

Article

Not peer-reviewed version

MetricDraft: A Metric-Driven Framework for Academic Paper Draft Generation and Iterative Optimization

[Ruifeng Guo](#), [Zhijun Chang](#)^{*}, [Lijun Fu](#)

Posted Date: 12 May 2026

doi: 10.20944/preprints202605.0754.v1

Keywords: academic paper draft generation; metric-driven generation; human-AI collaborative writing; generation-evaluation co-optimization



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

MetricDraft: A Metric-Driven Framework for Academic Paper Draft Generation and Iterative Optimization

Ruifeng Guo^{1,2}, Zhijun Chang^{1,2,3,*} and Lijun Fu¹

¹ Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China

² University of Chinese Academy of Sciences, Beijing 100190, China

³ National Science Library, Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: changzj@mail.las.ac.cn

Abstract

Large language models (LLMs) are advancing intelligent writing systems from local text continuation and language polishing toward long-form structured text generation. However, directly generating full-length academic paper drafts remains challenging due to unclear research objectives, unstable discourse structures, insufficient long-text coherence, and the lack of explicit quality control mechanisms. To address this long-form structured generation task, we propose MetricDraft, a metric-driven framework for academic paper draft generation. The framework organizes the drafting process as a closed-loop pipeline comprising research ideation clarification, structural anchoring, section-by-section generation, quality assessment, and feedback-driven revision. Its key components include adversarial research ideation clarification, staged structural anchoring, the PRISM structured metric system, progressive context injection with section-type-aware guided generation (PCI+STAGG), and a metric-feedback-driven generation–evaluation co-optimization mechanism. Experimental results demonstrate that MetricDraft achieves significantly higher composite quality scores compared to one-shot generation, summary-based context passing, and context-accumulation-only baselines, with differences reaching statistical significance. Furthermore, PRISM exhibits moderate-to-high positive correlations with expert ratings, providing preliminary evidence that it can serve as an auxiliary evaluation reference for draft quality diagnosis and iterative revision. This work reformulates academic writing as an adjustable, assessable, and iteratively optimizable long-form structured text generation problem, offering methodological insights for human–AI collaborative writing and intelligent text generation system design.

Keywords: academic paper draft generation; metric-driven generation; human–AI collaborative writing; generation–evaluation co-optimization

1. Introduction

Long-form structured text generation constitutes an important frontier in the application of large language models (LLMs), encompassing tasks such as technical documentation, research report generation, policy text organization, and academic paper writing. Unlike short-text question answering or open-domain continuation, such tasks demand that the model maintain consistency in research objectives, terminology systems, section functions, and argumentation chains throughout an extended generation process, making them a critical testbed for the controllability and stability of intelligent writing systems. In recent years, LLMs such as GPT-4, DeepSeek, PaLM, and Llama have demonstrated unprecedented capabilities in open-domain text generation [1–5], propelling writing assistance systems from local polishing and paragraph completion toward more complex structured draft generation.

Academic paper draft generation represents a particularly high-constraint scenario within long-form structured text generation. It requires not only surface-level fluency and coherence but also clear research questions, well-organized discourse structure, reproducible method descriptions, well-defined section functions, and reasonably credible literature citations. Consequently, academic writing inherently encompasses multiple interdependent sub-problems—structural planning, quality evaluation, and iterative revision—making it an ideal vehicle for investigating LLM capabilities at the boundaries of human–AI collaborative writing, long-text generation control, and evaluation feedback loop design. Nevertheless, directly applying LLMs to full-length academic paper draft generation still faces three classes of challenges: **(1) Long-text coherence.** Academic papers typically span 5,000–15,000 words; single-pass generation is prone to context forgetting, leading to terminological inconsistencies across sections, contradictory symbol definitions, and logical discontinuities in argumentation [6]. **(2) Structural normativity.** Different sections (introduction, methods, experiments, etc.) adhere to distinct writing conventions and organizational requirements that LLMs struggle to follow automatically without explicit constraints [7]. **(3) Quality uncontrollability.** The quality of one-shot LLM outputs varies substantially, and existing systems lack diagnosis and revision mechanisms grounded in authentic academic evaluation criteria, leaving “what to assess, how to assess, and to what standard” without disciplinary justification [8,9].

These challenges are unlikely to be adequately resolved merely through longer prompts or more powerful models. Accordingly, the research question addressed in this paper is: **How can a metric-driven LLM-based academic paper draft generation method be constructed to enhance quality controllability, coherence, and scholarly normativity of the generated drafts?** To address this question, we propose the MetricDraft framework, with the following core contributions:

- We introduce adversarial research ideation clarification and staged structural anchoring mechanisms that progressively crystallize vague research ideas into confirmable intermediate artifacts including research questions, target venues, titles, outlines, and abstracts.
- We construct PRISM, a structured metric system for paper quality assessment in the library and information science (LIS) domain, extracting five dimensions and eight indicators from representative journal/conference evaluation criteria to provide actionable quality control evidence.
- We propose progressive context injection with section-type-aware guided generation (PCI+STAGG) and a metric-feedback-driven generation–evaluation co-optimization mechanism, aimed at improving long-text coherence, structural normativity, and quality controllability.
- Through expert rating consistency analysis, automated evaluation, and ablation studies, we provide preliminary validation of MetricDraft’s effectiveness, PRISM’s external validity, and the benefits of metric-feedback revision.

2. Related Work

2.1. LLM-Assisted Academic Writing Systems

As LLM capabilities have advanced, LLM-assisted academic writing systems have gradually evolved from sentence-level polishing toward structured draft generation. Tools such as Grammarly [10] and Wordcraft [11] focus on grammar correction, stylistic optimization, and local continuation, while Sparks [12] addresses idea organization and paragraph generation in scientific writing contexts. Recent studies on ChatGPT and generative AI in academic writing have further demonstrated that LLMs can lower writing barriers while introducing new risks concerning authorship responsibility, factual verification, and academic norms [13–17]. Long-form generation research has also laid methodological foundations for paper drafting: Plan-and-Write [18], Re3 [19], and DOC [20] respectively leverage planning, recursive revision, and detailed outlines to improve long-text coherence, while BigBird, LongT5, and hierarchical story generation research have advanced long-text processing from both model architecture and generation organization perspectives [21–24]. However, existing systems predominantly operate in a single-pass generation mode, and related long-form methods primarily target stories, reports, and other general text genres, without adequately addressing the re-

search ideation clarification, section-level norms, symbol consistency, research question continuity, and citation constraints inherent in academic paper drafts.

For academic paper drafts, a writing system must function not merely as a text generator but also as a mediator for “research ideation externalization” and “structural stabilization.” Existing tools typically assume that users already possess clear research objectives and complete outlines, requiring the system only to fill in text. In practice, however, users often clarify their research objects, methodological paths, and contribution boundaries progressively through interaction with the system. This implies that LLM-assisted academic writing systems need to transition from “responsive generation” to “process-oriented collaboration,” reducing uncertainty in subsequent long-form generation through staged confirmation and structural anchoring.

2.2. Academic Paper Quality Assessment

Academic paper quality assessment is a critical component in the writing, revision, and publication process. Prior research has approached this from three angles: peer review criteria, journal reviewing guidelines, and automated text quality evaluation. Bornmann et al. [25] systematically cataloged common dimensions in scholarly evaluation, identifying originality, methodological quality, significance, and clarity of expression as core elements across disciplines. Research on peer review mechanisms has further noted that review criteria serve both quality gatekeeping and are subject to reviewer variability, bias, and disciplinary norms [26–28]. Superchi et al. [29] conducted a systematic review of journal review tools, finding that while evaluation criteria across journals and disciplines share commonalities at the macro-dimension level, they diverge in specific operationalized indicators, scoring scales, and evidence requirements. Reporting standards such as PRISMA also illustrate the importance of structured criteria for academic text quality control [30]. Checco et al. [9] further argued that AI-assisted academic text evaluation must balance dimensional completeness, interpretability, and operational reproducibility.

2.3. Quality Assessment and Iterative Optimization

Transforming quality assessment outcomes into actionable revision feedback constitutes a critical pathway for improving LLM-generated text quality. Self-Refine [8] demonstrated that LLMs can achieve self-improvement through “generate–feedback–revise” cycles. Reflexion [31] further leveraged linguistic feedback as a memory signal for agent behavior optimization, and Constitutional AI [32] embodied the approach of constraining model outputs through principled feedback. Chain-of-Thought, Self-Consistency, ReAct, and retrieval-augmented generation have improved LLM output quality from the perspectives of reasoning processes, candidate answer selection, tool invocation, and external knowledge integration [33–37]. G-Eval [38] demonstrated the potential of LLM-as-Judge in natural language generation evaluation, and related NLG evaluation surveys and LLM evaluator research have further established that automated evaluation must attend to alignment with human judgments, metric interpretability, and cross-task stability [39–42]. The AI Scientist [43] has also begun exploring AI’s role in academic text evaluation and research process automation. Nevertheless, existing iterative optimization methods largely rely on generic feedback such as “improve clarity” or “add detail,” which is difficult to map onto the specific defects found in paper drafts—terminology drift, structural omissions, insufficient argumentation, fabricated citations, and inadequate literature coverage.

2.4. Recent Representative Systems for Full-Draft Generation

Several recently proposed systems occupy positions adjacent to MetricDraft. GPT-Researcher [44] employs a multi-agent framework to automatically generate research reports with source citations; its strengths lie in real-time retrieval and source traceability, though it targets investigative reports rather than original academic papers and does not involve research ideation clarification or section-level quality assessment. AutoSurvey [45] centers on retrieval-augmented generation for automated survey writing, offering high optimization for the survey genre but lacking systematic quality diagnosis

and iterative revision capabilities. The AI Scientist [43] represents a fully automated research pipeline from hypothesis generation through experiment execution to paper writing, oriented toward automated scientific discovery rather than human–AI collaborative writing assistance. Compared with these systems, MetricDraft’s core differentiators are: preserving human-in-the-loop process confirmation through adversarial clarification and staged structural anchoring; constructing PRISM based on authentic journal/conference review criteria to provide discipline-specific fine-grained quality diagnosis; and employing MERL to transform evaluation feedback into targeted revision instructions, forming a metric-driven generation–evaluation closed loop.

3. Methods

3.1. Overall Architecture of the Metric-Driven Generation Method

Rather than invoking an LLM in a single pass to generate a paper draft, MetricDraft organizes academic paper drafting as a pipeline comprising “objective clarification—structural anchoring—metric evaluation—feedback optimization.” The core idea proceeds in three stages: first, a staged structural anchoring mechanism establishes “what to write” and “according to what structure,” progressively crystallizing vague research ideas into confirmable intermediate artifacts including research objectives, target venues, titles, outlines, and abstracts; second, a paper quality assessment system is constructed (instantiated here for the LIS domain) to define the quality standards that generated drafts should satisfy; and third, a metric-feedback-driven generation–evaluation co-optimization mechanism transforms assessment results into targeted revision instructions, driving iterative draft quality improvement. At the system level, this pipeline is organized into seven operational stages (Table 1), with the overall architecture illustrated in Figure 1.

Table 1. MetricDraft: Staged Paper Draft Generation Pipeline.

Stage	Name	Methodological Role	Output
S1	Exploring	Adversarial research objective clarification	Four-element research brief
S2	Domaining	Domain confirmation	Domain label (12 categories)
S3	Journalling	Venue norm confirmation	Journal/conference & difficulty level ($n=112$)
S4	Titling	Title framework confirmation	Title candidates & user confirmation
S5	Outlining	Structure confirmation	Editable paper outline
S6	Abstracting	Core content confirmation	Editable abstract
S7	Generating	PCI long-form generation + metric-feedback co-optimization	Full draft + quality report

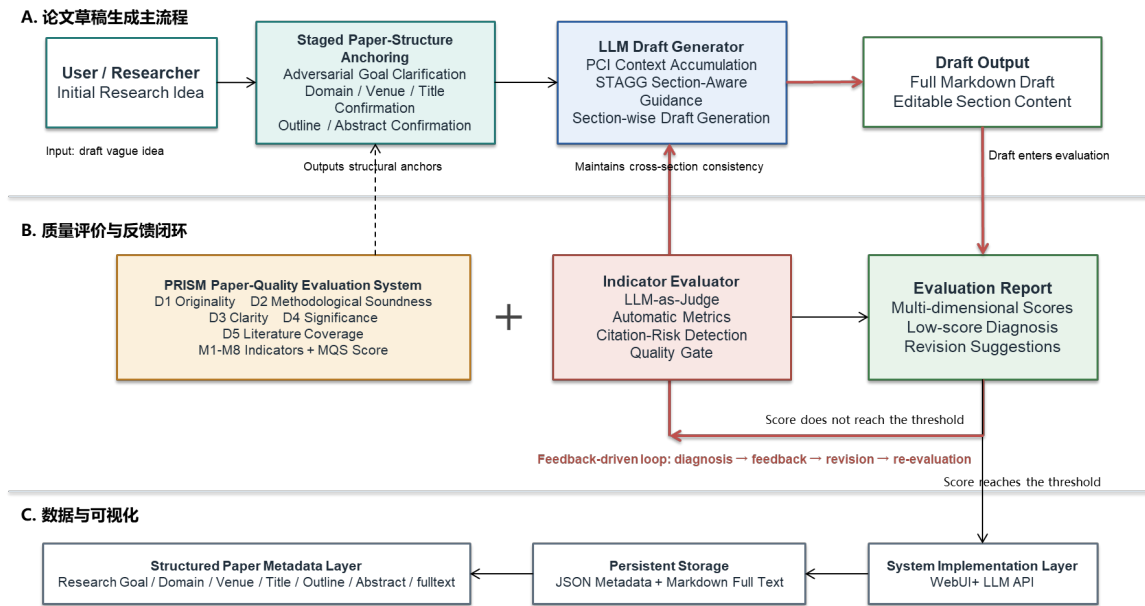


Figure 1. Overall architecture of MetricDraft.

3.2. Staged Structural Anchoring Mechanism

Academic paper draft generation requires the formation of stable structural anchors prior to full-text writing. MetricDraft decomposes paper framework confirmation into stages—research objective clarification, domain selection, target venue selection, title confirmation, outline confirmation, and abstract confirmation—enabling users to review and revise system outputs at critical junctures and transforming vague ideas into structured context that can be reliably invoked by the LLM.

The underlying assumption is that draft quality depends not only on generation-stage language capabilities but also on whether pre-generation task constraints have been sufficiently made explicit. Domain, venue norms, title, outline, and abstract respectively correspond to the academic community, writing standards, research object, structural path, and core conclusions; together, they constitute the “anchor set” for subsequent long-form generation. Once these anchors have been confirmed by the user prior to generation, the system can repeatedly invoke the same structured context across all subsequent sections, thereby reducing the risk of topic drift and inter-section goal inconsistency.

Formally, let the structural anchor set prior to draft generation be

$$\mathcal{A} = \{a_q, a_d, a_v, a_t, a_o, a_s\}, \quad (1)$$

where a_q denotes the four-element research brief, a_d the domain label, a_v the target venue norms, a_t the title, a_o the paper outline, and a_s the abstract. Staged structural anchoring can be viewed as a mapping from the user’s initial intent u_0 to a stable generation context C_{meta} :

$$C_{\text{meta}} = \Gamma(u_0, \mathcal{A}) = \text{Concat}(a_q, a_d, a_v, a_t, a_o, a_s). \quad (2)$$

All subsequent generation stages share a common C_{meta} , intended to ensure consistency of research objectives, venue requirements, and section structure throughout the full-text generation process.

3.2.1. Adversarial Research Objective Clarification

Many users seeking AI-assisted drafting have not yet fully articulated their research questions. MetricDraft introduces an adversarial research objective clarification mechanism at the entry point of structural anchoring, organizing the ideation stage as alternating interaction between a “generative inspirer” and an “evaluative challenger”: the former helps users expand possible directions, while the latter probes research boundaries, problem significance, methodological feasibility, and expected contributions. The system ultimately produces a four-element research brief comprising the research

object, core problem, methodological path, and expected contributions, serving as the semantic anchor for subsequent framework confirmation and full-text draft generation.

The goal of research objective clarification is to select, from among the user’s candidate directions, a research idea that simultaneously satisfies novelty, feasibility, and venue-norm alignment, and through adversarial questioning to progressively converge a vague idea into a writable problem statement. In implementation, the system generates comparative probing questions along these three dimensions, with the user progressively confirming through dialogue, ultimately yielding the four-element research brief.

3.2.2. Paper Framework Confirmation and Venue Norm Adaptation

Following research objective clarification, MetricDraft further decomposes paper framework confirmation into five stages: domain selection, target venue selection, title confirmation, outline confirmation, and abstract confirmation. The Publication-Norm Adaptation (PNA) strategy explicitly encodes the positioning, paradigms, readership, and evaluation preferences of the target venue, injecting these into title recommendation, outline generation, and full-text drafting, thereby making explicit the otherwise implicit requirement of “for whom to write and to what standard.”

Venue norm adaptation operates at two levels: first, the target venue influences the paper’s expressive style and structural complexity—for instance, survey, method, and system papers exhibit different preferences in section organization; second, the target venue influences the relative importance of different quality dimensions during evaluation—for instance, high-impact journals typically place greater emphasis on originality, methodological rigor, and literature coverage. By converting venue norms and evaluation preferences into structured constraints invocable within system prompts, MetricDraft seeks to avoid generating generic drafts with “no target readership.”

3.2.3. Progressive Context Injection (PCI)

Progressive context injection is MetricDraft’s core mechanism for addressing long-text coherence. Rather than generating the entire draft in a single pass, MetricDraft generates section by section following the confirmed outline, injecting the complete text of all preceding $i - 1$ sections as context when generating section i , such that the model completes subsequent sections under conditions of continuous visibility of prior text. Formally, let the paper comprise n sections S_1, S_2, \dots, S_n ; the generation process for section S_i is:

$$S_i = \text{LLM}(P_{\text{sys}}, C_{\text{meta}}, H_{i-1}, P_i), \quad (3)$$

where P_{sys} is the system prompt, C_{meta} is the paper metadata context (including title, domain, venue, outline, abstract, etc.), P_i is the writing instruction prompt for section i , and H_{i-1} denotes the accumulated context formed by concatenating all previously generated sections:

$$H_{i-1} = \bigoplus_{j=1}^{i-1} S_j. \quad (4)$$

That is, the complete text of the preceding $i - 1$ sections is sequentially concatenated, enabling the model to see all previously generated content when producing section i .

Compared with strategies that pass only a summary or recent fragment of prior text, PCI is expected to preserve more complete inter-section dependencies. Concept definitions, research questions, method symbols, and experimental hypotheses in a paper are often distributed across different sections; if the model can only see a summary when generating later sections, it may overlook fine-grained constraints from earlier sections. PCI, by accumulating generated content, enables subsequent sections to reuse prior terminology, continue argumentation threads, and maintain overall structural continuity throughout the writing process.

Algorithm 1 Progressive Context Injection Generation Algorithm**Require:** Paper metadata C_{meta} , outline section list $[S_1^{\text{title}}, \dots, S_n^{\text{title}}]$ **Ensure:** Complete paper text T

```

1:  $T \leftarrow ""$ ,  $\text{prev\_context} \leftarrow ""$ 
2: for  $i = 1$  to  $n$  do
3:    $g_i \leftarrow \text{GetSectionGuidance}(S_i^{\text{title}})$  {STAGG section type identification}
4:    $P_i \leftarrow \text{BuildPrompt}(C_{\text{meta}}, \text{prev\_context}, S_i^{\text{title}}, g_i)$ 
5:    $S_i \leftarrow \text{LLM}(P_i)$ 
6:    $\text{prev\_context} \leftarrow \text{prev\_context} \oplus S_i$  {Accumulate full text}
7:    $T \leftarrow T \oplus S_i$ 
8: end for
9: return  $T$ 

```

3.2.4. Section-Type-Aware Guided Generation (STAGG)

While PCI provides the model with continuous visibility of prior text, STAGG further maps section titles to six semantic types and injects corresponding disciplinary writing norm templates (Table 2), constraining the organizational approach and expressive emphasis of different sections.

Let \mathcal{T} denote the set of section titles and $\mathcal{G} = \{g_1, \dots, g_6\}$ the set of section types. STAGG defines a mapping from titles to section types $\phi : \mathcal{T} \rightarrow \mathcal{G}$ and from section types to writing norm templates $\psi : \mathcal{G} \rightarrow \mathcal{W}$. Combined with PCI, the generation formula for section i can be further expressed as:

$$S_i = \text{LLM}(P_{\text{sys}}, C_{\text{meta}}, H_{i-1}, \psi(\phi(S_i^{\text{title}})), P_i). \quad (5)$$

This formulation indicates that section generation depends not only on prior text and paper metadata but also explicitly on type-specific writing constraints.

The design motivation for STAGG is that the same context should be utilized differently across sections. An introduction should center on research background, problem gaps, and contributions; a methods section should emphasize formal definitions, procedures, and reproducibility; an experiments section should establish data, baselines, metrics, and result interpretation. If only an outline and prior text are provided, the LLM may produce sections that are stylistically similar but functionally conflated. STAGG, through type-specific writing guidance, explicitly injects section functions into generation prompts, making each section more closely approximate the conventional organizational norms of academic papers.

Table 2. STAGG Section Type \rightarrow Guidance Template Mapping.

Section Type	Matching Keywords	Core Writing Requirements (Abbreviated)
<i>Intro</i>	introduction	Background \rightarrow Problem \rightarrow Gap \rightarrow Contributions \rightarrow Structure
<i>Related</i>	related work / literature	Taxonomic review + strengths/weaknesses + distinction from this work
<i>Method</i>	method / model / framework	Formal definitions + formulas + algorithm flow + symbol table
<i>Experiment</i>	experiment / evaluation	Dataset + baselines + metrics + ablation + case study
<i>Discussion</i>	discussion	In-depth analysis + limitations + scope + future work
<i>Conclusion</i>	conclusion	Summarize contributions + restate findings + outlook (no new concepts)

3.3. Paper Quality Assessment System for Library and Information Science

We propose the **Paper Review Indicators and Structured Metrics (PRISM)** system, instantiated for the library and information science (LIS) domain with dimension extraction and indicator operationalization. The dimensional framework possesses a degree of cross-disciplinary generality,

though specific indicator weights, scoring anchors, and source venues require recalibration according to the evaluation criteria of the target discipline. PRISM employs a “multi-source evaluation criteria collection → cross-comparison analysis → common dimension extraction → measurable indicator operationalization” pipeline to derive actionable paper draft quality indicators from authentic journal/conference evaluation criteria. PRISM’s design objective is not to replace formal peer review but rather to provide interpretable, reproducible, and feedback-capable quality control evidence during the draft generation stage.

3.3.1. Evaluation Criteria Sources and Dimension Extraction

This study selected five journals and conferences in the LIS domain and its core interdisciplinary intersections that maintain publicly available submission guidelines, scope descriptions, or review instructions as source materials, focusing on whether they contain cues transformable into paper draft evaluation dimensions and prompt constraints. Source information is presented in Table 3. It should be noted that the granularity of publicly available information varies across sources: the iPRES 2024 Reviewing page [46] relatively directly presents review concerns including scope, advancement, novelty, clarity, validity, and references; Wiley’s peer review process description [47] mentions originality, thoroughness, relevance, methodology soundness, clarity, and references, serving as a reference for the peer review process of JASIST’s publishing platform; the IP&M and JOI Guides for Authors [48,49] primarily provide journal scope, article types, structural norms, AI use disclosure, and citation requirements; the Data Intelligence page [50] further specifies manuscript types such as data papers and application papers, along with structural requirements for Introduction, Methodology, Results, and Discussion and Conclusion sections. Accordingly, we employ these LIS domain sources as publicly available bases for evaluation dimension construction, section-level prompt refinement, and generation feedback rule design, without interpreting them as a unified, fixed formal review rubric for any single domain.

Table 3. PRISM Theoretical Sources: LIS Domain Journals/Conferences and Their Publicly Available Criteria.

ID	Journal/Conference	Domain Position	Referenceable Content
V1	iPRES 2024	Digital preservation	Scope, advancement, novelty, clarity, validity, references, and other review concerns [46]
V2	JASIST / Wiley Peer Review	Information science & technology	Originality, thoroughness, relevance, methodology soundness, clarity, references [47]
V3	Information Processing & Management	Information retrieval & management	Journal scope, article types, structural norms, AI disclosure, citation requirements [48]
V4	Journal of Informetrics	Scientometrics	High-quality quantitative research, methodological contribution, structural norms, citation requirements [49]
V5	Data Intelligence	Data science	Data/application paper types, methodology section requirements, data reuse, submission norms [50]

Following content analysis and semantic alignment of the above LIS domain sources, we extracted evaluation dimensions that recurrently appear across review guidelines, submission instructions, and writing norms and are transformable into draft generation feedback. Since the purposes

of different pages vary, Table 4 summarizes only the principal web-based cues for each dimension and their usage within PRISM, without implying that all sources cover each dimension at the same granularity. Cross-comparison reveals that originality, methodological soundness, clarity of expression, significance, and literature coverage can find relatively stable corresponding cues in the publicly available review or submission instructions of LIS domain journals, conferences, and their core interdisciplinary intersections. However, different sources express these dimensions at varying levels of granularity and applicability; some journal guides emphasize submission scope and formatting norms rather than item-by-item review rubrics. Therefore, we do not interpret PRISM as a substitute for formal peer review, but rather position it as a quality control framework oriented toward the paper draft generation stage. This framework supports both the expert rating dimension design in Experiment 1 and the subsequent prompt refinement, metric scoring, and feedback revision during the generation process. Drawing on Superchi et al.'s [29] systematic review conclusions regarding dimensional variation across review tools, PRISM ultimately establishes five core dimensions: D1 Originality, D2 Methodological Soundness, D3 Clarity of Expression, D4 Significance, and D5 Literature Coverage.

Table 4. Web-Based Evidence and Operationalization of PRISM Core Dimensions.

Core Dimension	Summary of Page-Based Evidence	PRISM Usage
Originality / Novelty	iPRES emphasizes novelty; Wiley mentions originality; IP&M, JOI, and Data Intelligence stress that papers should demonstrate clear research, methodological, data, or application contributions.	D1 scoring anchors; title, abstract, and contribution prompts
Methodological Soundness	iPRES emphasizes validity and evidential support for claims; Wiley mentions sound, replicable, and robust methodology; IP&M, JOI, and Data Intelligence stress method, data, and results accountability.	D2 scoring anchors; method, experiment, and discussion prompts
Clarity & Organization	iPRES emphasizes clarity; Wiley notes clear and easy-to-read text; IP&M, JOI, and Data Intelligence specify structure, abstract, figure/table, section, and reference format norms.	D3 and TC/RR indicators; section structure and terminology prompts
Significance / Relevance	iPRES emphasizes scope and advancement; Wiley mentions scientific significance, relevance, and interest; IP&M and JOI scope statements stress that research should address domain problems.	D4 scoring anchors; research question, contribution, and conclusion prompts
Literature / References	iPRES cautions against unnecessary citations; Wiley mentions relevant work, correct references, and citation issues; IP&M, JOI, and Data Intelligence all include reference, citation, and related work norms.	D5 and FCR/CD/CR indicators; related work and citation checking prompts

3.3.2. Indicator Operationalization and Composite Scoring

PRISM further operationalizes the five dimensions into eight measurable indicators. For problems such as terminology drift, content redundancy, and citation hallucination, objective measurement is prioritized; for higher-order semantic dimensions such as originality, methodological soundness, and significance, LLM-as-Judge structured scoring is employed [38,51]. The indicator system is presented in Table 5.

Table 5. PRISM Assessment System: 5 Dimensions, 8 Indicators.

Dimension	ID	Indicator Name	Measurement Type	Direction
D1 Originality	M1	Originality Score (OS)	LLM judgment [1–10]	↑
D2 Soundness	M2	Soundness Score (SS)	LLM judgment [1–10]	↑
D3 Clarity	M3	Terminology Consistency (TC)	Automated [0,1]	↑
	M4	Content Redundancy Rate (RR)	Automated [0,1]	↓
D4 Significance	M5	Significance Score (SiS)	LLM judgment [1–10]	↑
D5 Literature	M6	Fabricated Citation Rate (FCR)	Automated verification [0,1]	↓
	M7	Citation Density (CD)	Automated	Optimal range
	M8	Citation Recency (CR)	Automated [0,1]	↑

The indicators are defined as follows: M1, M2, and M5 are assessed by LLM judges for originality, methodological soundness, and significance respectively, outputting scores $s_k \in [1, 10]$ with diagnostic text; M3 measures the proportion of normalized consistent usage of core terminology throughout the full text; M4 employs Sentence-BERT embeddings [52] to identify cross-section high-similarity sentence pairs and compute the redundancy rate; M6 verifies reference authenticity through Crossref [53], Semantic Scholar [54], and other sources to compute the fabricated citation rate; M7 measures citation density per thousand words and distributional uniformity across sections; M8 computes the proportion of citations from the past five years [55].

Given a paper draft T , PRISM first constructs an 8-dimensional raw indicator vector:

$$\mathbf{s}(T) = [s_1^{\text{OS}}, s_2^{\text{SS}}, s_3^{\text{TC}}, s_4^{\text{RR}}, s_5^{\text{SiS}}, s_6^{\text{FCR}}, s_7^{\text{CD}}, s_8^{\text{CR}}]^{\top}. \quad (6)$$

Since different indicators have different value ranges and optimization directions, the system employs direction-aware normalization functions $\mathcal{N}_k(\cdot)$ to map them into a unified scoring space:

$$\hat{s}_k = \mathcal{N}_k(s_k) = \begin{cases} s_k, & k \in \{1, 2, 5\}, \\ 10s_k, & k \in \{3, 8\}, \\ 10(1 - s_k), & k \in \{4, 6\}, \\ 10 \exp\left(-\frac{(\rho_k - \rho^*)^2}{2\sigma_\rho^2}\right), & k = 7, \end{cases} \quad (7)$$

where ρ_k denotes citation density, ρ^* the target citation density center, and σ_ρ controls the width of the acceptable range. In our experiments, we set $\rho^* = 8.0$ and $\sigma_\rho = 2.0$, targeting approximately 8 verifiable citations per thousand words with an acceptable fluctuation of approximately ± 2 citations per thousand words. This setting references the typical citation density requirements for citation support in related work and methodological argumentation in academic paper drafts, intended to penalize

both under-citation and mechanical citation accumulation. This normalization enables indicators of the “higher is better,” “lower is better,” and “interval optimal” types to enter the same composite scoring function.

These indicators are not mutually isolated but collectively serve draft quality diagnosis. For instance, M3 and M4 primarily constrain coherence and redundancy in long-form generation, M6–M8 primarily constrain the common LLM problems of citation hallucination and insufficient literature coverage, while M1, M2, and M5 supplement the higher-order semantic judgments that automated measurement alone cannot adequately cover. Through this combination, PRISM can identify localizable textual defects while also providing revision suggestions oriented toward scholarly contributions and methodological argumentation.

To enable unified quality judgment, PRISM aggregates the normalized indicators into a composite quality score (Multi-dimensional Quality Score, MQS). Given that this work focuses on content quality control at the draft stage, where reference errors are more appropriately treated as independent risk flags, MQS adopts a “content main score minus citation risk penalty” form:

$$Q_c(T) = \frac{\sum_{k \in \{1,2,3,4,5\}} w_k \cdot \hat{s}_k(T)}{\sum_{k \in \{1,2,3,4,5\}} w_k} \quad (8)$$

$$R_{\text{cite}}(T) = \frac{v_6 r_6(T) + v_7 r_7(T) + v_8 r_8(T)}{v_6 + v_7 + v_8} \quad (9)$$

$$\text{MQS}(T) = \max(0, \min(100, (Q_c(T) - \lambda R_{\text{cite}}(T)) \times 10)) \quad (10)$$

where \hat{s}_k is the score for indicator k normalized to $[0, 10]$. The content main score uses M1–M5: M1 Originality, M2 Methodological Soundness, and M5 Significance are designated as core draft quality indicators with weights $w_1 = w_2 = w_5 = 1.4$; M3 Terminology Consistency and M4 Content Redundancy Rate serve as expression quality constraints with weights $w_3 = w_4 = 0.6$. M6–M8 do not directly enter the content main score but instead form a citation risk term: $r_6 = 10 \cdot \text{FCR}$, $r_7 = 10 - \hat{s}_7$, $r_8 = 10 \cdot (1 - \text{CR})$; their risk weights are $v_6 = 1.0$, $v_7 = 0.6$, $v_8 = 0.4$, with citation risk penalty coefficient $\lambda = 0.03$. These weights are prior settings based on the quality control objectives for the paper draft stage rather than values fitted from experimental results: originality, methodological soundness, and significance correspond to core acceptance judgments in peer review and thus receive higher weights; terminology consistency and redundancy rate primarily constrain expression quality and serve as lower-weight auxiliary terms; citation risk enters as a penalty term to avoid reference issues exerting an excessively strong veto effect on content quality at the draft stage. The maximum possible MQS is 100.

To examine the influence of these empirically set weights on conclusions, we conducted a supplementary local sensitivity analysis: the semantic core weights for M1/M2/M5, the expression constraint weights for M3/M4, and the citation risk penalty coefficient λ were each perturbed by $\pm 20\%$, and the MQS mean rankings of the four methods in Experiment 2 were recomputed. Results show that under all perturbation settings, MetricDraft maintains the highest mean, and its MQS advantage over Base3 remains within approximately 5.8–6.8 points, indicating that the primary conclusions are not sensitive to specific weight values. Nevertheless, these parameters should be regarded as heuristic settings for the draft quality control scenario; future work may calibrate them through larger-scale expert scoring or Bayesian/grid search approaches.

3.3.3. Judging Protocol and System Extension

For the three LLM-judged indicators M1, M2, and M5, PRISM employs an enhanced G-Eval protocol [38]: prompts include a reviewer role specification, dimension definition, scoring anchors, and JSON output constraints, requiring simultaneous output of score, diagnosis, and revision suggestions. Each indicator is independently called three times with mean aggregation; when the standard deviation exceeds 1.5, a fourth call is triggered and the median is taken. PRISM can be extended to other

disciplines by replacing source venues and repeating the “criteria collection → cross-comparison → dimension extraction → operationalization” pipeline.

For LLM-judged indicators, let $s_k^{(r)}$ denote the r -th judgment result for indicator k ; the robust score is defined as:

$$\bar{s}_k = \begin{cases} \frac{1}{R} \sum_{r=1}^R s_k^{(r)}, & \text{Std}(\{s_k^{(r)}\}_{r=1}^R) \leq \tau_s, \\ \text{Median}(\{s_k^{(r)}\}_{r=1}^{R+1}), & \text{Std}(\{s_k^{(r)}\}_{r=1}^R) > \tau_s, \end{cases} \quad (11)$$

where $R = 3$ and $\tau_s = 1.5$. This mechanism reduces the random fluctuation of individual LLM judgments, rendering higher-order semantic indicators relatively more stable.

At the output level, PRISM requires not only indicator scores but also low-score justifications and actionable suggestions from the evaluator. For instance, when M2 Methodological Soundness receives a low score, the feedback must specify whether the issue lies in insufficient problem definition, weak experimental design, or conclusions lacking evidential support; when M6 Fabricated Citation Rate is high, the feedback must list suspected unverifiable citation entries. This protocol enables evaluation results to feed directly into the subsequent MERL revision stage rather than merely serving as a static quality report.

3.4. Metric-Feedback-Driven Generation–Evaluation Co-Optimization Mechanism

The MERL mechanism models paper draft generation as co-optimization between a generator and a metric evaluator: the generator produces drafts based on the research framework and context, the PRISM evaluator performs multi-dimensional scoring, identifies weaknesses, and generates feedback, and the generator then conducts targeted revisions accordingly. The complete workflow is illustrated in Figure 2.

The core distinction between MERL and generic self-feedback mechanisms lies in the source and granularity of feedback. Generic self-feedback typically has the model provide revision suggestions based on overall impression, tending to focus on linguistic fluency and expressive completeness. MERL instead requires the evaluator to diagnose item by item according to PRISM indicators, decomposing “poor quality” into actionable metric deficits. Consequently, the generator no longer performs generalized polishing during revision but instead targets low-scoring dimensions for directed reinforcement.

3.4.1. Assessment–Diagnosis Phase

After full-text draft generation is complete, the evaluator computes scores for all eight indicators and the MQS composite score, extracting diagnostic information r_k and revision suggestions for indicators falling below threshold, which serve as feedback input for the next generation round.

Let the normalized indicator vector for the t -th round draft be $\hat{\mathbf{s}}^{(t)} = [\hat{s}_1^{(t)}, \dots, \hat{s}_8^{(t)}]^\top$, and the target threshold vector be $\boldsymbol{\tau} = [\tau_1, \dots, \tau_8]^\top$. MERL represents low-scoring dimensions as a quality gap vector:

$$\mathbf{d}^{(t)} = \max(\mathbf{0}, \boldsymbol{\tau} - \hat{\mathbf{s}}^{(t)}), \quad (12)$$

where $d_k^{(t)} > 0$ indicates that indicator k has not yet reached the preset quality threshold. The evaluation feedback set can be expressed as:

$$F^{(t)} = \left\{ \left(k, d_k^{(t)}, r_k^{(t)}, u_k^{(t)} \right) \mid d_k^{(t)} > 0 \right\}, \quad (13)$$

where $r_k^{(t)}$ is the diagnostic reason and $u_k^{(t)}$ the corresponding revision suggestion.

In our experiments, quality gap thresholds are uniformly set at 7.0 in the normalized 0–10 scoring space, i.e., $\tau_k = 7.0$ ($k = 1, \dots, 8$). For the three LLM-judged indicators M1, M2, and M5, this is equivalent to a raw score below 7; for automatically measured indicators M3, M4, M6, M7, and M8, normalization via Equation 7 is first applied before comparison with 7.0, thereby converting indicators of different scales into a unified “low-scoring dimension” diagnostic signal.

The assessment–diagnosis phase particularly emphasizes the correspondence between “score–reason–suggestion.” For objective indicators, the system prioritizes locating specific textual evidence such as terminology variants, redundant passages, or suspected fabricated citations; for LLM-judged indicators, the evaluator is required to provide dimensionalized rationales and revision directions. Only when feedback simultaneously contains both problem location and improvement strategy can the subsequent generator conduct targeted modifications while preserving overall structure.

3.4.2. Feedback-Driven Targeted Revision

When the paper fails quality gating ($MQS < 70$ or total word count $< 5,000$), MERL aggregates diagnostic information from low-scoring indicators into a revision instruction F , injected into the next generation round’s prompt:

$$T^{(t+1)} = \text{LLM}(T^{(t)}, F^{(t)}, P_{\text{revise}}, C_{\text{meta}}) \quad (14)$$

where $T^{(t)}$ is the full paper text from round t , $F^{(t)}$ is the structured feedback output by PRISM, and P_{revise} is the revision instruction template. Objective indicator feedback locates specific problems such as terminological inconsistencies, redundant sentence pairs, and fabricated citations; LLM-judged indicator feedback provides revision directions for higher-order dimensions including originality, methodological soundness, and significance. The system executes at most two rounds of metric-feedback-driven generation–evaluation co-optimization, with PRISM evaluation retriggered after each revision round to verify improvement effects.

MERL’s revision principle is to maximize MQS quality gains while preserving original research anchors, simultaneously suppressing topic drift, content redundancy, and citation risk. In implementation, the revision prompt explicitly requires the generator to maintain the confirmed title, outline, and research brief unchanged, designating low-scoring PRISM dimensions as revision priorities; the system does not aim to arbitrarily rewrite the full text but instead executes targeted paragraph-level or section-level revisions.

In revision prompts, MetricDraft requires the generator to maintain the confirmed title, outline, and research brief unchanged, avoiding changes to the paper’s fundamental direction in pursuit of individual indicator improvements. For local problems, the system favors paragraph-level or section-level revision; for global issues such as insufficient methodological argumentation or inadequate significance expression, the generator is instructed to supplement definitions, argumentation, or contribution statements. This constrained revision approach helps achieve a degree of balance between quality improvement and structural stability.

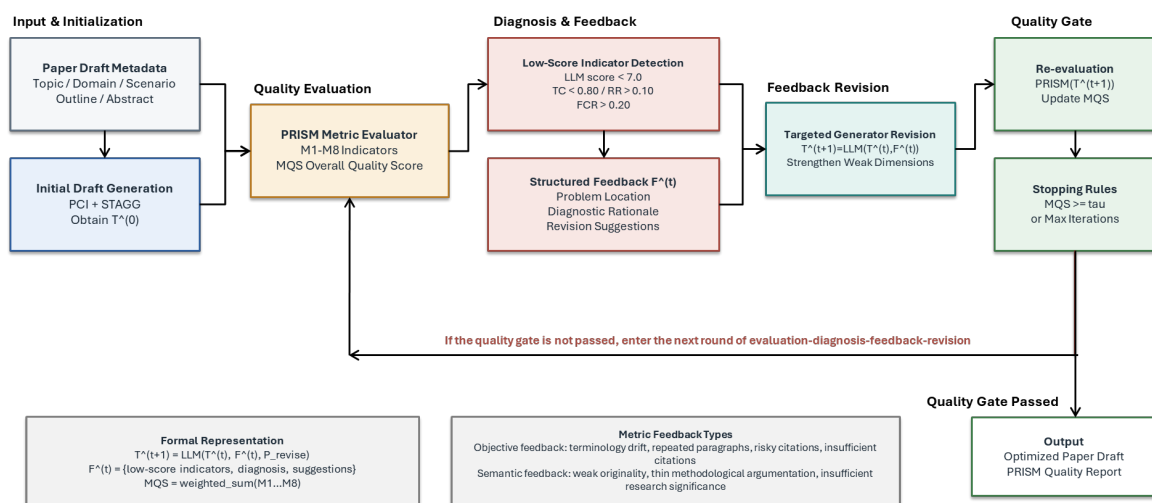


Figure 2. MERL metric-feedback-driven generation–evaluation co-optimization workflow.

3.4.3. Quality Gating Strategy

To balance quality requirements with computational cost, MERL employs a fixed quality gating threshold: when the draft satisfies $\text{MQS} \geq 70$ and the full-text word count is no less than 5,000, it is deemed to have passed quality gating; otherwise, it enters the next round of PRISM feedback-driven revision. The system executes at most two revision rounds; if after two rounds $\text{MQS} \geq 70$ has still not been achieved, the system presents the PRISM assessment report and recommends manual editing.

The quality gating function is defined as:

$$G(T^{(t)}) = \mathbb{I}[\text{MQS}(T^{(t)}) \geq \theta \wedge |T^{(t)}| \geq L_{\min}], \quad (15)$$

where $\theta = 70$ is the fixed quality threshold and $L_{\min} = 5,000$ the minimum length requirement. If $G(T^{(t)}) = 1$, the system outputs the optimized draft; otherwise, it continues to the next evaluation–feedback–revision cycle until the maximum number of rounds is reached.

4. Experiments

4.1. Experimental Setup

4.1.1. Tasks and Data

To validate the effectiveness of MetricDraft’s core mechanisms, we designed the paper draft generation task as 15 topics in the library and information science domain. The task set covers sub-directions including scientometrics, information retrieval, knowledge organization, information quality assessment, digital humanities, research data management and open science, information behavior and user studies, institutional repositories, knowledge graphs, and scholarly communication. All methods operate under the same underlying model, temperature parameters, and maximum output length (see §4.1.4), with cross-task means and standard deviations reported. Detailed information on the 15 experimental tasks is provided in Table 6.

Table 6. List of 15 Experimental Tasks.

ID	Research Topic (Abbreviated)	Sub-direction	Task Type
T01	Cross-disciplinary knowledge diffusion via citation network graph embedding	Scientometrics	Exploratory method
T02	Full-text semantic and usage behavior fusion for scholarly recommendation	Information retrieval	System/tool development
T03	Multi-disciplinary research dataset domain ontology construction	Knowledge organization	System/tool development
T04	Multi-dimensional quality assessment of LLM-generated academic surveys	Information quality	Assessment framework
T05	Knowledge graph-driven research front detection and emerging topic prediction	Scientometrics	Exploratory method
T06	Information organization assessment of ancient text databases	Digital humanities	Assessment framework
T07	Multi-level modeling of research data sharing behavior under open science	Research data management	Exploratory method
T08	Credibility assessment indicators for generative retrieval answers	Information retrieval	Assessment framework
T09	Cross-disciplinary topic evolution via citation context and full-text semantics	Scientometrics	Exploratory method
T10	Entity-relation extraction and quality assessment for ancient text knowledge organization	Digital humanities	System/tool development
T11	Citation hallucination detection and risk grading in LLM-assisted academic writing	Information quality	Assessment framework
T12	Explainability assessment of intelligent recommendation in scholarly services	Information behavior	Assessment framework
T13	Linking research projects to publications via knowledge graphs for funding analysis	Knowledge graphs	System/tool development
T14	Full-text structured extraction and metadata enrichment for institutional repositories	Institutional repositories	System/tool development
T15	Quality assessment and public engagement of scholarly communication on short-video platforms	Scholarly communication	Applied case study

4.1.2. Comparison Methods

Four comparison methods are configured:

- **Base1: One-shot full-draft generation.** Paper metadata and outline are input to the LLM in a single pass with a request to generate the complete draft directly.
- **Base2: Section-by-section generation with summary-based context.** Generation proceeds section by section, but each section receives only a summary of prior text as context.
- **Base3: Progressive context injection without co-optimization.** Generation proceeds section by section with full prior text injected at each step, but without metric-feedback-driven generation–evaluation co-optimization.
- **MetricDraft: Full method.** Integrates adversarial research ideation clarification, staged structural anchoring, PRISM structured metric system, progressive context injection with STAGG, and metric-feedback-driven generation–evaluation co-optimization.

4.1.3. Evaluation Metrics

Quality evaluation employs PRISM's 5-dimension, 8-indicator system. Primary results focus on reporting TC (Terminology Consistency), RR (Content Redundancy Rate), FCR (Fabricated Citation Rate), M7 (Citation Density), and MQS (composite quality score).

4.1.4. Underlying Model and Parameter Configuration

All experiments uniformly use DeepSeek-V4-Pro (API model identifier `deepseek-v4-pro`, accessed via the official DeepSeek API) as the underlying LLM. Generation parameters are uniformly set to: temperature = 0.7, maximum output length (`max_tokens`) = 32,768. For MetricDraft's metric-feedback-driven generation-evaluation co-optimization stage, the maximum number of iteration rounds is set to 2. All four comparison methods (Base1, Base2, Base3, and MetricDraft) and all ablation configurations run independently under these same model and parameter settings to eliminate confounds from model selection and inference parameter differences.

4.2. Experiment 1: Preliminary Validation of the PRISM Structured Metric System

Experiment 1 examines whether PRISM can serve as an auxiliary evaluation framework for LIS paper draft quality. Fifteen paper drafts were selected as evaluation objects, and 10 experts with LIS backgrounds were invited to rate each along the D1–D5 dimensions, yielding 150 valid expert rating records.

Expert composition. All 10 experts are active researchers in library and information science or related directions, holding graduate degrees or above. Their professional ranks comprise 3 senior (full professor/senior research fellow) and 7 associate senior (associate professor/senior engineer). Their research directions span scientometrics, information retrieval, knowledge organization, digital libraries, and information quality assessment, matching the thematic scope of the 15 experimental tasks. All experts have experience in academic paper writing or journal reviewing.

Rating training and procedure. Written rating instructions were provided to each expert via email, including: definitions and scoring anchors for D1–D5 (1–5 Likert scale), rating examples (one high-score and one low-score sample), and rating guidelines. The instruction document was revised after a pilot rating by two experts to ensure dimensional definitions were clear and operable. In the formal rating phase, all 10 experts conducted **independent ratings** of the 15 drafts after reading the instructions and examples, with no communication between experts during the rating process. Each draft's final expert score is the mean of 10 experts' ratings on each D1–D5 dimension, with a separate composite quality total (sum of D1–D5 dimension means) computed for comparison with MQS. Pearson correlation coefficients and Spearman rank correlation coefficients are computed between expert mean scores and PRISM automated evaluation results.

Results across 15 drafts indicate that PRISM automated scores exhibit moderate-to-high positive correlations with expert ratings overall (Table 7). Specifically, D3 Clarity and D5 Literature Coverage achieve Pearson correlation coefficients of 0.848 and 0.765 respectively; the Pearson correlation between MQS composite score and expert overall quality rating is 0.871 ($p < 0.001$), with a Spearman rank correlation of 0.926 ($p < 0.001$). D2 Methodological Soundness (Pearson $r = 0.796$, $p < 0.001$) and D4 Significance (Pearson $r = 0.644$, $p = 0.010$) also reach statistical significance. D1 Originality achieves a Pearson correlation of 0.709 ($p = 0.003$), indicating significant linear association; however, its Spearman rank correlation of 0.493 ($p = 0.062$) does not reach the 0.05 significance level, suggesting that while PRISM's linear association holds for the originality dimension, ranking consistency requires further verification with larger samples. Originality assessment is inherently highly subjective, and this result is consistent with findings in peer review research where inter-rater reliability for this dimension is generally low [25].

Table 7. Consistency Validation Between PRISM Automated Evaluation and Expert Ratings Across 15 Drafts.

Dimension	PRISM Indicator	Pearson r	p	Spearman ρ	p
D1 Originality	M1 OS	0.709	0.0031	0.493	0.0618
D2 Soundness	M2 SS	0.796	0.0004	0.748	0.0013
D3 Clarity	M3 TC + M4 RR	0.848	0.0001	0.831	0.0001
D4	M5 SiS	0.644	0.0096	0.557	0.0310
Significance					
D5 Literature	M6 FCR + M7 CD + M8 CR	0.765	0.0009	0.886	0.0000
MQS composite vs. expert total		0.871	0.0000	0.926	0.0000

Expert rating consistency is at an acceptable level: $ICC(2,1) = 0.341$ based on overall quality scores, with $ICC(2,k) = 0.838$ based on 10-expert averaged scores; Cronbach's α across the five dimension scores is 0.737, indicating acceptable internal consistency. It should be noted that $ICC(2,1) = 0.341$ falls in the poor-to-fair range, reflecting substantial variability between individual raters on the same draft—a common finding in evaluation tasks involving higher-order semantic judgments such as originality and significance [25]. By adopting the 10-expert mean ($ICC(2,k) = 0.838$) as the final expert score, individual evaluation noise has been substantially attenuated. These results provide preliminary evidence that PRISM maintains relatively stable positive consistency with domain expert overall quality judgments, supporting its use as an auxiliary evaluation framework for paper draft quality diagnosis and iterative revision, though its ranking consistency on highly subjective dimensions such as originality still requires further verification with larger samples.

Expert ratings employ a 1–5 scale primarily to reduce the evaluation burden and enhance scoring anchor operability; PRISM's LLM judgments employ a 1–10 scale to preserve finer-grained differentiation in automated scoring. The difference in scale granularity may compress variability in expert scores; consequently, the correlation coefficients reported should be interpreted as conservative estimates of consistency.

4.3. Experiment 2: Validation of Long-Form Generation and Feedback Co-Optimization

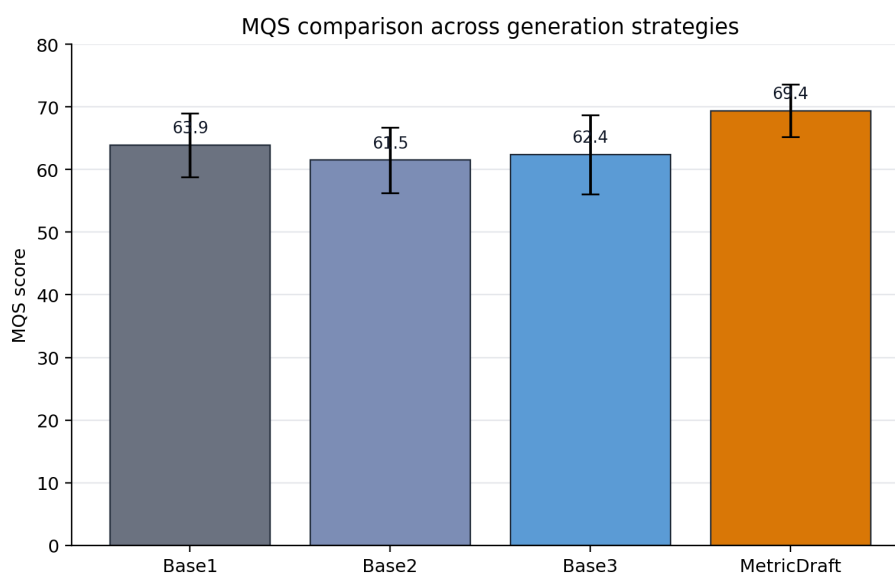
This experiment compares Base1, Base2, Base3, and the full MetricDraft method to simultaneously observe the effects of long-form generation strategies and metric feedback revision mechanisms on paper draft quality. Base1 tests one-shot full-draft generation, Base2 tests section-by-section generation with summary-based context passing, Base3 tests progressive context injection alone, and MetricDraft further adds PRISM metric-feedback-driven generation–evaluation co-optimization.

Table 8 and Figure 3 show that under PRISM's composite scoring combining content main score and citation risk penalty, the full MetricDraft method achieves the highest MQS (69.4 ± 4.2), exceeding Base1 (63.9 ± 5.1), Base2 (61.5 ± 5.2), and Base3 (62.4 ± 6.3). Comparing Base3 with MetricDraft further reveals that, on the same foundation of progressive context injection, introducing PRISM metric feedback raises MQS from 62.4 to 69.4, indicating that evaluation feedback yields relatively stable quality gains. Regarding individual indicators, MetricDraft's FCR (56.0%) is lower than Base1 (69.3%) and Base2/Base3 (both 100.0%), suggesting that its citation constraints and revision mechanism reduce the proportion of suspected fabricated citations to some extent; however, M7 citation density remains low, indicating that the current generation pipeline has not yet stably produced sufficient verifiable literature citations. MetricDraft's TC (21.7%) is lower than all three alternatives (causal analysis in the Discussion section), likely attributable to its longer generated length and the introduction of more synonymous expressions during revision. Overall, MetricDraft demonstrates a relatively clear relative score advantage under PRISM measurement at the draft stage.

Table 8. Comprehensive Comparison Across 15 Tasks: Different Generation Strategies vs. Full MetricDraft.

Method	TC(%) \uparrow	RR(%) \downarrow	FCR(%) \downarrow	M7 \uparrow	MQS \uparrow	Word Count
Base1: One-shot full draft	25.9 \pm 12.8	0.2 \pm 0.2	69.3 \pm 42.1	0.5 \pm 0.5	63.9 \pm 5.1	17,545 \pm 6,394
Base2: Section-wise + summary context	23.4 \pm 14.6	0.2 \pm 0.1	100.0 \pm 0.0	0.0 \pm 0.0	61.5 \pm 5.2	18,439 \pm 3,102
Base3: Progressive context injection	28.2 \pm 14.7	0.2 \pm 0.2	100.0 \pm 0.0	0.0 \pm 0.1	62.4 \pm 6.3	21,713 \pm 3,996
MetricDraft: Full method	21.7 \pm 8.8	0.1 \pm 0.1	56.0 \pm 49.4	0.5 \pm 0.6	69.4\pm4.2	27,078 \pm 7,068

TC: Terminology Consistency; RR: Content Redundancy Rate; FCR: Fabricated Citation Rate; M7: Citation Density; MQS: PRISM composite quality score.

**Figure 3.** MQS composite quality score comparison across generation strategies (15 tasks).

To further examine whether the observed differences have statistical support, we conducted paired two-sided tests on MQS between MetricDraft and each comparison method, with the 15 generation tasks serving as pairing units. Given that quality scores may not satisfy normality assumptions, we primarily report Wilcoxon signed-rank test results, supplemented with paired t -test p -values as sensitivity analysis. Results are presented in Table 9.

Table 9. Paired Significance Tests of MQS: MetricDraft vs. Comparison Methods (15 Tasks).

Paired Comparison	Mean Difference	Wilcoxon p	Paired t -test p
MetricDraft vs. Base1	+5.5	0.0084	0.0052
MetricDraft vs. Base2	+7.9	0.0006	0.0002
MetricDraft vs. Base3	+7.0	0.0063	0.0072

Paired tests across 15 tasks demonstrate that MetricDraft's MQS advantages over all three comparison methods reach statistical significance: the mean improvement over Base1 is +5.5 (two-tailed Wilcoxon $p = 0.0084$), over Base2 is +7.9 ($p = 0.0006$), and over Base3 is +7.0 ($p = 0.0063$). Paired t -test results also reach the 0.05 significance level in all cases. These results, under the current experimental setup, support the judgment that the combination of staged structural anchoring, progressive context injection, and PRISM feedback revision contributes to improved draft-stage composite quality scores.

4.4. Experiment 3: Ablation Study

Experiment 3 individually removes staged structural anchoring, progressive context injection, the PRISM metric evaluator, and metric-feedback-driven generation–evaluation co-optimization from the full method to observe each mechanism’s independent contribution.

Table 10 and Figure 4 show that removing each core mechanism is accompanied by significant MQS decline. Specifically, removing staged structural anchoring reduces MQS by 8.4 points (Wilcoxon $p = 0.0020$), suggesting that section-level structural management plays a relatively stable role in maintaining content coverage and overall quality. Removing progressive context injection reduces MQS by 7.7 points ($p = 0.0001$), indicating that complete prior text context may make an important contribution to long-text coherence and argumentation continuity. Removing the PRISM metric evaluator reduces MQS by 3.3 points ($p = 0.0125$), suggesting that a purpose-built paper draft quality evaluator may outperform generic evaluation signals. Removing metric-feedback-driven co-optimization reduces MQS by 7.6 points ($p = 0.0076$), further supporting the independent contribution of the feedback revision component. Ablation results also demonstrate that the full method can reduce FCR to 56.0%, while most ablation configurations approach or reach 100.0%, indicating that citation authenticity constraints and revision mechanisms retain practical utility in reducing suspected fabricated citations. Overall, all four core mechanisms demonstrate positive contributions to MetricDraft’s draft quality improvement, with staged structural anchoring, progressive context injection, and feedback revision contributing most prominently.

Table 10. Ablation Study of MetricDraft Key Mechanisms (15 Tasks).

Configuration	TC(%) \uparrow	FCR(%) \downarrow	MQS \uparrow	Δ MQS	Wilcoxon p
Full method	21.7 \pm 8.8	56.0 \pm 49.4	69.4 \pm 4.2	–	–
– Staged structural anchoring	20.3 \pm 10.7	100.0 \pm 0.0	61.0 \pm 7.8	–8.4	0.0020**
– Progressive context injection	20.0 \pm 11.6	98.9 \pm 4.3	61.7 \pm 3.7	–7.7	0.0001**
– PRISM evaluator (generic eval.)	28.9 \pm 18.0	100.0 \pm 0.0	66.1 \pm 2.0	–3.3	0.0125*
– Metric-feedback co-optimization	24.3 \pm 14.6	100.0 \pm 0.0	61.8 \pm 5.5	–7.6	0.0076**

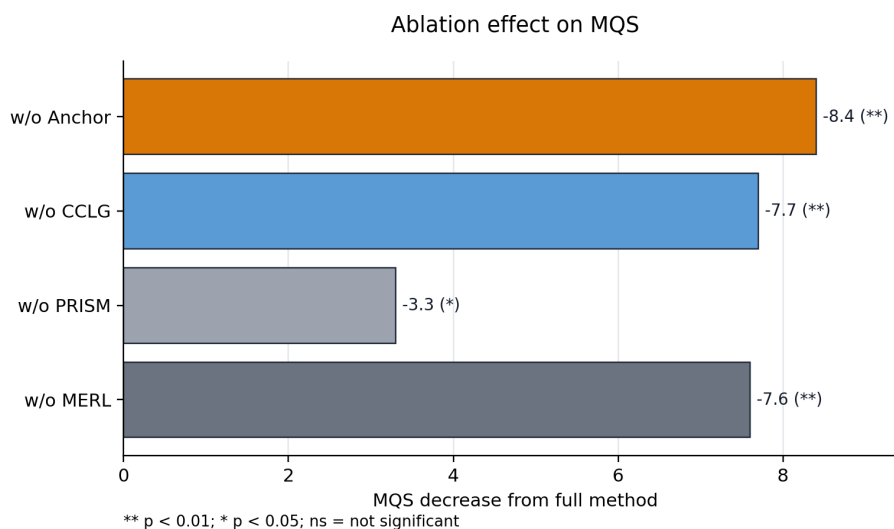


Figure 4. MQS decline upon ablation of key mechanisms (15 tasks).

5. Case Study

To concretely illustrate MetricDraft’s interaction process and quality control effects, this section uses Task T01 (Cross-disciplinary knowledge diffusion path identification via temporal citation network graph embedding) as an example, presenting an adversarial research ideation clarification dialogue excerpt, a PRISM first-round evaluation report summary, and a before/after paragraph comparison from MERL revision.

5.1. Adversarial Research Ideation Clarification Dialogue Excerpt

The following dialogue is excerpted from MetricDraft’s clarification stage for T01 (3 rounds total, condensed to highlight core questioning logic).

Round 1

User: I want to study knowledge diffusion in citation networks using graph embedding methods, targeting a JASIST paper.

MetricDraft (probing): Your idea has a basic direction but requires further clarification: (1) Are you focusing on static citation networks or temporally evolving networks? (2) Does “knowledge diffusion” specifically refer to cross-disciplinary citation propagation, topic migration, or methodological borrowing? (3) Is the intended technical contribution a novel embedding architecture design, or the application of existing graph embedding methods to diffusion measurement tasks? Please specify each point to delineate the research boundaries.

Round 2

User: Primarily cross-disciplinary knowledge diffusion, using citations as proxy, temporal networks are more valuable, contribution focus is on embedding framework design.

MetricDraft (probing): The direction is clearer, but two questions remain: (1) Existing paper-level embedding methods (e.g., SPECTER [56]) already support semantic similarity computation; what is the core difference of your temporal framework—is it the introduction of directed temporal edge propagation constraints, or the incorporation of cross-disciplinary citation weights in graph convolution? (2) What class of baselines will you compare against, and what metrics will measure diffusion path identification performance?

Round 3

User: The difference lies in introducing temporal constraints and cross-disciplinary citation weights; baselines include bibliographic coupling and co-citation analysis; evaluated by path identification accuracy.

MetricDraft (confirmation): The research objective is now sufficiently clear. Four-element research brief: **Research object**—cross-disciplinary knowledge diffusion paths in temporal citation networks; **Methodological core**—graph embedding framework incorporating directed temporal edge constraints and cross-disciplinary citation weight matrices; **Evaluation strategy**—comparison against bibliographic coupling and co-citation analysis baselines using path identification accuracy as the primary metric; **Venue norms**—JASIST, emphasizing methodological rigor and clear articulation of LIS disciplinary contributions. Upon confirmation, the system will proceed to title and outline confirmation stages.

This dialogue exemplifies the core mechanism of adversarial clarification: the system persistently probes “what differentiates your approach from existing methods” and “how will you validate it,” progressively transforming the user’s broad description into an actionable four-element research brief. In Round 2, the system directly identifies a key credibility gap in the methodological contribution, prompting the user to make implicit contributions explicit before entering the writing stage.

5.2. PRISM Evaluation Report and MERL Before/After Comparison

Following T01 draft generation and first-round revision, we conducted post-hoc diagnosis using the same PRISM computation as Experiment 2; results are shown in Table 11. It should be noted that during the online MERL revision stage, LLM judgment scores serve as diagnostic signals to

locate revisable issues, but MQS in experimental reporting is uniformly recomputed according to the content main score and citation risk penalty formula defined in §3.3; thus, the table below reports the same MQS consistent with Experiment 2 rather than the temporary total score returned by the LLM evaluator during revision.

Table 11. T01 Post-Revision Draft: PRISM Evaluation Report Summary.

Dimension	Indicator	Score	Core Diagnosis
D1 Originality	M1 OS (LLM)	8.0	Post-revision, the differentiation of TD-GAT and DABS from traditional main path analysis and temporal GNNs is clearly articulated
D2 Soundness	M2 SS (LLM)	7.0	Formal method description is relatively complete, but key experimental results and ablation evidence remain insufficient
D3 Clarity	M3 TC (automated)	14.5%	Numerous terminology variants in long text; normalized term proportion remains low
	M4 RR (automated)	0.1%	Sentence-BERT detection shows low cross-section semantic redundancy
D4 Significance	M5 SiS (LLM)	8.0	Domain value and application prospects of cross-disciplinary knowledge diffusion path identification clearly expressed
D5 Literature	M6 FCR (external)	100.0%	Under strict external database verification, extracted citations fail to form a verifiable citation set
	M7 CD (automated)	0.0	Verifiable citation density insufficient; does not reach recommended coverage level
	M8 CR (automated)	0.0	Due to insufficient verifiable citations, recent-5-year citation proportion cannot form effective support
MQS (PRISM)		69.3	Approaches but does not reach $MQS \geq 70$ quality gate

Diagnostic signals from the online revision stage indicate that T01's first-round draft weaknesses concentrated on methodological soundness and differentiated originality expression. After the first MERL revision round, LLM-judged dimensions improved: M1 rose from 7.0 to 8.0, M2 from 4.0 to 7.0, and M5 from 7.0 to 8.0. Taking the M1-feedback-driven originality revision as an example, the before/after comparison is as follows:

Before revision (first-round draft):

This paper proposes a graph embedding-based knowledge diffusion measurement method that encodes citation relationships as node vectors to capture cross-disciplinary knowledge migration. Compared with traditional bibliometric methods, the proposed method possesses higher representational capacity and can automatically discover latent diffusion paths.

After revision (MERL Round 1):

Unlike static semantic embedding methods such as SPECTER [56], this paper introduces two classes of temporal constraints in embedding learning: first, directed temporal citation edges serve as propagation directions, enabling node representations to retain temporal precedence semantics of knowledge flow; second, the cross-disciplinary citation frequency matrix serves as a message-passing prior in graph convolution, enabling embedding space distances to simultaneously reflect knowledge content similarity and diffusion directional intensity. This design can not only identify which publications are cited cross-disciplinarily but also quantify the intensity and temporal characteristics of knowledge diffusion from source to target disciplines, providing a methodological foundation for mapping disciplinary evolution.

The revised paragraph transforms the contribution from the vague assertion of “higher representational capacity” into two specific technical differentiators comparable with existing embedding methods, directly addressing the M1 diagnostic of “insufficient differentiation.” This case provides some evidence that MERL can improve local argumentation quality and LLM-judged dimensions. Upon PRISM recomputation, T01’s final MQS is 69.3, approaching but not reaching MetricDraft’s internal $\text{MQS} \geq 70$ quality gate. This result demonstrates that feedback revision can push drafts close to the quality gate boundary, but the current system still exhibits clear room for improvement in verifiable citation coverage and density.

6. Discussion

6.1. Methodological Significance of the Metric-Driven Generation Paradigm

The core methodological significance of this work lies in advancing academic paper draft generation from “prompt-driven text generation” toward “metric-driven quality control.” MetricDraft organizes objective clarification, structural confirmation, quality evaluation, and iterative revision as a continuous pipeline, with PRISM providing disciplinarily grounded fine-grained diagnosis, thereby reducing dependence on generic feedback and individual experience to some extent. PRISM’s value manifests in three aspects: disciplinary anchoring, dual-track operationalization, and extensibility—its dimensions derive from cross-comparison of authentic evaluation criteria, its indicators simultaneously cover objective measurement and semantic judgment, and it can be extended to other disciplines by replacing source venues.

This paradigm shift is significant because it situates LLM generation capability within a checkable, feedback-capable quality control framework. Traditional prompt engineering emphasizes how to write better prompts, whereas MetricDraft emphasizes how to transform the evaluation requirements of the academic community into constraint conditions within the generation process.

It should be acknowledged that PRISM simultaneously serves two roles in this paper: as the feedback source for MERL revision and as the unified evaluation yardstick when comparing different generation strategies in Experiment 2. This design does exhibit evaluation loop characteristics, but it is not intended to construct an internal standard favoring MetricDraft; rather, it serves the core research question: when the paper draft generation process is explicitly aligned with a structured quality requirement set derived from authentic journals, conferences, and peer review standards, does generation quality improve correspondingly? Thus, Experiment 2 tests the effect of “standards-driven generation” relative to non-standards-driven generation, rather than claiming that PRISM is equivalent to independent formal peer review. To mitigate the interpretive risks of circular evaluation, Experiment 1 introduces domain expert ratings to examine PRISM’s consistency with human quality judgments, and Experiment 3 separately examines the independent contribution of the PRISM evaluator and feedback revision through ablation. Accordingly, PRISM results should be understood as reproducible internal quantitative evidence of draft-stage quality control effectiveness, requiring joint interpretation with expert ratings, ablation experiments, and manual review.

Although our experiments employ a fixed $\text{MQS} \geq 70$ quality gate, the MetricDraft framework itself supports progressively raising the target gate value in subsequent writing rounds or practical application scenarios, enabling the system to transition from “achieving a basically usable draft” toward “approaching a higher-quality draft” in a progressive optimization manner.

6.2. Applicability Boundaries of Long-Form Generation and Feedback Optimization

PCI, by retaining complete prior text, facilitates consistency in terminology, symbols, research questions, and section logic, but its cost grows linearly with context length, and a “lost in the middle” effect may emerge in long texts [57]. Consequently, PCI may be most suitable for medium-length or well-structured paper drafts; for extremely long texts, combining key information summaries with section memory indexing could reduce context burden. The evaluation feedback mechanism may also exhibit diminishing returns: experiments show that MQS improvement plateaus after multiple

revision rounds, likely because the LLM tends toward local fine-tuning and factual issues such as fabricated citations are difficult to resolve through self-revision alone. Therefore, MERL should be regarded as an auxiliary tool for quality diagnosis and targeted revision rather than a fully automated quality assurance mechanism.

MetricDraft's experimental FCR remains at 56.0%, while Base2 and Base3 reach 100.0%, indicating that reference verifiability remains a critical bottleneck affecting composite quality interpretability. This issue primarily stems from the current system's lack of integration with real literature retrieval and citation verification modules—a structural limitation of LLM draft generation rather than a failure of the PRISM feedback revision mechanism itself. Future versions will incorporate RAG or retrieval-augmented citation generation to constrain references through externally verifiable sources rather than relying on model memory.

Furthermore, the effectiveness of feedback optimization depends on whether the evaluator can provide sufficiently specific and credible diagnostics. If feedback is too generic, the generator tends toward surface-level rewriting; if too local, it may neglect the paper's overall argumentation structure. Future work could introduce hierarchical feedback mechanisms: first judging at the full-text level whether research questions, structure, and contributions hold, then locating terminology, redundancy, and citation issues at the section level, and finally executing specific modifications at the paragraph level. Such layered feedback may help alleviate the tension between local optimization and global quality.

6.3. Analysis of MetricDraft's Lower Terminology Consistency

Experimental results show that MetricDraft's TC (Terminology Consistency, 21.7%) is lower than Base1 (25.9%), Base2 (23.4%), and Base3 (28.2%). However, this does not necessarily indicate weaker terminology management capability; it can likely be explained from three perspectives.

First, **statistical bias from text length**. TC measures the proportion of normalized consistent core terminology usage across the full text. MetricDraft drafts are longest on average (27,078 words), significantly exceeding Base1 (17,545), Base2 (18,439), and Base3 (21,713). Increased length raises the probability of terminology variants appearing, thereby depressing TC. The indicator may thus be subject to systematic text-length effects; future work could introduce length-normalized processing.

Second, **MERL revision enhances expressive diversity**. During the MERL stage, MetricDraft injects evaluation feedback into revision prompts, prompting the LLM to rewrite or expand content. While this process improves fluency and information completeness, it may also introduce more synonymous expressions, thereby lowering TC. The TC decline may partly reflect revision-induced linguistic diversification rather than terminology control failure.

Third, **TC's limited suitability for high-quality long texts**. For structurally more complete and content-rich drafts, terminology variants do not necessarily indicate quality degradation but may represent reasonable lexical variation. Future work should distinguish "normative synonymous substitution" from "confusing terminology variants" to improve indicator interpretability.

In summary, MetricDraft's lower TC is likely primarily attributable to the joint influence of text length, revision mechanisms, and indicator design, and should not be simplistically interpreted as a system deficiency. This finding also indicates that a single lexical consistency indicator has measurement limitations in long-text and iterative revision scenarios; future work could introduce length-normalized TC and further distinguish "reasonable synonymous substitution" from "confusing terminology variants" to enhance the indicator's explanatory power for high-quality long texts.

6.4. Limitations and Ethical Considerations

Several limitations remain. PRISM's current sources concentrate on the LIS domain; cross-disciplinary generalization requires re-collecting evaluation criteria and verifying dimensional stability. LLM-as-Judge indicators depend on the underlying model and scoring prompts, and higher-order semantic dimensions retain a degree of subjectivity [25]. Citation verification is constrained by external API coverage and may underperform on non-English literature, early publications, and

grey literature. The task set size remains limited compared to general NLP benchmarks; thus, our results should be understood as exploratory validation for LIS paper draft generation rather than definitive conclusions for all disciplines and writing scenarios. We did not directly compare with GPT-Researcher, AutoSurvey, or The AI Scientist in experiments, primarily because these systems respectively target investigative reports, survey generation, or fully automated research pipelines, with input/output interfaces and task objectives not fully consistent with MetricDraft's paper draft generation setting, making fair comparison within the current experimental framework infeasible. Future work will expand task scale, introduce additional disciplines, underlying models, real user writing scenarios, and cross-system comparison frameworks to examine external validity.

On the ethical plane, MetricDraft should be positioned as an **auxiliary tool for academic paper draft writing** rather than an automated writing system that replaces researchers in producing papers. The system retains human-in-the-loop design and prompts users to verify factual risks through fabricated citation identification. Future work will extend PRISM's disciplinary coverage, introduce real literature retrieval and citation verification modules, and conduct real-world human-AI collaborative usage validation.

It must be particularly emphasized that paper draft generation tools should not diminish researchers' principal position regarding research authenticity, originality, and scholarly responsibility. MetricDraft aims to help users more efficiently organize ideas, generate initial drafts, and identify quality issues, not to complete research judgments or circumvent scholarly review on users' behalf. System outputs should be viewed as editable drafts and diagnostic suggestions; the research design, data sources, citation authenticity, and conclusion reliability in final papers still require researcher confirmation.

7. Conclusion

Addressing the research question of "how to construct a metric-driven LLM-based academic paper draft generation method," this paper proposes MetricDraft. The method integrates adversarial research ideation clarification, staged structural anchoring, the PRISM quality assessment system, progressive context injection for long-form generation, and metric-feedback-driven generation-evaluation co-optimization, providing a pipeline-based quality control framework for academic paper draft generation. Under the current experimental setup, MetricDraft achieves the highest MQS composite quality score mean, with differences relative to all three baseline methods reaching statistical significance. PRISM exhibits moderate-to-high positive correlations with expert ratings, providing preliminary evidence that it can offer auxiliary support for paper draft quality diagnosis and feedback revision. These results still require further validation through larger-scale task sets, additional disciplinary scenarios, and real user experiments.

Author Contributions: Conceptualization, R.G.; methodology, Z.C.; investigation, L.F.; writing—original draft preparation, Z.C. and L.F.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund of China, grant number 21BTQ106.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the experts who participated in the draft quality evaluation. During the preparation of this manuscript, the authors used large language models for language polishing and format conversion. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
FCR	Fabricated citation rate
LLM	Large language model
LIS	Library and information science
MERL	Metric-feedback-driven generation–evaluation co-optimization loop
MQS	MetricDraft quality score
PCI	Progressive context injection
PRISM	Paper Review Indicators and Structured Metrics
RR	Redundancy rate
STAGG	Section-type-aware guided generation
TC	Terminology consistency

References

1. Achiam, J.; Adler, S.; Agarwal, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
2. DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv* **2024**, arXiv:2405.04434. <https://doi.org/10.48550/arXiv.2405.04434>.
3. Anil, R.; Dai, A.M.; Firat, O.; et al. PaLM 2 Technical Report. *arXiv* **2023**, arXiv:2305.10403.
4. Touvron, H.; Martin, L.; Stone, K.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
5. OpenAI. GPT-4o System Card. *arXiv* **2024**, arXiv:2410.21276.
6. Bai, Y.; Lv, X.; Zhang, J.; et al. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*; 2024; pp. 3119–3137. <https://doi.org/10.18653/v1/2024.acl-long.172>.
7. Lee, M.; Liang, P.; Yang, Q. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of CHI*; 2022.
8. Madaan, A.; Tandon, N.; Gupta, P.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In *Proceedings of NeurIPS*; 2023.
9. Checco, A.; Bracciale, L.; Loreti, P.; et al. AI-Assisted Peer Review. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 25. <https://doi.org/10.1057/s41599-020-00703-8>.
10. Grammarly Inc. Grammarly: Free Writing AI Assistance. Available online: <https://www.grammarly.com> (accessed on 6 May 2026).
11. Yuan, A.; Coenen, A.; Reif, E.; Ippolito, D. Wordcraft: Story Writing with Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*; 2022; pp. 841–852. <https://doi.org/10.1145/3490099.3511105>.
12. Gero, K.I.; Liu, V.; Chilton, L.B. Sparks: Inspiration for Science Writing Using Language Models. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants*; 2022; pp. 100–106. <https://doi.org/10.18653/v1/2022.in2writing-1.12>.
13. Kasneci, E.; Sessler, K.; Küchemann, S.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>.
14. Salvagno, M.; Taccone, F.S.; Gerli, A.G. Can Artificial Intelligence Help for Scientific Writing? *Crit. Care* **2023**, *27*, 75. <https://doi.org/10.1186/s13054-023-04380-2>.
15. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; et al. So What If ChatGPT Wrote It? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>.
16. Lund, B.D.; Wang, T. Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries? *Libr. Hi Tech News* **2023**, *40*, 26–29.

17. Perkins, M. Academic Integrity Considerations of AI Large Language Models in the Post-Pandemic Era: ChatGPT and Beyond. *J. Univ. Teach. Learn. Pract.* **2023**, *20*, 7.
18. Yao, L.; Peng, N.; Weischedel, R.; et al. Plan-and-Write: Towards Better Automatic Storytelling. In *Proceedings of AAAI*; 2019.
19. Yang, K.; Klein, D.; Peng, N.; et al. Re3: Generating Longer Stories with Recursive Reprompting and Revision. In *Proceedings of EMNLP*; 2022.
20. Yang, K.; Klein, D.; Peng, N.; Tian, Y. DOC: Improving Long Story Coherence with Detailed Outline Control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*; 2023; pp. 3378–3465. <https://doi.org/10.18653/v1/2023.acl-long.190>.
21. Zaheer, M.; Guruganesh, G.; Dubey, K.A.; et al. Big Bird: Transformers for Longer Sequences. In *Proceedings of NeurIPS*; 2020.
22. Guo, M.; Ainslie, J.; Uthus, D.; et al. LongT5: Efficient Text-to-Text Transformer for Long Sequences. In *Proceedings of NAACL*; 2022.
23. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical Neural Story Generation. In *Proceedings of ACL*; 2018.
24. Holtzman, A.; Buys, J.; Du, L.; et al. The Curious Case of Neural Text Degeneration. In *Proceedings of ICLR*; 2020.
25. Bornmann, L.; Mutz, R.; Daniel, H.D. A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS ONE* **2010**, *5*, e14331. <https://doi.org/10.1371/journal.pone.0014331>.
26. Smith, R. Peer Review: A Flawed Process at the Heart of Science and Journals. *J. R. Soc. Med.* **2006**, *99*, 178–182. <https://doi.org/10.1177/014107680609900414>.
27. Lee, C.J.; Sugimoto, C.R.; Zhang, G.; Cronin, B. Bias in Peer Review. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 2–17. <https://doi.org/10.1002/asi.22784>.
28. Tennant, J.P.; Dugan, J.M.; Graziotin, D.; et al. A Multi-Disciplinary Perspective on Emergent and Future Innovations in Peer Review. *F1000Research* **2017**, *6*, 1151.
29. Superchi, C.; González, J.A.; Solà, I.; et al. Tools Used to Assess the Quality of Peer Review Reports: A Methodological Systematic Review. *BMC Med. Res. Methodol.* **2019**, *19*, 48. <https://doi.org/10.1186/s12874-019-0728-x>.
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. <https://doi.org/10.1136/bmj.n71>.
31. Shinn, N.; Cassano, F.; Gopinath, A.; et al. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of NeurIPS*; 2023.
32. Bai, Y.; Kadavath, S.; Kundu, S.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073.
33. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of NeurIPS*; 2022.
34. Wang, X.; Wei, J.; Schuurmans, D.; et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of ICLR*; 2023.
35. Yao, S.; Zhao, J.; Yu, D.; et al. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR*; 2023.
36. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS*; 2020.
37. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of EACL*; 2021.
38. Liu, Y.; Iter, D.; Xu, Y.; et al. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. In *Proceedings of EMNLP*; 2023.
39. Novikova, J.; Dušek, O.; Curry, A.C.; Rieser, V. Why We Need New Evaluation Metrics for NLG. In *Proceedings of EMNLP*; 2017.
40. Sai, A.B.; Mohankumar, A.K.; Khapra, M.M. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.* **2022**, *55*, 1–39.
41. Kocmi, T.; Federmann, C. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In *Proceedings of WMT*; 2023.
42. Fabbri, A.R.; Kryściński, W.; McCann, B.; et al. SummEval: Re-Evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 391–409. https://doi.org/10.1162/tacl_a_00373.

43. Lu, C.; Lu, C.; Lange, R.T.; et al. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv* **2024**, arXiv:2408.06292.
44. Elovic, A. GPT Researcher. Available online: <https://github.com/assafelovic/gpt-researcher> (accessed on 8 May 2026).
45. Wang, Y.; Guo, Q.; Yao, W.; et al. AutoSurvey: Large Language Models Can Automatically Write Surveys. *arXiv* **2024**, arXiv:2406.10252. <https://doi.org/10.48550/arXiv.2406.10252>.
46. iPRES 2024 Organizing Committee. Reviewing. Available online: <https://ipres2024.pubpub.org/reviewing> (accessed on 6 May 2026).
47. Wiley. Peer Review Process. Available online: <https://authors.aws.wiley.com/author-resources/Journal-Authors/submission-peer-review/peer-review.html> (accessed on 6 May 2026).
48. Elsevier. *Information Processing & Management: Guide for Authors*. Available online: <https://www.sciencedirect.com/journal/information-processing-and-management/publish/guide-for-authors> (accessed on 6 May 2026).
49. Elsevier. *Journal of Informetrics: Guide for Authors*. Available online: <https://www.sciencedirect.com/journal/journal-of-informetrics/publish/guide-for-authors> (accessed on 6 May 2026).
50. MIT Press. *Data Intelligence: Submission Guidelines*. Available online: <https://direct.mit.edu/dint/pages/submission-guidelines> (accessed on 6 May 2026).
51. Zheng, L.; Chiang, W.L.; Sheng, Y.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of NeurIPS*; 2023.
52. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of EMNLP*; 2019.
53. Hendricks, G.; Tkaczyk, D.; Lin, J.; Feeney, P. Crossref: The Sustainable Source of Community-Owned Scholarly Metadata. *Quant. Sci. Stud.* **2020**, *1*, 414–427. https://doi.org/10.1162/qss_a_00022.
54. Kinney, R.; Anastasiades, C.; Authur, R.; et al. The Semantic Scholar Open Data Platform. *arXiv* **2023**, arXiv:2301.10140. <https://doi.org/10.48550/arXiv.2301.10140>.
55. Larivière, V.; Archambault, É.; Gingras, Y. Long-Term Variations in the Aging of Scientific Literature: From Exponential Growth to Steady-State Science (1900–2004). *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 288–296. <https://doi.org/10.1002/asi.20744>.
56. Cohan, A.; Feldman, S.; Beltagy, I.; et al. SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020; pp. 2270–2282. <https://doi.org/10.18653/v1/2020.acl-main.207>.
57. Liu, N.F.; Lin, K.; Hewitt, J.; et al. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.