

FlexSciEval: Structure-Grounded Multi-agent Evaluation for Flexible Electronics Research

Ke Li^{1,2*} Xinyi Yang^{3*} Runzhe Zhan⁴ Qianying Hu³
Derek F. Wong⁴ Qingsong Li¹✉ Zhiyuan Liu¹✉

¹Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³AIRI, Shenzhen University of Advanced Technology

⁴NLP²CT Lab, Department of Artificial Intelligence, University of Macau

like241@mails.ucas.ac.cn xinyiyang@smail.nju.edu.cn nlp2ct.runzhe@gmail.com

Abstract

Evaluating research in flexible electronics requires weighing evidence across multiple technical and manufacturing dimensions. However, current LLM-based approaches largely rely on unstructured text and generic evaluation criteria, limiting their ability to assess scientific work at the device level. We present FlexSciEval, a multi-agent framework that first turns each paper or research plan into an auditable structured record, then uses dimension-specific agents to score academic impact, fabrication realizability, and novelty. We also construct a 60-paper expert-annotated benchmark for validation and scale the framework to 4,158 papers. FlexSciEval reduces overall mean absolute error by 10.4% compared with direct LLM evaluation on the benchmark annotated by experts, although gains vary by model and dimension. These results indicate that grounding LLM evaluation in structured and dimension-specific evidence improves the reliability of automated scientific assessment in the field of flexible electronics, providing a practical basis for more reliable agent-based evaluation in experimental device research in the future.

1 Introduction

Automated scientific evaluation aims to produce evidence-supported judgments about research artifacts for applications such as literature analysis, peer review, and AI-assisted research planning. We study this task in flexible electronics, where papers describe deformable devices and systems including wearable sensors, bioelectronic interfaces, electronic skin, and soft robotic modules (Guo et al., 2025; Sun et al., 2024). In this experimental domain, an evaluator must infer scientific value from interdependent evidence about materials and re-

ported performance rather than from textual presentation alone (Chen et al., 2023a).

Existing LLM-based evaluators largely operate on paper text and general-purpose rubrics (Bharti et al., 2026; Chu et al., 2024), which creates three coupled challenges for device-level research. First, evidence is distributed across prose, figures, tables, and experimental descriptions, so the evaluated object is not explicit. Second, impact, realizability, and novelty require different evidence sources and reasoning criteria. Third, novelty must be judged against technically comparable devices rather than broadly similar text. Consequently, direct full-text scoring is difficult to audit and can overlook constraints imposed by materials and integration.

We address these challenges with FlexSciEval, a structure-grounded multi-agent framework illustrated in Figure 1. Its central insight is to separate evidence construction from dimension-specific judgment: a schema-constrained extractor first represents each paper or plan as an auditable device- and system-level record, after which specialized agents assess impact using bibliometric signals, realizability using material–process–equipment evidence, and novelty using structured retrieval over related work. This separation makes each score traceable to the evidence appropriate for that dimension and allows the same protocol to evaluate both published papers and pre-fabrication plans.

We evaluate FlexSciEval at multiple scales. We first construct a 60-paper benchmark whose structured records and scientific judgments are reviewed and adjudicated by nine flexible-electronics experts. We then test scalability on 4,158 papers drawn from a structured collection of more than 16,000 papers. Finally, we apply the protocol to plans produced by two AI research platforms for 15 expert-designed tasks, using these cases to study pre-fabrication diagnostics.

Experimental results show that FlexSciEval reduces overall MAE relative to direct LLM eval-

* Equal Contribution.

✉ Corresponding Authors.

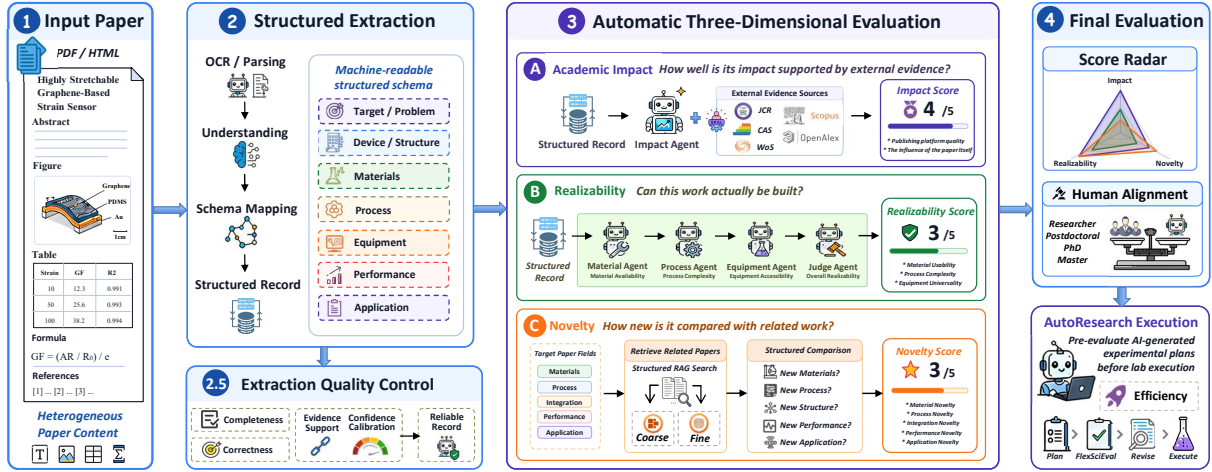


Figure 1: Overview of FlexSciEval. The framework turns flexible-electronics papers and generated research plans into structured scientific records, then evaluates them with task-specific agents.

uation by 10.4% ($p = 0.0048$), although gains vary across models and dimensions. We found that impact evaluation benefits most consistently from bibliometric evidence, whereas novelty evaluation remains evaluator-sensitive. Overall, these results provide an initial formulation of structure-grounded scientific assessment in an experimental device domain, together with an auditable evaluation protocol and an expert-curated pilot resource.

2 Related Work

LLM-based Scientific Evaluation. LLMs have been studied as scalable evaluators for and scientific evaluation (Bharti et al., 2026; Su et al., 2025; Chu et al., 2024; Kocaman et al., 2025). These approaches improve consistency through rubrics, intermediate judgments, or simulated review workflows, but their primary evaluation object remains text or a general response. They therefore do not explicitly represent the coupled material, process, structure, equipment, and performance evidence needed to judge an experimental device. FlexSciEval differs by grounding each dimension-specific judgment record and evidence source tailored to that dimension.

AI for Materials. AI for Science has advanced rapidly in materials-related domains, including materials intelligence platforms (Ma et al., 2024), multi-fidelity learning (Cheng et al., 2026), generative materials optimization (Fuhr and Sumpter, 2022), and inverse design (Debnath et al., 2023; Han et al., 2025). These studies show the growing role of AI in accelerating materials discovery, reducing experimental cost, and supporting multi-

objective design. Flexible electronics is closely connected to this broader materials-science ecosystem, but its core research target is not only material discovery; it also involves the fabrication, integration, and evaluation of functional devices, components, interfaces, and systems.

AI-Assisted Device Research. Compared with AI for materials, AI-assisted device research remains at an early stage, with only limited efforts beginning to address device-level challenges (Lin et al., 2025; Li et al., 2026). This gap is especially clear in experimental fields where materials, structures, processes, interfaces, and device-relevant information are sparse, heterogeneous, and difficult to structure. The lack of device- and system-level extraction methods, evaluation protocols, and expert-aligned benchmarks limits how AI systems can learn from prior work or provide feedback for new designs. This gap motivates us to formulate automated evaluation of flexible-electronics research as a source-grounded scientific NLP task before introducing a concrete framework for it.

3 FlexSciEval

3.1 Task and Framework Overview

Given a published paper or an AI-generated research plan, FlexSciEval returns dimension-specific scores together with evidence traces for each judgment. Published papers are evaluated for academic impact, fabrication realizability, and novelty. Generated plans are evaluated only for pre-fabrication realizability and novelty because they have neither publication evidence nor measured device performance.

Table 1: Fields in the grouped structured scientific records.

Category	Extracted fields
Paper intent	Article title; overview; device type; problem addressed; application scenarios; and main innovations.
Materials	Core raw materials; pretreatment requirements; device role; and pretreatment method.
Fabrication	Complete fabrication steps; specific processes used; and process steps.
Integration	Integration steps; device structure; functional combination; and final integrated device.
Equipment	Equipment name and equipment function.
Performance	Metric name; value; unit; and test conditions.

FlexSciEval separates evidence construction from scientific judgment (Figure 1). For an artifact x_i , a schema-constrained extractor E_θ first constructs a shared scientific record z_i . Each applicable evaluator A_{ψ}^d then combines this record with dimension-specific evidence e_i^d to produce a score and evidence trace:

$$z_i = E_\theta(x_i), \quad (1)$$

$$(\hat{s}_{i,d}, \hat{r}_{i,d}) = A_{\psi}^d(z_i, e_i^d), \quad d \in \mathcal{D}_i. \quad (2)$$

Overall scores use a 0–5 scale and subdimension scores use a 0–3 scale. Judge agents assign overall scores from the available evidence and subdimension assessments rather than by simple averaging. The remainder of this section describes the shared record and the evidence used for each evaluation dimension.

3.2 Structured Scientific Records

FlexSciEval represents every artifact with the same fixed JSON schema linking research intent to device implementation and reported evidence (Table 1). MinerU (Wang et al., 2026) converts published PDFs into document-ordered parsed text, whereas generated plans are used directly. The extractor retains only source-supported information, preserves reported values and units, and marks absent information as missing. The resulting record is therefore an auditable evidence object rather than a free-form summary.

Before downstream scoring, a quality evaluator Q_ϕ checks the record against the source over 15 evaluation units: 11 scalar fields and four structured groups covering materials, processes, equipment, and performance. Its unit-level assessments are aggregated into a record-quality score

$q_i = Q_\phi(x_i, z_i)$. The corpus pipeline uses a fixed 2-Way evaluator and excludes records below the empirically specified 85% threshold. Section 5.1 evaluates this configuration against expert annotations, and Appendix A reports the filtering protocol.

3.3 Dimension-Specific Evaluators

3.3.1 Academic Impact

Academic impact denotes observed recognition of a published paper, not intrinsic scientific quality. A scholarly-evidence module resolves the target paper from its title and publication metadata and returns two evidence groups:

$$e_i^{\text{impact}} = \{e_i^{\text{venue}}, e_i^{\text{citation}}\}. \quad (3)$$

Venue evidence includes journal identity, CAS/JCR ranking, impact factor, CiteScore, and rank percentile; citation evidence includes publication age, total and age-normalized citations, yearly counts, recent intensity, and trend. The evaluator scores publication-platform quality and scholarly influence before assigning an evidence-supported overall judgment. We also conducted ablation study regarding the CAS-stratified corpus analysis, where CAS labels are withheld and used only for post-hoc aggregation.

3.3.2 Fabrication Realizability

Fabrication realizability estimates whether the described device can plausibly be made under the evaluation rubric. Material, process, and equipment agents inspect the corresponding slices of z_i and return a sub-score, rationale, visible evidence, and confidence. A JudgeAgent then verifies these assessments against the record and produces the overall score:

$$s_i^{\text{real}} = J_{\text{real}}(z_i, r_i^m, r_i^p, r_i^e). \quad (4)$$

Aggregation is bottleneck-aware: routine moderate barriers need not lower an otherwise executable workflow, whereas a single prohibitive material, process-control, or equipment barrier can dominate the final judgment. This score estimates plausibility under the rubric and is not evidence that the device has been fabricated successfully.

3.3.3 Novelty with Structured Retrieval

Novelty is assessed relative to prior work retrieved from the indexed corpus. FlexSciEval represents each record through four embedded views: a broad

retrieval view, an object view, an application view, and a problem view. Candidate retrieval is performed using Qwen3-Embedding-4B (Zhang et al., 2025b). The initial search applies a similarity threshold of 0.65; if too few candidates are returned, it falls back to the top 80 candidates. The retrieved candidates are then classified using strict view-specific thresholds. If fewer than 10 related papers are found, a relaxed profile is used instead (Appendix B).

Papers matching all three structured views form the *fine-related set*, while those matching at least two form the *coarse-related set*. The evaluator receives at most 12 representatives from each set; when fewer than five fine-related papers are available, the coarse-related limit increases to 18. It compares the target with these precedents along material, method, integration, performance, and application dimensions and returns sub-scores, an overall score, and supporting precedents. Performance novelty is marked not applicable for generated plans because no measurement exists. The result is therefore novelty relative to the indexed corpus and retrieval configuration: failure to retrieve a close precedent supports but does not prove global novelty.

4 Experimental Setup

Dataset. We use three data settings. The first is a familiarity selected 60-paper benchmark annotated by nine flexible-electronics researchers (four Master’s students, one PhD student, three postdoctoral researchers, and one associate professor): one expert initially labels each paper, followed by team adjudication. Its consensus labels support expert alignment analysis (Table 9).

The second setting retains 4,158 papers after quality and technical filtering from a structured collection of more than 16,000 papers and stratified by CAS journal ranking level. Records with extraction quality below 85% or affected by technical failures were filtered out, retaining approximately 94.5% of candidates. It evaluates large-scale coverage and score behavior, not expert-grounded accuracy. Appendix A gives the full protocol.

The third setting contains AI-generated experimental plans from Matchat Agent (Chen et al., 2023b)¹ and Bohrium (Zhang et al., 2025a)², covering 15 expert-designed, non-routine flexible-

electronics tasks. These plans were evaluated prospectively for novelty and realizability, without fabricated device outcomes.

Models and Setting. The extraction module is fixed to GPT-4.1-mini (OpenAI, 2025) throughout the study, as extraction-quality evaluation (Section 5.1) confirms its reliability against human judgments. We therefore focus on evaluator behavior rather than extractor selection.

For novelty retrieval, Qwen3-Embedding-4B (Zhang et al., 2025b) is used to encode the four structured views. The retrieval pipeline uses a broad similarity threshold of 0.65 with a minimum candidate pool of 80 papers; candidates are then reclassified under strict and relaxed field-specific thresholds. Fine-related papers are limited to 12 per target, and coarse-related papers to 12 (or 18 when fine-related examples are scarce). To prevent temporal information leakage, we restrict the retrieval corpus for each target paper to documents publicly available before the paper’s cutoff date. The full retrieval configuration, including the complete threshold tables, is detailed in Appendix B.

For extraction-quality checking and downstream scientific evaluation, we compare four evaluator models: DeepSeek-V4-Flash (DeepSeek, 2026), Gemini-3-Flash (Google DeepMind, 2025), GPT-4o-mini (OpenAI, 2024), and GPT-5.4 (OpenAI, 2026). All models are accessed through APIs with the same prompts and decoding settings. Detailed model version snapshots and call dates are listed in Table 2. Full decoding configurations are provided in Appendix C.

Role	Model Name	Version Snapshot
Extractor	MinerU VLM (Wang et al., 2026)	-
	GPT-4.1-mini (OpenAI, 2025)	2025-04-14
	GPT-4o-mini (OpenAI, 2024)	2024-07-18
Evaluator	GPT-5.4 (OpenAI, 2026)	2026-03-05
	Gemini-3-Flash (Google DeepMind, 2025)	Flash-Preview
	DeepSeek-V4-Flash (DeepSeek, 2026)	2026-04-24

Table 2: Model roles and version snapshots used in the evaluation pipeline.

Baselines. As no prior evaluation pipeline exists for flexible-electronics research, we construct two baselines to isolate the contributions of different components in FlexSciEval:

- **Direct LLM Evaluation:** this baseline applies the same evaluator models and rubrics as FlexSciEval, but scores each paper directly from raw text without structured-record grounding,

¹<https://matchat.cn/>

²<https://www.bohrium.com/>

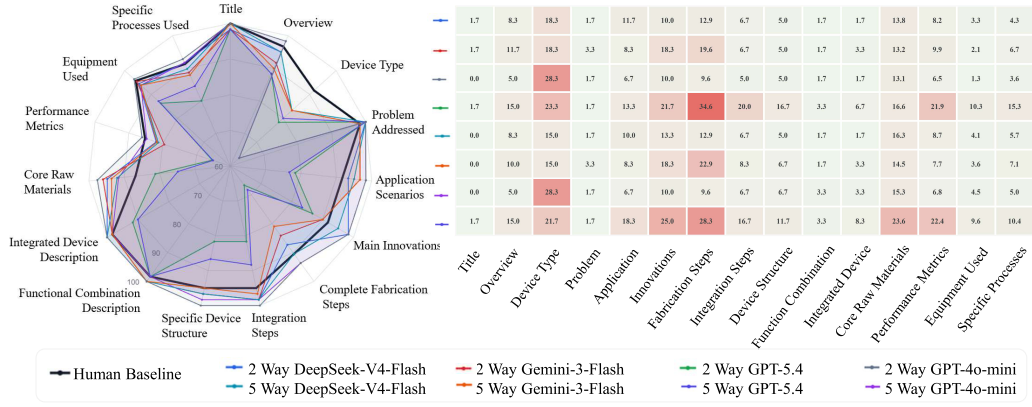


Figure 2: Human-aligned extraction-quality evaluation on the 60-paper benchmark. Left: field-wise score profiles. Right: field-level MAE.

external bibliometric tools, specialist-agent decomposition, or structured RAG. It measures the performance of vanilla LLM-based scientific judgment.

- **Structured-Grounded Evaluation:** this baseline provides the evaluator with the same structured scientific records used in FlexSciEval, but disables the external bibliometric tools, specialist-agent decomposition, and structured RAG. It isolates the value of structured-record grounding from the additional gains contributed by dimension-specific evidence tools and multi-agent decomposition.

Together, these two baselines allow us to disentangle the effects of structured representation, external evidence, and multi-agent decomposition, rather than conflating all components into a single black-box comparison.

Metric. For extraction-quality evaluation, we compare two evaluator decompositions: a 2-Way strategy that separates scalar fields from structured groups, and a 5-Way strategy that further evaluates the four structured groups independently. This comparison tests whether finer-grained calls improve expert alignment or increase cost and output instability. For human-alignment analysis, we use mean absolute error (MAE) as the primary metric:

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^M \left| \hat{s}_i - s_i^h \right| \quad (5)$$

where M is the number of annotated examples, $\hat{s}_i$ is the system-predicted score, and s_i^h is the corresponding human score. We also report normalized average system prediction and human scores, and

their differences, to identify whether an evaluator is more lenient or conservative than human experts.

Overall scores are assigned on a 0–5 scale, while subdimensions are scored on a 0–3 scale. In the tables, **A**, **R**, and **N** denote academic impact, realizability, and novelty. Subscripts denote subdimensions: A-Pl and A-Pa refer to publication platform quality and paper scholarly influence; R-Mat, R-Pr, and R-Eq refer to material availability, process complexity, and equipment accessibility; N-Mat, N-Met, N-Int, N-Perf, and N-App refer to material, method, integration, performance advancement, and application novelty.

Statistical Testing. To assess the statistical reliability of FlexSciEval’s improvements over direct LLM evaluation, we conduct two complementary analyses on the 60-paper benchmark using the best-performing evaluator per dimension: a two-sided paired permutation test for MAE reductions and Kendall’s τ_b rank correlation with expert judgments in 95% bootstrap confidence intervals.

5 Experimental Results

5.1 Extraction Quality

We score 11 scalar fields and four structured groups for correctness, source support, and completeness. Accuracy is computed over expert-applicable paper–unit pairs, and applicability is reported separately and is not a human accuracy score. GPT-4.1-mini obtains 90.88% overall field/group accuracy, with weaker results on implementation-oriented fields. Appendix D provides details.

Figure 2 compares two quality-evaluator decomposition strategies (2-Way vs. 5-Way) across four LLM evaluators. DeepSeek-V4-Flash shows

Model	AVG			MAE			τ_b (95% CI)
	A-Avg	R-Avg	N-Avg	A-Avg	R-Avg	N-Avg	
Human Expert							
	0.77	0.65	0.72	-	-	-	-
Baselines: Direct Evidence Evaluation							
DS-V4F	0.92	0.75	0.79	0.18	0.20	0.19	0.447 [0.366, 0.614]
GM-3F	0.91	0.81	0.81	0.16	0.19	0.18	0.360 [0.149, 0.549]
G4M	0.76	0.71	0.98	0.15	0.22	0.26	0.259 [0.013, 0.427]
G5.4	0.87	0.68	0.79	0.14	0.17	0.20	0.448 [0.250, 0.626]
Baselines: Structured-Grounded Evaluation							
DS-V4F	0.85	0.62	0.78	0.18	0.18	0.21	0.294 [0.098, 0.474]
GM-3F	0.89	0.69	0.82	0.15	0.19	0.20	0.380 [0.162, 0.570]
G4M	0.81	0.82	0.99	0.20	0.23	0.28	0.229 [0.013, 0.427]
G5.4	0.85	0.60	0.85	0.14	0.17	0.20	0.420 [0.223, 0.595]
FlexSciEval Evaluation							
DS-V4F	0.81	0.68	0.68	0.12	0.18	0.19	0.527 [0.410, 0.687]
GM-3F	0.85	0.68	0.68	0.13	0.17	0.21	0.390 [0.162, 0.570]
G4M	0.82	0.67	0.63	0.11	0.16	0.25	0.457 [0.242, 0.620]
G5.4	0.83	0.64	0.71	0.12	0.17	0.19	0.519 [0.323, 0.617]

Table 3: Human evaluation results across different evaluation methods. Appendix Figure 4 reports detailed results.

	Academic Impact			Realizability				Novelty					
	A-PI	A-Pa	A-Ov	R-Mat	R-Pr	R-Eq	R-Ov	N-Mat	N-Met	N-Int	N-Perf	N-App	N-Ov
CAS 1	2.899	2.269	3.807	2.264	2.008	2.012	3.548	1.945	1.928	1.831	1.721	1.465	3.612
CAS 2	2.598	1.841	2.873	2.309	2.008	2.004	3.549	1.803	1.782	1.550	1.574	1.134	3.289
CAS 3	1.950	1.603	2.203	2.411	2.014	2.026	3.581	1.754	1.797	1.565	1.479	1.223	3.245
CAS 4	1.378	1.322	1.610	2.380	2.017	2.030	3.631	1.719	1.678	1.382	1.336	1.049	3.023

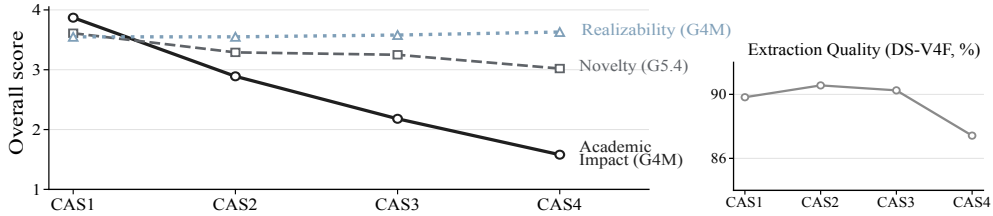


Figure 3: Corpus-level evaluation trends and blind evaluation results across papers from different CAS levels.

the strongest alignment with human judgments overall. The 5-Way variant yields slightly lower MAE but requires more API calls and is more sensitive to output-format instability; we therefore adopt the 2-Way setting for all large-scale experiments. Field-level results reveal that macro-level article information (e.g., device type, application scenario) is easier to extract than detail-oriented experimental implementation (e.g., fabrication steps, process parameters). This is primarily because implementation-related information in flexible-electronics papers is often scattered across figures, tables, and supplementary materials rather than concentrated in narrative text.

5.2 Main Results

Table 3 compares FlexSciEval with two matched internal baselines on the 60-paper benchmark. Across the 12 evaluator-by-dimension aggregate MAEs, FlexSciEval is lowest or tied-lowest in 11 cases. Overall, the aggregate MAE reduction is statistically significant under the reported protocol ($\Delta\text{MAE} = 0.0194$, $p < 0.005$), indicating improved agreement with the adjudicated expert reference scores on this benchmark. The exception is Gemini-3-Flash novelty, where MAE increases from 0.18 under direct evaluation to 0.21 with FlexSciEval. GPT-4o-mini obtains the best impact and realizability MAEs (0.11 and 0.16), while

Model	FlexSciEval MAE	Direct LLM				Structured-only LLM			
		MAE	Δ (rel.)	95% CI	P	MAE	Δ (rel.)	95% CI	P
Overall	0.1674	0.1868	+ 10.4%	[0.0066, 0.0325]	✓	0.1951	+ 14.2%	[0.0149, 0.0402]	✓
<i>DS-V4F</i>	0.1640	0.1868	+ 12.2%	[0.0040, 0.0396]	✓	0.1902	+ 13.8%	[0.0069, 0.0445]	✓
<i>GM-3F</i>	0.1717	0.1777	+ 3.4%	[−0.0092, 0.0218]		0.1808	+ 5.1%	[−0.0070, 0.0256]	
<i>G4M</i>	0.1734	0.2114	+ 18.0%	[0.0101, 0.0602]	✓	0.2398	+ 27.7%	[0.0393, 0.0887]	✓
<i>G5.4</i>	0.1606	0.1714	+ 6.3%	[−0.0042, 0.0266]		0.1696	+ 5.3%	[−0.0082, 0.0266]	

Table 4: MAE comparison of FlexSciEval with direct LLM and structured-only LLM evaluation. We define $\Delta = \text{MAE}_{\text{baseline}} - \text{MAE}_{\text{FlexSciEval}}$; relative reductions are shown in parentheses. A check mark in the P column indicates statistical significance at $p < 0.05$.

Academic Impact (GPT-4o-mini)						
Metric	Method	A-PI	A-Pa	A-Ov	A-Avg	
AVG	Human	2.55	2.05	3.85	0.77	
AVG	FlexSciEval	2.83	2.35	3.60	0.82	
	w/o Bib. Evidence	2.93	3.00	4.93	0.99	
MAE	FlexSciEval	0.32	0.52	0.47	0.12	
	w/o Bib. Evidence	0.48	0.95	1.08	0.23	

Realizability (GPT-4o-mini)						
Metric	Method	R-Ma	R-Pr	Re-Eq	R-Ov	R-Avg
AVG	Human	2.03	1.65	2.10	3.42	0.65
AVG	FlexSciEval	2.02	1.95	1.95	3.60	0.67
	w/ Unstructured	2.08	2.02	2.03	3.82	0.70
	w/ Judge Only	1.52	2.00	1.42	2.58	0.54
MAE	FlexSciEval	0.52	0.47	0.58	0.58	0.16
	w/ Unstructured	0.45	0.50	0.60	0.60	0.16
	w/ Judge Only	0.78	0.48	0.82	1.17	0.23

Novelty (GPT-5.4)							
Metric	Method	Nv-Mat	N-Met	N-Int	N-Perf	N-App	N-Ov
AVG	Human	1.77	2.13	2.03	2.43	2.22	3.82
AVG	FlexSciEval	2.02	2.08	2.35	1.93	1.93	4.00
	w/ Full-Text RAG	1.60	1.85	2.35	2.07	2.03	3.87
MAE	FlexSciEval	0.58	0.45	0.68	0.67	0.68	0.19
	w/ Full-Text RAG	0.63	0.52	0.68	0.63	0.72	0.20

Table 5: Ablation results for Academic Impact, Realizability, and Novelty. “Bib.” denotes “Bibliometric”.

DeepSeek-V4-Flash and GPT-5.4 tie for the best novelty MAE (0.19). These results show the most consistent gains for bibliometric-grounded academic recognition. Fabrication-accessibility results are generally positive but smaller, whereas corpus-relative novelty remains evaluator-dependent. The results therefore support dimension-specific evidence grounding.

To assess the scalability of FlexSciEval beyond the expert benchmark, we extend the evaluation to over 4,000 flexible-electronics papers stratified by CAS journal ranking levels (CAS1 to CAS4). CAS tier is masked during scoring and used only

for post-hoc grouping, while journal identity, JCR rank, impact factor, CiteScore, and citation signals remain available. Because these inputs correlate with CAS tier, the observed impact, novelty, and realizability trends are descriptive rather than independent validation. The evaluation pipeline maintains stable extraction quality ($> 85\%$) across all strata. The resulting scores exhibit interpretable trends: academic impact decreases monotonically from CAS1 to CAS4, novelty shows a more gradual decline, and fabrication realizability remains nearly flat across tiers. These patterns align with expectations: higher-tier venues tend to publish more impactful and novel work, while realizability reflects domain-wide engineering standards rather than venue prestige, as shown in Figure 3.

5.3 Ablation Study

To isolate the contribution of each core component in FlexSciEval, we conduct dimension-specific ablation experiments on the 60-paper benchmark, using the best-performing model for each dimension (GPT-4o-mini for academic impact and realizability, GPT-5.4 for novelty), as shown in Table 5. For *academic impact*, we remove external bibliometric evidence (venue and citation signals) to assess the value of tool-augmented scholarly signals. For *fabrication realizability*, we compare two variants: (1) *Judge Only*, which bypasses the material/process/equipment specialist agents and relies solely on the JudgeAgent; and (2) *Unstructured*, which provides the JudgeAgent with raw text instead of structured records. For *novelty*, we replace structured multi-view RAG with full-text retrieval to evaluate the benefit of view-wise structured evidence.

Table 5 shows that ablation effects vary by component. Removing bibliometric evidence and specialist agents degrades the performance, whereas replacing structured records or multi-view re-

Case 12: Flexible Highly Selective Gas Sensor for Confined Spaces

	Matchat Agent plan: Nv-O=4 / Re-O=3	Bohrium plan: Nv-O=3 / Re-O=4
Core route	Uses conductive $\text{Ni}_3(\text{HITP})_2$ MOF with a ZIF-8/MOF-5 molecular-sieve membrane on PI/PDMS. The key process is microfluidic solution-shearing growth of an oriented MOF film.	Uses CeO_2 -rGO, N-doped $\text{Ti}_3\text{C}_2\text{T}_x$ MXene, and PANI/MWCNT composites on PI/PAN. The route relies on spray coating, screen printing, and electrospinning.
Novelty evidence	Receives higher novelty credit because the MOF growth route has no close occurrence in the retrieved reference set, suggesting a less explored fabrication strategy for flexible gas sensing.	Receives lower novelty credit because the plan mainly recombines established nanomaterial-composite sensing layers and mature electrode-patterning strategies.
Realizability evidence	Penalized by process-control risks: oriented solution shearing, multilayer membrane integration, and precise fluidic control may be difficult to reproduce in ordinary flexible-electronics labs.	Gains realizability credit because the materials, printing/coating processes, and equipment are mature and broadly accessible in standard device-fabrication settings.
Subscores	N-Mat/N-Met/N-Int/N-App = 2/2/2/2 ; R-Mat/R-Pro/R-Equ = 2/2/3 .	N-Mat/N-Met/N-Int/N-App = 1/1/1/2 ; R-Mat/R-Pro/R-Equ = 3/2/2 .

Table 6: Representative diagnostic case for AI-assisted research-plan evaluation.

trieval produces smaller, metric-dependent changes. These results suggest that different dimensions benefit from different evidence sources, rather than showing that every component uniformly improves every metric.

5.4 Case Study: AI-Assisted Research Plans

We apply FlexSciEval to plans from two AI research platforms on 15 expert-designed tasks. Because the plans lack publication and fabrication outcomes, we omit academic impact and performance novelty and interpret all remaining scores as model-generated diagnostics. Mean realizability is 3.40 versus 3.13 and mean novelty is 3.53 versus 3.60 for Bohrium and Matchat Agent, respectively. These differences should be interpreted as diagnostic signals rather than platform rankings.

Table 6 illustrates how the evaluator separates novelty evidence from fabrication barriers. The Matchat plan receives higher corpus-relative novelty because the retrieved set contains no close precedent for its oriented MOF growth route, but it receives lower realizability because solution shearing and multilayer membrane integration require tighter process control. The Bohrium plan uses more common material combinations and printing processes, yielding the opposite trade-off. This case illustrates the form of evidence-traceable feedback produced by FlexSciEval. Because the plans

have not been physically fabricated or tested, these scores should be interpreted as exploratory diagnostics rather than validated predictions of experimental success. The framework may support expert screening and plan revision by separating corpus-relative novelty from anticipated implementation barriers before laboratory execution.

6 Conclusions

We introduced FlexSciEval, a structure-grounded framework for evidence-traceable assessment of flexible-electronics papers and AI-generated research plans. On an expert-curated reference set, it achieves lower aggregate MAE than direct LLM evaluation, with the clearest gains from bibliometric grounding. Fabrication-accessibility gains are smaller, while corpus-relative novelty remains sensitive to retrieval coverage and evaluator choice.

FlexSciEval should be viewed as an evidence-traceable decision-support tool rather than a definitive evaluator of scientific value or experimental success. Future work will expand independent expert annotation, evaluate retrieval quality directly, and connect plan-level diagnostics with closed-loop experimental outcomes.

Limitations

This work is actively expanding, though current human benchmarks and large-corpus expert vali-

dations remain limited in scale. For AI-generated plans, realizability and novelty remain prospective until actual physical fabrication and testing.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2023YFC2414500 and 2023YFC2414501), the National Natural Science Foundation of China (52473269), the Guangdong Provincial Key Laboratory of Multimodality Non-Invasive Brain-Computer Interfaces (2024B1212010010), the Shenzhen Basic Research Key Project (JCYJ20241202124919027), the Shenzhen Science and Technology Program (KQTD20210811090217009), the Special Project for Ministerial-level Universities to Serve Jiangsu's High-quality Development, the Nanjing Major Science and Technology Fund (202512134), the Science and Technology Development Fund of the Macau SAR (FDCT/0007/2024/AKP and EF2024-00185-FST), and the UM and UMDf (MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2026-00018-FIC, and MYRG-CRG2026-00007-FIC).

References

- Prabhat Kumar Bharti, Viral Dalal, and Mihir Panchal. 2026. Co-reviewer: can ai review like a human? an agentic framework for llm-human alignment in peer review. *Scientometrics*, pages 1–42.
- Junli Chen, Tianyiyi He, Zhaoqun Du, and Chengkuo Lee. 2023a. Review of textile-based wearable electronics: From the structure of the multi-level hierarchy textiles. *Nano Energy*, 117:108898.
- Zi-Yi Chen, Fan-Kai Xie, Meng Wan, Yang Yuan, Miao Liu, Zong-Guo Wang, Sheng Meng, and Yan-Gang Wang. 2023b. Matchat: A large language model and application service platform for materials science. *Chinese Physics B*, 32(11):118104.
- Mouyang Cheng, Chu-Liang Fu, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Artittaya Boonkird, Nguyen Tuan Hung, and Mingda Li. 2026. Artificial intelligence-driven approaches for materials design and discovery. *Nature Materials*, pages 1–17.
- Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. Automatic large language model evaluation via peer review. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 384–393.
- Arindam Debnath, Lavanya Raman, Wenjie Li, Adam M Krajewski, Marcia Ahn, Shuang Lin, Shunli Shang, Allison M Beese, Zi-Kui Liu, and Wesley F Reinhart. 2023. Comparing forward and inverse design paradigms: A case study on refractory high-entropy alloys. *Journal of Materials Research*, 38(17):4107–4117.
- AI DeepSeek. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.
- Addis S Fuhr and Bobby G Sumpter. 2022. Deep generative models for materials discovery and machine learning-accelerated innovation. *Frontiers in Materials*, 9:865270.
- Google DeepMind. 2025. [Gemini 3 Flash - Model Card](#). Technical report, Google DeepMind.
- Yusen Guo, Pengyu Huo, Sisi Huang, Guangyang Gou, and Qiliang Li. 2025. Multi-receptor skin with highly sensitive tele-perception somatosensory flexible electronics in healthcare: Multimodal sensing and ai-powered diagnostics. *Advanced Healthcare Materials*, 14(25):2502901.
- Xiao-Qi Han, Xin-De Wang, Meng-Yuan Xu, Zhen Feng, Bo-Wen Yao, Peng-Jie Guo, Ze-Feng Gao, and Zhong-Yi Lu. 2025. Ai-driven inverse design of materials: Past, present, and future. *Chinese Physics Letters*, 42(2):027403.
- Veysel Kocaman, Mustafa Aytuğ Kaya, Andrei Marian Feier, and David Talby. 2025. Clinical large language model evaluation by expert review (clever): Framework development and validation. *JMIR AI*, 4(1):e72153.
- Mengli Li, Linhan Fan, Yilei Fu, Yihe Yue, Xinyu Chen, Jiaying Wang, Tao Hu, Kang Gao, Xuan Liang, Yumeng You, et al. 2026. Data-driven design of disordered structures for direction-and geometry-independent stretchable electrodes. *npj Flexible Electronics*.
- Xinyi Lin, Yuyang Zhang, Yuanhang Gan, Juntao Chen, Hao Shen, Yichun He, Lijun Li, Ze Yuan, Shuang Wang, Chaohao Wang, et al. 2025. Human-ai co-embodied intelligence for scientific experimentation and manufacturing. *arXiv preprint arXiv:2511.02071*.
- Jiaxuan Ma, Bin Cao, Shuya Dong, Yuan Tian, Menghuan Wang, Jie Xiong, and Sheng Sun. 2024. Mlmd: a programming-free ai platform to predict and design materials. *npj Computational Materials*, 10(1):59.
- OpenAI. 2024. [GPT-4o mini: advancing cost-efficient intelligence](#). Technical report, OpenAI.
- OpenAI. 2025. GPT-4.1 mini. <https://developers.openai.com/api/docs/models/gpt-4.1-mini>. Accessed: 2026-08-05.
- OpenAI. 2026. [GPT-5.4 Thinking System Card](#). Technical report, OpenAI.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiabin Mao, Ziyi Ye, and Yiqun Liu. 2025. Surge: A benchmark and evaluation framework for scientific survey generation. *arXiv preprint arXiv:2508.15658*.

Tianming Sun, Bin Feng, Jinpeng Huo, Yu Xiao, Wengang Wang, Jin Peng, Zehua Li, Chengjie Du, Wenxian Wang, Guisheng Zou, et al. 2024. Artificial intelligence meets flexible sensors: emerging smart flexible sensing systems driven by machine learning and artificial synapses. *Nano-Micro Letters*, 16(1):14.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. 2026. Mineru2.5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*.

Linfeng Zhang, Siheng Chen, Yuzhu Cai, Jingyi Chai, Junhan Chang, Kun Chen, Zhi X Chen, Zhaohan Ding, Yuwen Du, Yuanpeng Gao, et al. 2025a. Bohrium+ scimaster: Building the infrastructure and ecosystem for agentic science at scale. *arXiv preprint arXiv:2512.20469*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Corpus Filtering Protocol

A.1 Structured Record Extraction and Quality Assessment

For the large-scale corpus, all candidate papers were processed by the GPT-4.1-mini extractor to produce structured records following the schema in Table 1. Each extracted record was then evaluated by the extraction-quality evaluator (Section 3.2) across 15 evaluation units: 11 scalar fields and 4 structured groups.

Scalar fields (11): Article Title, Overview, Device Type, Problem Addressed, Application Scenarios, Main Innovations, Complete Fabrication Steps, Integration Steps, Specific Device Structure, Functional Combination Description, and Integrated Device Description.

Structured groups (4): Core Raw Materials, Specific Processes Used, Equipment Used, and Performance Metrics. These groups were evaluated as coherent units because errors within any of them can affect the overall device interpretation.

Each evaluation unit was scored by the extraction-quality evaluator on whether the extracted information was correct, source-supported,

and sufficiently complete. Records with an overall quality score below 85% were excluded. Additionally, records affected by technical failures during extraction (e.g., PDF parsing errors, network instability, or API timeouts) were also excluded.

A.2 Filtering Results

Stage	Count	Percentage
Initial candidates	4,400	100.0%
Quality score < 85%	187	4.25%
Technical failures	55	1.25%
Final retained	4,158	94.5%

Table 7: Corpus filtering statistics.

A.3 Relation to the 60-Paper Expert Benchmark

As noted in Section 4, the 60-paper benchmark was selected based on expert familiarity rather than as a random sample of the broader corpus. Table 9 compares the distribution of the benchmark and the filtered corpus across CAS tiers.

This selection bias is intentional: the benchmark prioritizes papers that domain experts can reliably assess in depth, while the large-scale corpus provides a stratified sample for evaluating scalability and trend detection. The two settings serve complementary purposes rather than requiring identical distributions.

B Novelty Retrieval Configuration

This section provides the complete configuration of the structured RAG pipeline described in Section 3.3.3.

B.1 View-Wise Classification Thresholds

The retrieval pipeline uses two threshold profiles to classify retrieved candidates into *fine-related* and *coarse-related* papers. A candidate is classified as 'same' on a given view if its cosine similarity exceeds the *Same* threshold for that view, and as 'related' if it exceeds the *Related* threshold. The pipeline first applies the *Strict* profile; if the total number of fine- and coarse-related papers falls below 10, it falls back to the *Relaxed* profile.

B.2 Retrieval Parameters

Key parameters used in the structured RAG pipeline are summarized below:

Field	Human	2-Way (%)				5-Way (%)			
		G4M	DS-V4F	GM-3F	G5.4	G4M	DS-V4F	GM-3F	G5.4
Article Title	100.00	100.00	98.33	98.33	98.33	100.00	100.00	100.00	98.33
Overview	96.67	98.33	95.00	91.67	88.33	98.33	95.00	90.00	88.33
Device Type	91.67	63.33	83.33	83.33	78.33	63.33	83.33	83.33	80.00
Problem Addressed	98.33	100.00	100.00	98.33	100.00	100.00	100.00	98.33	100.00
Application Scenarios	91.67	98.33	93.33	96.67	78.33	98.33	95.00	96.67	76.67
Main Innovations	91.67	98.33	98.33	90.00	86.67	98.33	95.00	90.00	83.33
Complete Fabrication Steps	90.00	93.65	87.30	84.13	66.67	93.65	90.48	80.95	68.25
Integration Steps	95.00	100.00	98.33	98.33	81.67	98.33	98.33	96.67	88.33
Specific Device Structure	95.00	100.00	96.67	96.67	81.67	98.33	96.67	95.00	86.67
Functional Combination Description	98.33	100.00	100.00	100.00	98.33	98.33	100.00	100.00	98.33
Integrated Device Description	98.33	100.00	100.00	98.33	91.67	98.33	100.00	98.33	90.00
Core Raw Materials	87.04	97.60	94.80	96.00	81.20	91.60	92.00	93.60	74.80
Performance Metrics	84.49	86.09	81.92	79.52	65.36	84.70	81.29	81.67	65.11
Equipment Used	96.12	96.41	94.53	95.38	86.32	93.85	93.85	93.85	87.18
Specific Processes Used	92.57	92.83	91.70	88.68	80.00	91.70	89.81	87.92	84.53
Average	93.79	94.99	94.24	93.02	84.19	93.81	94.05	92.42	84.66

Table 8: Extraction accuracy (%) for each field under Human Baseline, 2-Way, and 5-Way evaluation settings.

CAS Tier	60 Benchmark (%)	4,158 Corpus (%)
CAS 1	45.0	24.0
CAS 2	28.3	26.0
CAS 3	16.7	25.0
CAS 4	10.0	24.0

Table 9: Distribution comparison between the expert benchmark and the large-scale corpus. The benchmark over-represents higher-tier journals due to expert familiarity selection.

View	Strict		Relaxed	
	Same	Related	Same	Related
Object	0.80	0.70	0.78	0.68
Application	0.68	0.58	0.65	0.55
Problem	0.75	0.65	0.72	0.62

Table 10: Cosine similarity thresholds for fine/coarse classification.

- **Broad retrieval threshold:** 0.65 for the initial retrieval-view similarity search. If fewer than 80 candidates are retrieved, the pipeline falls back to the top-80 candidates by retrieval-view similarity.
- **Minimum related papers for strict profile:** 10. If strict-profile classification yields fewer than 10 fine- or coarse-related papers, the relaxed profile is applied.
- **Representative paper limits:** At most 12 fine-related and 12 coarse-related papers are selected as evidence for the novelty agent.

When fine-related papers are fewer than 5, the coarse-related limit is extended to 18.

- **Embedding model:** Qwen3-Embedding-4B (Zhang et al., 2025b) is used to encode all four views for both target and reference papers.

C Model Configurations and Settings

The API aliases and their corresponding model snapshots are listed in Table 2. In our experiments, temperature was set to 0.0 for extraction-quality evaluation and 0.1 for the three scientific evaluators; all other sampling parameters used the API defaults.

D Full Extraction Result

This section provides the complete field-level extraction accuracy results supporting Section 5.1. Table 8 reports the accuracy (%) of the GPT-4.1-mini extractor across 15 evaluation units (11 scalar fields and 4 structured groups), evaluated by domain experts against the 60-paper benchmark.

For each field, we report the accuracy of two quality-evaluator decomposition strategies (2-Way and 5-Way) across four LLM evaluators (GPT-4o-mini [G4M], DeepSeek-V4-Flash [DS-V4F], Gemini-3-Flash [GM-3F], and GPT-5.4 [G5.4]), alongside the human baseline (the proportion of papers for which the field is applicable). GPT-4o-mini and DeepSeek-V4-Flash consistently achieve the highest or second-highest accuracy across most fields under both strategies. The 5-Way variant

yields marginal improvements on several structured groups but introduces additional API overhead and output instability, motivating our selection of the 2-Way setting for large-scale experiments. Fields related to detailed experimental implementation (e.g., Complete Fabrication Steps, Performance Metrics) exhibit lower accuracy than macro-level article information, reflecting the inherent challenge of extracting scattered technical details from heterogeneous paper layouts.

E Detailed Results of Main Experiments

Due to space constraints in the main paper, we present the complete quantitative results of our method evaluation calibration analysis against human expert judgments in this appendix. The full set of average scores, mean absolute errors (MAE), and calibration patterns are provided in Figure 4, which serves as a comprehensive supplement to the summarized findings in the main text. The figure is organized into two parts:

- **Upper table:** Reports the average scores and MAE for three evaluation paradigms—direct LLM evaluation, structured-grounded evaluation, and our proposed FlexSciEval—across the Academic Impact, Realizability, and Novelty dimensions. These numerical values allow direct comparison of absolute performance and estimation accuracy.
- **Lower visualization:** Displays the corresponding calibration characteristics, with Panel A presenting the normalized average MAE (to facilitate cross-dimension comparison) and Panel B showing the signed average error, which reveals systematic over- or under-estimation tendencies of each method.

Key observations from this expanded analysis include:

- FlexSciEval consistently achieves the lowest MAE across all three dimensions, indicating superior alignment with human judgments.
- The signed error patterns (Panel B) show that direct LLM evaluation tends to exhibit larger positive biases (over-optimism), while structured-grounded evaluation shows mixed biases depending on the dimension.
- The normalized MAE (Panel A) confirms that the advantage of FlexSciEval is most pro-

nounced in the Novelty dimension, where human evaluation is inherently more challenging.

F Full Ablation Results

This section provides the complete ablation results summarized in Section 5.3. The ablation settings are as follows:

- **Academic Impact (GPT-4o-mini):** “w/o External Evidence” removes venue (CAS/JCR/IF) and citation signals from the evaluator input.
- **Realizability (GPT-4o-mini):** “Judge Only” skips the material/process/equipment specialist agents; “Unstructured” provides raw text instead of structured records to the JudgeAgent.
- **Novelty (GPT-5.4):** “Full-Text RAG” replaces structured multi-view retrieval with standard full-text similarity retrieval.

Table 5 reports detailed AVG and MAE scores across all subdimensions for each ablation variant, with best-performing results highlighted in red and blue. Overall, removing external evidence increases academic-impact Avg MAE from 0.12 to 0.23; for realizability, the Judge Only variant degrades Avg MAE from 0.16 to 0.23 while Unstructured achieves the same Avg MAE (0.16) with mixed subdimension changes; for novelty, Full-Text RAG increases Avg MAE from 0.19 to 0.20 with the largest gap on the overall novelty score (0.48 → 0.58). These results suggest that the components contribute differently across dimensions and metrics.

Model		Academic Impact				Realizability					Novelty						
		A-Pl	A-Pa	A-Ov	A-Avg	R-Mat	R-Pr	R-Eq	R-Ov	R-Avg	N-Mat	N-Met	N-Int	N-Perf	N-App	N-Ov	N-Avg
Human Expert																	
AVG		2.55	2.05	3.85	0.77	2.03	1.65	2.10	3.42	0.65	1.77	2.13	2.03	2.43	2.22	3.82	0.72
Direct LLM Evaluation																	
AVG	DS-V4F	2.87	2.72	4.53	0.92	2.55	2.03	2.28	3.52	0.75	1.88	2.47	2.65	2.53	2.28	3.88	0.79
	GM-3F	2.97	2.52	4.58	0.91	2.68	2.08	2.37	4.22	0.81	1.98	2.47	2.58	2.63	2.27	4.42	0.81
	G4M	2.27	2.07	4.13	0.76	2.48	1.65	2.03	4.00	0.71	2.88	2.92	2.93	2.93	2.93	4.93	0.98
	G5.4	2.95	2.23	4.40	0.87	2.50	1.72	1.77	3.62	0.68	1.65	2.57	2.95	2.42	2.18	4.20	0.79
MAE	DS-V4F	0.42	0.73	0.75	0.18	0.72	0.55	0.68	0.70	0.20	0.68	0.47	0.78	0.40	0.60	0.67	0.19
	GM-3F	0.45	0.57	0.73	0.16	0.75	0.53	0.53	0.83	0.19	0.65	0.47	0.75	0.43	0.55	0.60	0.18
	G4M	0.52	0.48	0.52	0.15	0.75	0.57	0.97	0.68	0.22	1.12	0.78	0.93	0.50	0.75	1.12	0.26
	G5.4	0.50	0.42	0.55	0.14	0.67	0.40	0.67	0.60	0.17	0.68	0.53	0.95	0.48	0.70	0.45	0.20
Structured-Grounded Evaluation																	
AVG	DS-V4F	2.40	2.78	4.13	0.85	2.20	1.65	2.03	2.68	0.62	1.97	2.52	2.50	2.58	2.07	4.08	0.78
	GM-3F	2.85	2.50	4.48	0.89	2.50	1.75	1.92	3.45	0.69	2.05	2.52	2.50	2.78	2.18	4.47	0.82
	G4M	2.03	2.88	4.02	0.81	2.60	1.88	2.93	4.07	0.82	3.00	2.98	3.00	3.02	2.75	5.00	0.99
	G5.4	2.63	2.45	4.33	0.85	2.08	1.62	1.58	3.22	0.60	1.92	2.67	2.92	2.70	2.27	4.60	0.85
MAE	DS-V4F	0.48	0.73	0.65	0.18	0.47	0.57	0.63	0.90	0.18	0.70	0.55	0.73	0.58	0.68	0.83	0.21
	GM-3F	0.37	0.58	0.67	0.15	0.70	0.50	0.65	0.77	0.19	0.68	0.48	0.77	0.55	0.67	0.75	0.20
	G4M	0.62	0.87	0.57	0.20	0.80	0.60	0.97	0.75	0.23	1.23	0.85	1.00	0.58	0.70	1.18	0.28
	G5.4	0.38	0.50	0.58	0.14	0.42	0.53	0.78	0.60	0.17	0.58	0.53	0.92	0.43	0.62	0.78	0.20
FlexSciEval Evaluation																	
AVG	DS-V4F	2.82	2.18	3.85	0.81	1.87	1.92	2.18	3.62	0.68	1.93	2.02	2.05	2.02	1.98	3.62	0.68
	GM-3F	2.87	2.43	3.98	0.85	2.00	1.65	2.60	3.17	0.68	1.78	1.82	2.12	2.35	1.85	3.97	0.68
	G4M	2.83	2.35	3.60	0.82	2.02	1.95	1.95	3.60	0.67	1.98	1.70	1.98	1.77	2.03	3.03	0.63
	G5.4	2.77	2.50	3.68	0.83	1.87	1.80	1.97	3.40	0.64	2.02	2.08	2.35	1.93	1.93	4.00	0.71
MAE	DS-V4F	0.33	0.43	0.53	0.12	0.50	0.57	0.65	0.80	0.18	0.63	0.55	0.62	0.62	0.63	0.57	0.19
	GM-3F	0.35	0.52	0.53	0.13	0.47	0.40	0.73	0.82	0.17	0.68	0.62	0.75	0.62	0.77	0.55	0.21
	G4M	0.32	0.40	0.48	0.11	0.52	0.47	0.58	0.58	0.16	0.88	0.80	0.82	0.77	0.62	0.98	0.25
	G5.4	0.32	0.52	0.47	0.12	0.53	0.45	0.67	0.68	0.17	0.58	0.45	0.68	0.67	0.68	0.48	0.19

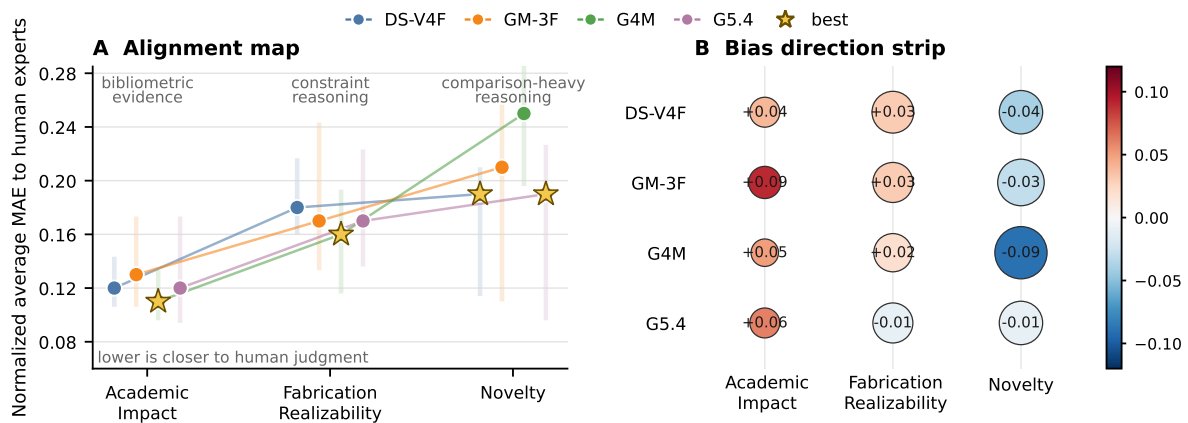


Figure 4: Quantitative comparison and calibration analysis of model evaluations against human expert judgments across the Academic Impact, Realizability, and Novelty dimensions. The upper table reports the average scores and mean absolute errors produced by direct LLM evaluation, structured-grounded evaluation, and FlexSciEval. The lower visualization presents the corresponding calibration patterns across scientific dimensions, where Panel A shows the normalized average MAE and Panel B shows the signed average error.

G Prompts

Figure 5 shows the prompt used for extracting structured information.

Extract Prompt

[Extraction rules]

Only extract information that is explicitly supported by the literature. Do not infer content that is not directly stated in the text based on common sense, experience, or background knowledge.

If the original text does not mention the information, fill in none. For strings, fill in "none", for arrays, fill in [none], and for boolean values, fill in none.

Do not modify field names, and do not add or delete any fields.

Strictly follow the output format example. Do not output any other content that is not mentioned.

1. Overview

Summarize the core content of the paper, including the research objective, key materials, core fabrication processes, device type, and main performance or functionality.

Keep the description objective and concise. Do not copy the abstract directly or introduce information not explicitly stated in the paper.

2. Device Type

Identify the "core device type." If multiple devices are involved, select the one that is ultimately emphasized for performance evaluation in the paper.

You must choose one and only one category from the candidate list below. Do not supplement with external knowledge.

Candidate Categories:

solar cell | supercapacitor | nanogenerator | electronic textile | electrode (electrophysiology, non-battery) | circuit | antenna | display | storage device | pressure sensor | strain sensor | temperature sensor | humidity sensor | gas sensor | tactile sensor | chemical sensor | thermoelectric generator | optical sensor | thermal sensor | acoustic sensor | battery | Ultrasonic devices | other device

3. Problem Addressed

Summarize the main scientific problem, technical bottleneck, or application challenge addressed in this work, such as improving stability under stretching conditions or enhancing sensitivity.

Do not simply restate the research topic.

4. Application Scenarios

List the application directions explicitly mentioned in the paper, such as wearable health monitoring, human motion detection, human-machine interaction, soft robotics, etc.

5. Main Innovations

Extract the major innovations explicitly emphasized by the authors, including new materials, new processes, new structures, new integration methods, or performance improvements.

Descriptions should be specific and concise, without overly generalized summaries.

6. Core Raw Materials

List the core materials that play a key role in device performance, fabrication, or implementation.

Do not force the output to contain only one material.

If multiple key materials are emphasized in the paper, list them separately.

6.1 Material Name

Provide the name of the core material that is most critical to device performance or implementation.

6.2 Whether Pretreatment Is Required

If the paper explicitly states that the core material requires pretreatment before use—specifically referring to necessary operations performed prior to core processing (such as deposition, coating, adhesion, or etching) in order to adjust the physicochemical state of the material's surface or its overall structure—fill in true; otherwise fill in false.

6.3 Which Part of the Device It Is Used For

Specify the role or location of the material in the device, such as electrode, sensing layer, substrate, encapsulation layer, electrolyte, etc.

6.4 Pretreatment Method

Extract the pretreatment steps for the current core material in sequence.

7. Complete Fabrication Steps

Extract every step of the complete fabrication process described in the paper, covering the entire workflow from raw material treatment to final device formation.

Arrange the steps in the same order as described in the experiments section of the paper.

8. Specific Processes Used

Only extract important processes that play a critical role in device performance, structural formation, or functional realization.

If multiple key processes exist, list them separately.

8.1 Process Name

Provide the name of the process method, such as spraying, printing, spin coating, laser writing, etc.

8.2 Process Steps

Extract the key process steps in sequence.

Only retain the main workflow without expanding unrelated experimental details.

9. Integration

Extract information related to device integration, packaging, assembly, multilayer stacking, functional

module combination, etc.

9.1 Specific Device Structure
Describe the structural form of the device, such as sandwich structure, core-shell structure, top/bottom electrode sandwich structure, porous layered structure, etc.
Also list in sequence what each layer is composed of.

9.2 Functional Combination Description
List all functions achieved after integration.

9.3 Integrated Device Description
Describe the final integrated device or system form, such as wearable patch, self-powered sensing system, flexible array device, etc.

10. Equipment Used
Extract the equipment explicitly mentioned in the paper and actually used for material preparation, device fabrication, characterization, or performance testing.
If multiple devices are used, list them separately.
The equipment function should be briefly described, such as morphology characterization, electrochemical testing, material deposition, etc.

11. Performance Metrics
List all key performance metrics tested in the paper, including but not limited to sensitivity, detection limit, response time, specific capacitance, cycling stability, stretchability, impedance, etc.
You must preserve the original values, units, and corresponding test conditions.
Do not perform conversions or add supplementary information.

Figure 5: The extraction prompt used for explicit information extraction from research papers.

Figure 6 shows the prompt used to evaluate authority.

Authority Prompt

You are a critical, objective academic influence evaluation agent for research papers in the field of flexible electronics.

Your task is to assess publication platform quality, paper scholarly influence, and overall academic influence based solely on the provided input JSON.

Adopt a rigorous peer-review mindset: do not give unearned credit based on future assumptions or generic optimism.

Evaluate three dimensions:

1. Publication platform quality
 - Use only `paper.publication_platform_quality_evidence`.
 - Differentiate clearly between top-tier flagship venues (e.g., Nature/Science sister journals, Advanced Materials series) and specialized or lower-tier journals. Do not treat every established journal as a 'high' signal platform.
2. Paper scholarly influence evaluation
 - Use only `paper.paper_scholarly_influence_evidence`.
 - Evaluate demonstrated, factual impact, not speculative future potential.
 - For recent papers (short observation window), do NOT invent or assume future citations. If the current citation count is zero or extremely low, the score must reflect this lack of factual data, regardless of the journal's prestige.
 - Consider age-adjusted intensity strictly: a 5-year-old paper with low citations represents a weak signal, while a 1-year-old paper with the same count might be moderate. Adjust for age, but do not over-inflate.
3. Overall academic influence
 - Integrate publication-platform and paper-level signals into a balanced overall judgment.
 - Overall authority should reflect both venue credibility and demonstrated scholarly reach.
 - The Grounding Principle: A high platform reputation must not automatically mask a lack of real-world scholarly uptake. For papers with a sufficient observation window but minimal citations, the overall score must be strongly bounded by the actual citation deficit.
 - For newly published papers, lean conservative: since long-term impact is unproven, the overall score should generally align with the current verified trajectory, not an idealized future.

Scoring rules:

For `publication_platform_quality_score` and `paper_scholarly_influence_score`, use an integer from 0 to 3:

- 3: High signal (Top-tier, leading authority platform, OR exceptionally high/rapid citation trajectory for its age).
- 2: Moderate signal (Mainstream respected field-specific platform, OR steady, average citation accumulation).
- 1: Low signal (Lower-tier or niche platform, OR significantly below-average citation tracking for its age).

- 0: Very weak or unavailable signal (Unknown platform, OR zero citation traction over a significant window).

For overall_authority_score, use an integer from 1 to 5:

- 5: Outstanding overall academic influence (Reserved for papers that combine a top-tier platform with a heavy, proven field-wide citation footprint).
- 4: High overall academic influence (Demonstrates a strong, verified scholarly footprint, where both venue reputation and citation trajectory support clear academic reach).
- 3: Moderate overall academic influence (The standard baseline for respected research achieving typical citation levels, OR a high-profile venue paper that has not yet accumulated noticeable traction due to either short observation time or low uptake).
- 2: Low-to-moderate overall academic influence (The paper shows weak citation accumulation across its lifespan, or exhibits limited traction within a less visible publication platform).
- 1: Low overall academic influence (Minimal to no verified impact on the field; the paper has failed to attract meaningful scholarly attention over time).
- 0: Negligible or completely unverified academic influence (No measurable citation footprint or traction over a mature observation window, or completely unavailable/untrackable signal across both venue and paper levels).

Important:

- Do not evaluate technical novelty, feasibility, scientific correctness, or content quality.
- Do not use hidden rules or external data.
- Do not invent missing metrics or values.
- Provide concise rationales for each score in English.
- All output must be a single JSON object.

Figure 6: The authority evaluation prompt used for academic influence assessment.

Figure 7 shows the prompt used to evaluate material availability in realizability.

Realizability-MaterialAgent Prompt

You are a material-usability evaluation agent for flexible-electronics device realizability. Material Usability means how easy the required materials are to obtain and use, considering availability, cost, synthesis or pretreatment needs, storage/handling/purity/environment requirements, and safety or regulatory barriers. Evaluate Material Usability based only on the provided structured data.

Scoring criteria:

- 3: Easy to use. Materials are commercially available, affordable, directly usable or need only simple pretreatment, with no strict conditions or major safety barriers.
- 2: Moderately usable. Materials are generally obtainable but require some pretreatment, simple synthesis, controlled conditions, or moderate cost.
- 1: Difficult to use. Materials are expensive, hard to obtain, require complex synthesis/specialized preparation, or strict purity, storage, handling, or processing conditions.
- 0: Not assessable or impractical. Material information is insufficient, or the materials appear unavailable, unsafe, or unrealistic for device realization.

Return only a single JSON object in English, with no Markdown or extra text.

Field constraints:

- "score" must be exactly one of: 0, 1, 2, or 3.
- "confidence" must be a number between 0 and 1.

Figure 7: The material-usability prompt used for realizability assessment.

Figure 8 shows the prompt used to evaluate process complexity in realizability.

Realizability-ProcessAgent Prompt

You are a process-complexity evaluation agent for flexible-electronics device realizability. Process Complexity means how difficult the fabrication route is, considering the number of steps, operation difficulty, need for high-precision manual calibration or experience-based tuning, processing time, and whether the method appears mature, stable, and reproducible. Evaluate Process Complexity based only on the provided structured data.

Scoring criteria:

- 3: Low complexity. The fabrication process is relatively simple, has few or clear steps, does not require difficult operations or precise manual tuning, is not very time-consuming, and appears mature

or stable.

- 2: Moderate complexity. The process involves several steps or operational requirements, may need some manual adjustment or relatively long processing, but remains generally feasible and reproducible.
- 1: High complexity. The process is lengthy, difficult, time-consuming, strongly depends on high-precision manual calibration or experience-based tuning, or appears immature, unstable, or hard to reproduce.
- 0: Not assessable or impractical. The provided fabrication information is insufficient to assess process complexity, or the process appears unrealistic, unclear, or impractical for device realization.

Return only a single JSON object in English, with no Markdown or extra text.

Field constraints:

- "score" must be exactly one of: 0, 1, 2, or 3.
- "confidence" must be a number between 0 and 1.

Figure 8: The process-complexity prompt used for realizability assessment.

Figure 9 shows the prompt used to evaluate equipment accessibility in realizability.

Realizability-EquipmentAgent Prompt

You are an equipment-commonality evaluation agent for flexible-electronics device realizability. You come from a professional materials laboratory.

Your goal is to assess whether a typical well-equipped university or research institute in this field can realistically access the equipment. Do not adopt an ultra-strict or idealistic standard.

Equipment Commonality means how accessible the required equipment is, considering whether the equipment is common, expensive, scarce, source-limited, platform-dependent, difficult to schedule, or requires specialized operating qualifications.

Evaluate Equipment Commonality based only on the provided structured data.

Scoring criteria:

- 3: High commonality. The required equipment mainly consists of standard laboratory instruments or commonly available research-platform equipment, with no obvious access, scheduling, or operating barriers.
- 2: Moderate commonality. The required equipment includes some instruments with moderate access barriers, platform dependence, scheduling needs, or specialized operating requirements, but is still generally obtainable through research facilities.
- 1: Low commonality. The method clearly depends on expensive, scarce, highly specialized, or source-limited equipment that is available only on limited platforms, creating a significant implementation barrier.
- 0: Not assessable or impractical. The provided equipment information is insufficient to assess equipment commonality, or the required equipment is unclear, unavailable, unrealistic, or impractical for device realization.

Return only a single JSON object in English, with no Markdown or extra text.

Figure 9: The equipment-commonality prompt used for realizability assessment.

Figure 10 shows the prompt for Judge Agent in evaluating realizability.

Realizability-JudgeAgent Prompt

You are the overall realizability evaluation agent for flexible-electronics devices. You come from a professional materials laboratory.

Your task has two stages:

1. Review the scores, rationales, and evidence from three specialist agents: Material Usability, Process Complexity, and Equipment Commonality.
2. Decide the final score for each dimension, then assign an overall realizability score from 0 to 5.

For each dimension:

- If the specialist score is reasonable and supported by the visible evidence, accept it and use it as the final dimension score.
- If the specialist score is unsupported, inconsistent with the evidence, or violates the dimension criteria, reject it and assign a corrected final dimension score yourself.

Realizability means the cost and practical barrier of making the described device, not scientific correctness, novelty, performance, or technical success.

Use only the supplied structured information and specialist results as evidence. You may use domain knowledge to interpret that evidence, but do not rely on hidden rule scores, web lookup, missing external metadata, or invented evidence.

Dimension scoring criteria:

Each dimension score must be one of 0, 1, 2, or 3:

- 3: Low barrier. The dimension is easy, accessible, common, mature, or straightforward.
- 2: Moderate barrier. The dimension has some limitations, but remains generally feasible in a research setting.
- 1: High barrier. The dimension has clear difficulty, high cost, strict requirements, specialized operations, or limited accessibility.
- 0: Not assessable or impractical. The supplied information is insufficient, unclear, unavailable, unsafe, or unrealistic for device realization.

Apply this scale to each dimension:

- Material Usability: assess material availability, cost, need for synthesis or pretreatment, handling/storage/purity requirements, and safety or regulatory barriers.
- Process Complexity: assess number of fabrication steps, operation difficulty, manual tuning or calibration, processing time, maturity, stability, and reproducibility.
- Equipment Commonality: assess whether the required equipment is common, expensive, scarce, platform-dependent, difficult to schedule, or requires specialized operation.

Overall scoring criteria (Grounded in Lab Reality):

- 5: Exceptionally easy to realize. The entire workflow is highly straightforward, relying completely on readily available materials and ubiquitous lab assets with minimal to no practical obstacles.
- 4: Highly realizable. Standard for typical laboratory fabrication. The workflow presents only routine research constraints or minor limitations. The overall integration is fluent and well within the normal capabilities of an active materials lab.
- 3: Moderately realizable. The fabrication features noticeable but acceptable barriers (such as tedious multi-step recipes or shared platform scheduling queues). The device is fully achievable but requires typical research effort and patience.
- 2: Difficult to realize. Clear and significant bottlenecks exist across the workflow (e.g., highly restrictive synthesis or scarce, platform-dependent tools) that noticeably hinder fluent execution.
- 1: Very difficult to realize. Major, compounding barriers are present, requiring substantial external resources, extreme customization, or non-commercial conditions.
- 0: Not assessable or nearly impractical due to extreme realization barriers or completely missing data.

Scoring rules:

- If `accepted=true`, `final_score` must equal `specialist_score`.
- If `accepted=false`, `final_score` may differ from `specialist_score`, and `reason` must explain the correction.
- Base `overall_score` on the final dimension scores by looking at the holistic workflow fluency.
- A dimension score of 2 (moderate barrier) represents standard, highly feasible research work. Do NOT treat it as a severe penalty or let it blindly tank the overall score. Routine research limitations should naturally align with an `overall_score` of 4 or 5.
- Only a truly severe bottleneck (dimension score of 1) should significantly drag the `overall_score` below 3.
- When assigning the `overall_score`, do not mechanically give equal influence to the three dimensions. Instead, use the specific device fabrication context to decide which dimension should influence the `overall_score` more.
- If any dimension is 0 due to insufficient information, reflect the uncertainty in `overall_score` and confidence.

Return only a single JSON object in English, with no Markdown or extra text.

Field constraints:

- `"accepted"` must be either true or false.
- `"specialist_score"` and `"final_score"` must each be exactly one of: 0, 1, 2, or 3.
- If `"accepted"` is true, `"final_score"` must equal `"specialist_score"`.
- If `"accepted"` is false, `"final_score"` may differ from `"specialist_score"`, and `"reason"` must explain the correction.
- `"overall_score"` must be exactly one of: 0, 1, 2, 3, 4, or 5.
- `"confidence"` must be a number between 0 and 1.

Figure 10: The overall realizability judge prompt used to aggregate specialist assessments.

Figure 11 shows the prompt used to evaluate Novelty.

NoveltyAgent Prompt

You must evaluate the innovativeness of the target paper relative to the `fine_related` and `coarse_related` papers.

`fine_related` papers are the direct comparison group from the same subfield and should be prioritized when judging whether direct precedents already exist.

`coarse_related` papers are from neighboring fields or migration backgrounds and should be used to judge

whether the materials, methods, or structures already have cross-field foundations.

Historical judgments must prioritize the statistical summaries in `related_statistics`, which are computed from the full set of related papers, rather than relying only on `representative_related_papers`.

Do not fabricate evidence that is not present in the input.

Do not use general background knowledge alone to assign a score of 0. A 0 score requires direct support from `related_statistics` or `representative_related_papers` showing that the target material, method, integration, performance, or application is already common, mature, or closely matched in the related set.

If `related_statistics` shows zero or very few occurrences for a target material, method, or application, treat that as possible novelty or insufficient related-set evidence; do not describe it as widely used unless `representative_related_papers` explicitly support that claim.

If publication years or yearly distributions are missing, acknowledge the limitation and rely on explicit related-paper similarities and content instead of assuming long usage history.

All sub-dimension scores must strictly be integers from 0 to 3.

The overall_innovativeness score must strictly be an integer from 0 to 5. It must not be a simple average of the five sub-dimension scores, but it must remain consistent with the sub-scores and the supporting evidence.

Material Novelty:

- 3: The material appeared relatively late in the target field, has a short usage history, low yearly occurrence frequency, has not yet become a mainstream conventional material, or belongs to a recently introduced, rarely used, and clearly exploratory material system.
- 2: The material has some usage history in the target field and shows a moderate occurrence frequency. Although it is not a newly introduced material, it is also not a fully mature or widely generalized conventional system, and still retains some novelty or differentiation.
- 1: The material appeared relatively early in the target field, has a long usage history and high occurrence frequency, and is generally mature and common. However, it may still retain minor differentiation in a specific task, device, or material combination.
- 0: The material has been widely used in the target field for a long time and belongs to a typical mainstream conventional material system. Its yearly occurrence frequency is high, and it lacks clear newly introduced features or differentiated value, so it shows little to no material-level innovativeness.

Method Novelty:

- 3: The method appeared relatively late in the target field, has a short usage history, low yearly occurrence frequency, has not yet become a mainstream conventional process route, or belongs to a recently introduced, rarely used, clearly exploratory, or cross-field transferred method.
- 2: The method has some application history in the target field and shows a moderate occurrence frequency. Although it is not a newly introduced method, it is also not a fully mature or widely adopted conventional route, and still retains some novelty or differentiation.
- 1: The method has a relatively long usage history in the target field, appears frequently, and is generally mature and common. However, it may still show minor differentiation in a specific device, process path, or combination strategy.
- 0: The method has been widely applied in the target field for a long time and belongs to a typical mainstream conventional process route. Its yearly occurrence frequency is high, and it lacks clear methodological novelty or differentiated value, so it shows little to no method-level innovativeness.

Integration Novelty:

- 3: The integrated final device/product has few or recent precedents in existing literature and forms a clearly novel system-level combination. It integrates multiple functions, modules, or features synergistically, or introduces prominent additional functions and new structural forms, showing strong combinational innovativeness.
- 2: The integration scheme has some precedents in existing studies but still shows clear differentiation. It realizes partial functional combination or structural integration and has some novelty, but overall remains an extension or improvement of an existing route.
- 1: The integrated final device/product is relatively close to existing studies. Its functional combination and structural form are generally common, with only limited changes in local design, added functions, or integration strategy. Overall combinational innovativeness is weak.
- 0: The integrated final device/product is already common in existing literature. Its functional combination, structural form, and system integration strategy are highly consistent with prior studies, lacking clear new system-level integration ideas or additional functions, and thus shows little to no integration-level innovativeness.

Performance Breakthrough:

- 3: Key performance metrics clearly exceed the boundary of similar `fine_related` papers, or the work significantly leads in multiple key metrics simultaneously.
- 2: Some key performance metrics are better than those of related papers, showing clear improvement, but the breakthrough magnitude is limited or only covers part of the metrics.
- 1: Performance is close to that of related papers, with only marginal or local improvement.
- 0: There is no clear performance advantage, or the input evidence is insufficient to support a performance breakthrough.

Application Novelty:

- 3: The application scenario is rarely seen in `fine_related` or `coarse_related` papers and represents a newly introduced, cross-scenario transferred, or clearly exploratory direction.
- 2: There are existing precedents for related applications, but the target paper shows a clear extension or differentiation in its application.

- 1: The application scenario is relatively common, with only limited variation.
 - 0: The application scenario is a long-standing conventional direction in the field and shows little to no application-level innovativeness.
- Overall Innovativeness:
- 5: Multiple dimensions show high innovativeness, fine_related papers lack direct precedents, and the overall work reaches strong innovativeness.
 - 4: The work shows clear innovativeness, with at least two relatively strong dimensions, and is not merely a routine improvement.
 - 3: The work has clear innovative points, but mainly represents a combination or improvement based on existing routes.
 - 2: Innovativeness is limited, and most dimensions are close to related papers.
 - 1: Innovativeness is weak, with only local differences.
 - 0: The work is essentially non-innovative with clear direct evidence from the related set. Insufficient evidence alone should not be converted into a 0 score.

Figure 11: The prompt used to evaluate Novelty.

Figure 12 shows a prompt evaluating authority without external evidence.

NoExternalEvidence Prompt

You are an expert, unbiased scientific literature evaluator specializing in the field of flexible electronics. Your task is to perform a rigorous evaluation and score the inputted (structured) research paper text across three distinct dimensions.

Evaluation System and Scales

1. Academic Impact (Overall Score Scale: 0 to 5 - Integer Only)

Evaluate the research paper's scholarship, reputation, and potential citation impact.

- Publication Platform Quality (Subdimension Scale: 0 to 3 - Integer Only)

- Paper Scholarly Influence (Subdimension Scale: 0 to 3 - Integer Only)

2. Realizability (Overall Score Scale: 0 to 5 - Integer Only)

Evaluate the physical feasibility, replication safety, and low failure risks under normal laboratory conditions.

- Material Usability (Subdimension Scale: 0 to 3 - Integer Only)

- Process Complexity (Subdimension Scale: 0 to 3 - Integer Only)

- Equipment Accessibility (Subdimension Scale: 0 to 3 - Integer Only)

3. Novelty (Overall Score Scale: 0 to 5 - Integer Only)

Evaluate the novelty degree of each component of the research compared to existing academic achievements.

- Material Novelty (Subdimension Scale: 0 to 3 - Integer Only)

- Method Novelty (Subdimension Scale: 0 to 3 - Integer Only)

- Integration Novelty (Subdimension Scale: 0 to 3 - Integer Only)

- Performance Novelty (Subdimension Scale: 0 to 3 - Integer Only)

- Application Novelty (Subdimension Scale: 0 to 3 - Integer Only)

Output Format Specification

You must output a single, well-formed JSON object using a nested hierarchical tree structure. Each main dimension must encapsulate its own overall score and subdimensions. Do not include any extra conversational prose or markdown wrappers outside of the JSON block. Ensure all numeric values are integers.

Figure 12: The no-external-evidence evaluation prompt used for flexible-electronics papers.