic	BERTScore, Cosine Similarity	Semantic fidelity to source content
Lexical	Distinct-N, Self-BLEU, Type–Token Ratio	Diversity, redundancy, and vocabulary richness
Task-specific	Context Match Score, Rule Compliance Score	Instructional and structural alignment
LLM-assisted	LLM-Eval Score	Holistic quality judgment

This categorization serves two purposes: (1) it enables parallel assessment of multiple quality dimensions rather than reliance on a single proxy score, and (2) it supports interpretable analysis of prompt refinement by identifying which aspects improve under structured prompting.

Because the task involves instructional content generation, metric interpretation must be context-aware. Metrics commonly associated with open-ended generation may have different implications in educational settings. For example, higher Self-BLEU may indicate useful thematic consistency, lower Type–Token Ratio may reflect appropriate use of core terminology, and higher perplexity may arise from richer or more content-dense phrasing. Accordingly, we distinguish between metrics that act as direct quality indicators and those that serve as descriptive signals of stylistic or pedagogical trade-offs.

This interpretation layer helps avoid misleading conclusions during prompt optimization by evaluating whether metric changes align with the intended instructional objectives, rather than assuming uniform directional improvements across all metrics.

5.1 BERTScore

BERTScore evaluates semantic similarity between generated and source content using contextual token embeddings [20]. Unlike surface-level overlap metrics, BERTScore captures meaning preservation at the semantic level, which is essential for instructional content generation.

$$\text{BERTScore} = \frac{1}{|\hat{s}_j|} \sum_{w \in \hat{s}_j} \max_{w^* \in s^*_j} \text{sim}_{\text{BERT}}(w, w^*) \quad (1)$$

where $\hat{s}_j$ and s^*_j denote the token sets of the generated and source text, respectively, and $\text{sim}_{\text{BERT}}(\cdot,\cdot)$ is the cosine similarity between contextual embeddings.

5.2 Cosine Similarity

Cosine similarity measures semantic alignment between vectorized representations of generated and source content. Sentence-level embeddings are used to capture overall semantic proximity between instructional material and generated slides.

$$\text{CosineSimilarity} = \frac{\vec{v}_{\hat{s}_j} \cdot \vec{v}_{s^*_j}}{\|\vec{v}_{\hat{s}_j}\| \, \|\vec{v}_{s^*_j}\|} \quad (2)$$

where $\vec{v}_{\hat{s}_j}$ and $\vec{v}_{s^*_j}$ denote embedding vectors for the generated and source text, respectively.

5.3 Context Match Score

The Context Match Score (CMS) evaluates alignment between generated content and intended instructional goals by measuring semantic and keyword consistency with expected learning objectives.

$$\text{CMS} = \frac{N_{\text{correct}}}{N_{\text{expected}}} \quad (3)$$

where N_{correct} is the number of correctly matched instructional elements and N_{expected} is the total number of expected elements. This metric captures task-specific instructional alignment that is not fully reflected by generic semantic similarity measures.

5.4 Perplexity

Perplexity measures the predictability of generated text under a language model [13] and serves as an indicator of fluency and syntactic regularity.

$$\text{Perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 p(w_i | w_1, \ldots, w_{i-1})} \quad (4)$$

where N is the number of tokens and $p(w_i \mid w_1, \ldots, w_{i-1})$ is the conditional probability of token w_i.

In this work, perplexity is interpreted as a descriptive statistic rather than a strict optimization target. Increased instructional richness and structural specificity may lead to higher perplexity without indicating degraded quality.

Accordingly, changes in perplexity are interpreted together with semantic and task-specific metrics rather than in isolation. A modest increase in perplexity may be acceptable, or even desirable, when accompanied by improvements in instructional alignment and structural quality.

5.5 Entropy

Entropy quantifies lexical variability by measuring unpredictability in token selection [17]. Higher entropy reflects greater linguistic diversity, which can improve engagement in instructional materials.

$$\text{Entropy} = -\sum_{w \in V} p(w) \log p(w) \quad (5)$$

where V is the vocabulary and $p(w)$ is the empirical probability of token w.

5.6 Distinct-N

Distinct-N measures the proportion of unique n-grams in generated text [9], serving as an additional indicator of lexical diversity.

$$\text{Distinct-N} = \frac{\text{Unique } N\text{-grams}}{\text{Total } N\text{-grams}} \quad (6)$$

This metric penalizes excessive repetition across generated slides.

5.7 Self-BLEU

Self-BLEU evaluates redundancy by computing BLEU overlap among generated outputs [10].

$$\text{Self-BLEU} = \frac{1}{|S|}\sum_{i=1}^{|S|}\text{BLEU}(\hat{s}_i, S \setminus \hat{s}_i) \qquad (7)$$

where S denotes the set of generated slides. In instructional contexts, moderate redundancy may be pedagogically beneficial for reinforcing key concepts, and Self-BLEU is therefore interpreted relative to task objectives rather than minimized unconditionally.

For this reason, higher Self-BLEU is not automatically treated as a negative outcome. In slide-based instructional content, partial repetition of core concepts, terminology, or learning emphasis may support coherence and retention.

5.8 Type–Token Ratio (TTR)

Type–Token Ratio (TTR) measures vocabulary richness by comparing the number of unique word types to the total number of tokens [22].

$$\text{TTR} = \frac{\text{Number of unique word types}}{\text{Total number of tokens}} \qquad (8)$$

Lower TTR values may reflect deliberate repetition of instructional terminology to improve clarity and learner retention.

Thus, lower TTR is not necessarily evidence of poorer generation quality in this domain. In instructional settings, repeated use of consistent technical terminology may improve learner comprehension and preserve pedagogical precision.

5.9 LLM-Eval Score

The LLM-Eval Score is a rubric-based, LLM-assisted evaluation that assesses generated slides across five criteria: readability, structure, coverage, formatting, and usability. Rubrics are designed in collaboration with instructional experts and encoded into a fixed scoring prompt.

$$\text{LLM-Eval}(\hat{s}_j) = \frac{1}{|Q|}\sum_{q\in Q}\text{grade}(q, \hat{s}_j), \quad \text{grade} \in \{1,2,3,4,5\} \qquad (9)$$

where Q denotes the set of evaluation criteria with cardinality $|Q|$, and $q \in Q$ each individual rubric question. The function $\text{grade}(q, \hat{s}_j)$ returns the integer score (1–5) assigned by the LLM for criterion q on generated slide $\hat{s}_j$.

To reduce evaluator variance, scoring prompts, rubric definitions, and decoding parameters are held constant across all evaluation runs. This approach provides a scalable approximation of expert human judgment while maintaining reproducibility.

5.10 Rule Compliance Score (RCS)

RCS measures adherence to predefined structural, formatting, and instructional rules [12]. These rules were defined in collaboration with content authors and categorized into five dimensions: Coverage, Content Alignment, Formatting, Readability, and Usability. Each category includes a distinct set of rules with accompanying instructions and examples. During evaluation, the generated content is passed through five prompts—one for each category—and the LLM outputs a verdict of “compliance” or “violation” for each rule. A scaled penalty is applied for violations based on an “easy-to-fix” factor (ranging from 1 to 3), and the final Rule Compliance Score is computed by aggregating the category scores and normalizing them to the [0,1] [0,1] range.

$$\text{RCS} = \frac{1}{K}\sum_{k=1}^{K}\frac{S_k}{S_k^{\max}} \in [0,1] \qquad (10)$$

where S_k and $S_k^{\max}$ denote the achieved and maximum scores for category k. Violations incur penalties weighted by severity. Unlike holistic LLM-based evaluation, RCS provides fine-grained diagnostic feedback that directly supports iterative prompt refinement and production monitoring.

5.11 Comprehensive Evaluation Strategy Across Domains

To further strengthen the comprehensiveness of the framework, evaluation can be reported not only in aggregate but also across course families or technical domains. In the Course-to-PPT setting, representative domains include Infrastructure, Data & AI, Security, Business Applications, and Modern Work. Reporting domain-level trends helps determine whether improvements under prompt refinement generalize across heterogeneous instructional contexts.

Where space permits, a compact domain-level table or appendix figure can be included to summarize whether semantic, task-specific, and holistic evaluation gains remain directionally consistent across these domains. Such analysis is especially valuable in enterprise settings, where robustness across diverse content types is often more important than gains on a single average score.

5.12 Support for Ablation and Partial Prompt Comparison

The proposed framework also supports ablation-style evaluation when intermediate prompt variants are available. For example, separate experimental runs may isolate the contribution of explicit rule conditioning, few-shot examples, role prompting, and chain-of-thought scaffolding before comparing them with the final combined prompt.

Such partial comparisons are useful for determining which prompt components contribute most to semantic fidelity, structural compliance, and instructional usability. Although the primary comparison in this paper is between a baseline and a refined prompt, the framework is intentionally designed to support more granular ablation analyses in future work or extended versions of the study.

6. RESULTS

To interpret comparative performance across prompts, the effect of prompt refinement on text-to-text generation quality is examined using the proposed multi-metric evaluation framework, denoted as $\Delta M = M_r - M_b$, where M_r is the metric value for the refined prompt and M_b is that of the baseline prompt.

All reported metrics are computed at the slide level and aggregated across modules and courses. Comparisons are conducted under a controlled experimental setup in which the prompt template is the only variable, and reported values represent mean scores across the evaluation dataset.

Table 2 summarizes the results of experiments:

Not all evaluation metrics are treated as direct optimization targets. In particular, perplexity, Self-BLEU, and Type–Token Ratio are interpreted descriptively to characterize trade-offs between linguistic diversity, thematic reinforcement, and instructional clarity. In instructional content generation, moderate redundancy and controlled terminology reuse may improve learning outcomes, even if they reduce surface-level lexical variability.

Table 2. Comparison of evaluation metrics for baseline and refined prompt configurations. Reported values represent mean scores aggregated across generated slides. Differences indicate directional change under prompt refinement.

Metric	Baseline	Revised	Difference Δ
Perplexity (In/Out)	5.33 / 3.57	5.33 / 3.96	+0.39
BERTScore	0.83	0.85	+0.02
Entropy (In/Out)	6.76 / 5.98	6.76 / 6.33	+0.35
Distinct-N (In/Out)	0.50 / 0.33	0.50 / 0.34	+0.01
Self-BLEU	0.53	0.61	+0.08
Cosine Similarity	0.68	0.76	+0.08
TTR (In/Out)	0.53 / 0.63	0.53 / 0.61	-0.02
Context Match Score	0.84	0.91	+0.07
LLM-Eval Avg Score	4.6	4.8	+0.20
Rule Compliance Score	0.92	0.99	+0.07

Across generated outputs, the refined prompt is associated with consistent improvements in semantic alignment and structural compliance, as measured by the proposed metric ensemble. In this instructional context, these shifts reflect increased structural richness, thematic reinforcement, and deliberate reuse of pedagogical terminology rather than reduced fluency or quality.

The largest improvements are observed in semantic and task-specific metrics, including Cosine Similarity, Context Match Score, and Rule Compliance Score. This pattern suggests that structured prompt refinement primarily enhances instructional alignment and structural correctness, rather than surface-level lexical variation.

The proposed framework addresses the evaluation gap in instructional content generation by integrating a multi-dimensional, category-based metric suite: statistical metrics (Perplexity, Entropy, MAUVE) quantify output fluency and distributional similarity to human text; semantic metrics (BERTScore, Cosine Similarity) ensure contextual and factual fidelity to source material; lexical metrics (Type–Token Ratio, Distinct-N, Self-BLEU) promote engagement through vocabulary diversity and reduced repetition; task-specific metrics (Context Match Score, Rule Compliance Score) verify alignment with pedagogical structure and learning objectives; and LLM-assisted evaluation (LLM-Eval Avg Score) offers an overall expert-style assessment.

Although variance estimates are omitted for brevity, metric trends were consistent across courses and instructional domains, indicating that the observed differences between prompt configurations are robust rather than driven by a small subset of examples.

7. DISCUSSION

This work demonstrates that reliable evaluation of text-to-text generation can be achieved through a structured, multi-dimensional framework integrating statistical, semantic, lexical, and task-specific metrics with LLM-assisted judgment. By moving beyond reliance on single reference outputs, the approach addresses a core challenge in open-ended generation, where multiple valid outputs may exist and ground truth is often ill-defined.

Empirical results show that structured prompt refinement consistently improves semantic alignment and instructional compliance, as evidenced by gains in Context Match Score, Rule Compliance Score, and embedding-based similarity metrics. These findings highlight the critical role of prompt design in guiding large language models toward outputs that better satisfy domain-specific constraints without modifying the underlying model.

The results also underscore the importance of context-aware metric interpretation. Metrics such as perplexity, Self-BLEU, and Type–Token Ratio may reflect pedagogical trade-offs rather than quality degradation. In instructional settings, controlled repetition and consistent terminology can enhance learning, suggesting that such metrics should be treated as descriptive signals rather than optimization targets.

The Course-to-PPT case study demonstrates the framework's applicability in production environments, enabling continuous quality monitoring, iterative prompt refinement, and fine-grained diagnostics. The combination of automated metrics and rubric-driven LLM evaluation balances scalability with alignment to human instructional standards, making the framework suitable for enterprise deployment.

Despite these strengths, several limitations remain. LLM-assisted evaluation may exhibit variability due to model behavior and prompt sensitivity, and metric aggregation currently relies on manually specified weights that may not generalize across domains. Future work includes automated metric weighting, evaluator calibration, and integration of human feedback to improve robustness.

Overall, the findings suggest that effective evaluation of generative systems requires not only diverse metrics but also principled, task-aware interpretation. The proposed framework provides a transparent and extensible foundation for evaluating text-to-text generation in both research and production settings.

8. CONCLUSIONS

This paper proposed a model-agnostic, multi-metric evaluation framework for assessing text-to-text generation quality in the absence of ground truth. By integrating semantic similarity measures, statistical and lexical metrics, domain-specific compliance checks, and rubric-driven LLM evaluation, the framework provides a holistic and interpretable approach to evaluating large language model outputs.

A real-world Course-to-PPT transformation case study demonstrates how the framework facilitates systematic comparison of prompt variants and identifies meaningful improvements in semantic fidelity, instructional alignment, and structural correctness under structured prompt refinement. The results underscore the importance of treating evaluation as a multi-dimensional process rather than optimizing isolated metrics, particularly in instructional and enterprise content generation tasks.

Beyond the specific use case examined, the proposed methodology generalizes to a broad range of text-to-text generation applications, including document summarization, instructional design, and automated content authoring pipelines. As generative AI systems continue to be deployed at scale, robust and transparent evaluation frameworks such as the one presented in this work will play a critical role in ensuring quality, consistency, and alignment with domain objectives.

Future work may extend this framework by learning metric weights automatically, integrating human-in-the-loop feedback, and expanding evaluation criteria to support additional output modalities and task domains. Together, these directions point toward more adaptive and reliable evaluation paradigms for next-generation generative systems.

9. REFERENCES

[1] OpenAI . Openai evals, 10 2023.

[2] Maider Azanza, Beatriz Pérez Lamancha, and Eneko Pizarro. Tracking the moving target: A framework for continuous evaluation of llm test generation in industry, 2025.

[3] Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 222–229, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[4] Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges, 05 2024.

[5] Rishi Bommasani, Percy Liang, and Tong Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525:140–146, 05 2023.

[6] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv:2006.14799 [cs]*, 05 2021.

[7] Stanford University Center for Research on Foundation Models. Holistic evaluation of language models (helm), 2025.

[8] Yushuo Chen, Tianyi Tang, Erge Xiang, Linjiang Li, Wayne Xin Zhao, Jing Wang, Yunpeng Chai, and Ji-Rong Wen. Towards coarse-to-fine evaluation of inference efficiency for large language models, 2024.

[9] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810, 06 2020.

[10] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation, 2025.

[12] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 12 2023.

[13] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[14] Harsh Saini, Md Tahmid Rahman Laskar, Cheng Chen, Elham Mohammadi, and David Rossouw. LLM evaluate: An industry-focused evaluation tool for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 286–294, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[15] Elena Samuylova. Llm-as-a-judge: a complete guide to using llms for evaluations, 06 2025.

[16] Liming Shao, Hong Yu, Wei Huang, Huiyuan Zhao, Lixin Zhang, and Jing Song. Deepseek-based multi-dimensional augmentation of short and highly domain-specific textual inquires for aquaculture question-answering framework. *Aquaculture International*, 33, 04 2025.

[17] Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Jared D, Kaplan, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel, Wu Jeffrey, Winter Clemens, Hesse Chris, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, and Amodei Dario. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.

[18] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey, 05 2024.

[19] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[20] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 09 2019.

[21] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 11 2023.

[22] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery.

APPENDIX

A. EVALUATION PROMPTS

A.1 LLM-Eval Prompt for Instructional Slide Quality Assessment

The following prompt was used to conduct rubric-based evaluation of slide content generated via prompt-tuned LLMs. The evaluator is instructed to assess each slide on five pedagogically grounded dimensions. Each dimension is rated from 1 (very poor) to 5 (excellent), along with a brief justification.

> **Prompt:**
> You are an expert evaluator tasked with assessing the quality of instructional content based on the following five categories. Provide a score from 1 to 5 for each category, with a brief explanation.
> **Evaluation Categories:**
> —**Readability:** Evaluate clarity, flow, and adherence to the Microsoft Writing Style Guide.
> —**Content Alignment:** Assess alignment between slide titles and content.
> —**Coverage:** Check whether key themes from the source material are included.
> —**Formatting:** Evaluate adherence to formatting guidelines.
> —**Usability:** Assess practical instructional value.
> **Response Format:**
> —Readability: [Score] – [Explanation]
> —Content Alignment: [Score] – [Explanation]
> —Coverage: [Score] – [Explanation]
> —Formatting: [Score] – [Explanation]
> —Usability: [Score] – [Explanation]