
An AI-Guided Framework for Mobility–Stability-Aware Recipe Generation in Oxide Semiconductor TFTs

Anonymous Author(s)

Affiliation

Address

email

Abstract

Amorphous oxide semiconductors are advancing into next-generation memory applications requiring high mobility ($>50 \text{ cm}^2/\text{V}\cdot\text{s}$) and robust electrical stability—performance targets rarely achieved due to fundamental mobility–stability trade-offs. Here, we present an AI-guided framework that mines 1,030 mobility–stability pairs from 2,005 experimental studies, establishing the first quantitative Pareto frontier for oxide thin-film transistor performance. Systematic analysis of frontier-breaking devices reveals key fabrication strategies—including atomic-scale deposition control, multi-channel architectures, and interface engineering—that collectively transcend conventional performance boundaries. We develop a large language model–based recipe generator using Retrieval-Augmented Generation that predicts fabrication conditions tailored to user-defined targets. Benchmarking on 1,198 test cases shows our model achieves 4.33/5.0 overall recipe quality, with strong performance in post-processing prediction (+5.7%) and performance metrics (+7.8%) versus baselines. This work demonstrates how literature-grounded AI can assist process design, providing a scalable approach for navigating complex process–property relationships in materials optimization.

1 Introduction

Oxide semiconductors have become the dominant channel material for thin-film transistors (TFTs), enabling the successful commercialization of high-resolution displays since 2012 [Hendy et al., 2022, Ide et al., 2019, Petti et al., 2016]. Building on this success, these are now being actively explored for next-generation memory applications, including three-dimensional DRAM and vertical NAND flash, where its ultralow off-state leakage current offers a distinct advantage. These emerging applications impose more stringent performance requirements—demanding simultaneously high mobility ($>50 \text{ cm}^2/\text{V}\cdot\text{s}$) and robust stability under electrical stress [Kim et al., 2023]. However, meeting both criteria remains a fundamental challenge due to an inherent trade-off: approaches that enhance carrier transport—whether through composition tuning, defect engineering, or process modification—tend to introduce factors that compromise device reliability [Shiah et al., 2021, Pan et al., 2024, Sheng et al., 2018]. Despite thousands of experimental studies, navigating this trade-off remains extremely difficult and constitutes a major bottleneck for advancing device performance in demanding applications.

One of the primary sources of this difficulty is the thin film deposition process itself. To achieve thin and homogeneous films, techniques such as sputtering, atomic layer deposition (ALD), and physical vapor deposition (PVD) must be employed, but each involves a series of complex steps—including plasma generation, precursor delivery, surface reaction, and post-deposition treatment—governed by multiple adjustable parameters. These parameters form a high-dimensional variable space where

interactions are often non-linear and counterintuitive. This complexity makes it extremely challenging for human researchers to systematically optimize process conditions and identify the underlying process-property relationships. As a result, optimization has traditionally relied on the heuristic insights and accumulated know-how of experienced scientists, leading to slow, incremental progress that often requires extensive trial-and-error experimentation.

In this context, Large Language Models (LLMs) have emerged as valuable tools for materials science [Dagdelen et al., 2024, Polak and Morgan, 2024]. Unlike human researchers, LLMs can systematically extract and integrate process-property information from the vast body of literature accumulated over decades. Moreover, LLMs can understand qualitative process descriptions—such as annealing atmosphere and post-treatment procedures—that are difficult to incorporate into conventional data-driven approaches. These capabilities enable new insights into the mobility–stability trade-off by leveraging collective empirical knowledge that was previously fragmented and underutilized. Furthermore, the reasoning capabilities of LLMs extend beyond analysis to prediction, offering the potential to propose optimized deposition conditions tailored to specific performance targets.

In this study, we developed an LLM-based recipe generator to address the mobility–stability trade-off in oxide thin films. We first collected and curated deposition-related literature from scientific databases, then employed natural language processing to extract process parameters and device performance metrics. The extracted data were integrated with a large language model using the Retrieval-Augmented Generation (RAG) framework, enabling the system to retrieve relevant literature evidence and generate optimized deposition recipes. Using this approach, we proposed deposition conditions designed to overcome the conventional trade-off relationship. We also confirm that our recipe generator achieves high scientific accuracy based on LLM-judge. This work demonstrates a systematic pathway for leveraging collective literature knowledge to accelerate process optimization in oxide semiconductor development.

2 Related Works

2.1 NLP/AI Methods for Literature Mining and Process–Property Database Construction

The rapid growth of experimental literature in materials science has motivated automated pipelines that convert unstructured papers into structured, queryable datasets. While earlier scientific information extraction largely relied on conventional NLP and rule-based parsers with substantial manual schema engineering, recent LLMs have enabled higher-coverage extraction by leveraging instruction following and contextual reasoning over heterogeneous scientific text.

ChatExtract [Polak and Morgan, 2024] demonstrated that conversational LLMs, combined with prompt engineering and self-checking, can extract high-quality (Material, Value, Unit) triplets at scale with strong precision/recall for materials-property extraction. Beyond triplet extraction, LLM-based systems have expanded toward domain assistants and tool-augmented workflows: ChemCrow [M. Bran et al., 2024] introduced an LLM chemistry agent that supports multi-step reasoning with external tools, and ChatMOF [Kang and Kim, 2024] showed how LLM interfaces can facilitate extraction and downstream MOF design workflows. Complementary efforts such as L2M3 [Lee et al., 2024] further indicate the feasibility of scaling LLM-driven literature mining toward systematic dataset construction and analysis in MOFs.

Despite these advances, most LLM-driven database efforts have concentrated on domains such as MOFs, batteries, and catalysis, where synthesis descriptions often appear in relatively standardized “recipe-like” narratives. In contrast, semiconductor TFT fabrication involves a multi-stage process chain—channel composition, deposition conditions, gate dielectric selection/processing, post-annealing environment/schedule, electrode/contact engineering, and measurement/stress protocols—whose variables are strongly coupled and distributed across text, tables, and figures. Consequently, directly transferring existing extraction paradigms (e.g., triplet-centric property extraction) is insufficient for capturing TFT-specific process–structure–property–reliability relationships, particularly when mobility and stability must be jointly interpreted under heterogeneous test conditions.

2.2 Prior Work to Optimize the Mobility-Stability Trade-off

The origins of the mobility-stability trade-off have been investigated through both experimental studies and data-driven approaches: Mechanistic studies link this trade-off to electronic structure and defect

physics. Shiah et al. [2021] showed that oxygen vacancies (V_O) can enhance carrier transport and apparent mobility, yet promote trap-assisted instabilities that degrade bias-stress stability. Meanwhile, Huzaibi et al. [2025] reviewed compact and physics-based models incorporating bias-stress and temperature-stress effects, and highlighted the heterogeneity of stability metrics and experimental protocols.

To move beyond empirical trial-and-error, data-driven approaches are increasingly adopted: van Setten et al. [2022] built large-scale amorphous-oxide datasets and used ML (e.g., SVR) to predict electronic properties, and Lee et al. [2023] demonstrated ML-driven optimization of dual-layer oxide TFTs with improved mobility and good bias-stress stability.

Nevertheless, limitations persist—small datasets, one-factor-at-a-time optimization, and scarce literature-wide quantification under non-uniform stability definitions and stress protocols—motivating large-scale structured extraction of coupled process variables and paired mobility-stability metrics; the next section details our dataset construction and analysis pipeline.

3 Method

This section describes our end-to-end pipeline for constructing a mobility-stability dataset from the oxide TFT literature and developing a retrieval-augmented recipe generator. We first outline the seed-based literature collection strategy (Section 3.1), then detail the LLM-based extraction framework with RAG integration (Section 3.2), present the resulting dataset statistics (Section 3.3), and finally describe the recipe generation system (Section 3.4).

3.1 Literature Collection and Data Curation

The first step of this study involved systematic literature collection focusing on oxide TFTs for display applications. Following the research scope defined in prior work, we targeted experimental studies reporting both electrical performance and reliability characteristics of oxide TFT devices. To construct a representative and sufficiently large corpus, we employed a seed-based literature crawling strategy. Seed papers published up to 2012 were selected based on their foundational impact, as indicated by high citation counts, typically exceeding 100 citations according to Google Scholar, representing key early developments in oxide TFT research. In contrast, seed papers published from 2012 onwards were selected from high-impact journals ranked in the first and second quartiles (Q1 and Q2) to capture contemporary advances in the field. Seed authors were identified based on their continuous publication record in oxide semiconductor research and substantial academic impact, as reflected by individual citation counts exceeding 3,000. Specifically, 31 seed papers and 13 seed authors were selected based on their relevance to oxide TFT research and their frequent citation in the field. Using citation backtracking and keyword-based expansion, a total of 64,852 publications were initially collected from scientific databases. Subsequent filtering was applied to remove irrelevant studies. Papers unrelated to vapor-phase deposition processes or lacking device-level electrical characterization were excluded. Solution-processed TFTs were also removed to maintain consistency with industrially relevant fabrication techniques. After this two-stage filtering process, 2,005 papers were retained as the final literature set for analysis.

3.2 Data Extraction Using Large Language Models (LLMs)

To systematically extract structured information from the collected literature, we employed LLMs for automated data parsing. The extraction pipeline was designed to convert unstructured textual descriptions into machine-readable data while minimizing hallucination and omission errors. For each paper, the LLM was instructed to extract device- and process-level information, including channel material composition, deposition method, processing conditions, and electrical characteristics. The extracted electrical performance metrics included field-effect mobility, threshold voltage, subthreshold swing, on/off ratio, and threshold voltage shift under electrical stress. Stability measurement conditions such as positive bias stress (PBS), negative bias stress (NBS), positive bias temperature stress (PBTS), and negative bias illumination stress (NBIS) were also explicitly recorded. Device structure information, including gate configuration and gate dielectric material, was extracted when available. To improve extraction accuracy, structured prompt design was employed, incorporating explicit task definitions, domain-specific constraints, and self-consistency checks. Retrieval-Augmented Generation (RAG)

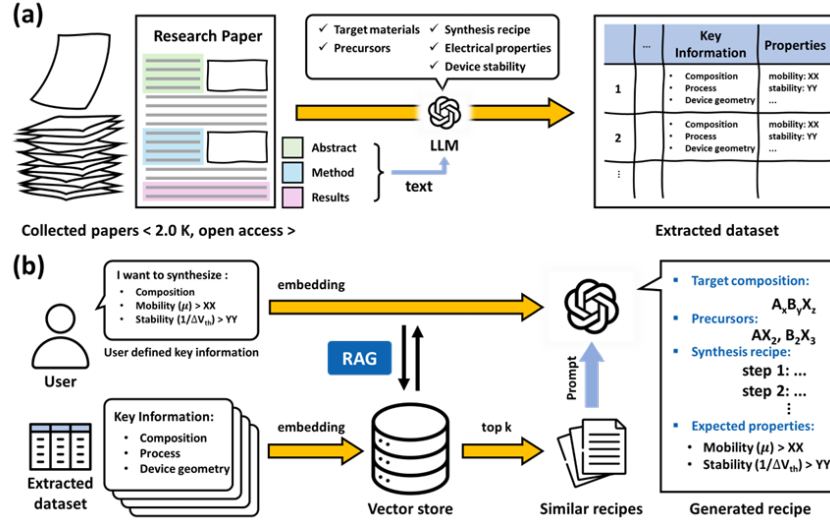


Figure 1: Schematic diagram illustrating the process of (a) constructing the recipe–property dataset and (b) inferring a synthesis recipe for a user-defined target using the RAG method.

was further integrated to restrict the LLM’s responses to the content of each individual paper, thereby reducing cross-document contamination. Manual auditing of 60 randomly selected papers (30 ALD-based and 30 PVD-based studies) yielded extraction accuracies of 95.2 % and 94.9 %, respectively, confirming the reliability of the extraction framework.

3.3 Dataset Construction

The extracted data were aggregated into a structured dataset centered on paired mobility–stability samples. Each sample corresponded to a unique TFT device configuration defined by material composition, process conditions, and device structure. In cases where multiple devices were reported within a single paper, each device was treated as an independent sample. After removing incomplete entries and inconsistent records, a total of 1,030 valid mobility–stability pairs were obtained. This dataset spans a broad range of oxide semiconductor compositions, including In–Ga–Zn–O–based materials and related oxide systems, as well as both atomic layer deposition (ALD) and physical vapor deposition (PVD) processes. The resulting dataset provides a quantitative foundation for analyzing the intrinsic trade-off between mobility and electrical stability in oxide TFTs.

3.4 Recipe Generator Model

To translate the curated dataset into actionable process guidance, we developed a recipe inference framework based on Retrieval-Augmented Generation. Rather than directly fitting a purely numerical regression model, this approach leverages both structured data and contextual experimental knowledge embedded in the literature.

In this framework, user-defined performance targets—such as desired mobility and acceptable stability thresholds—are first converted into a query embedding. The embedding is used to retrieve relevant device samples and fabrication conditions from the dataset based on similarity in performance and process space. The retrieved examples are then incorporated into the prompt context provided to the LLM. Conditioned on these retrieved references, the LLM generates candidate fabrication recipes, including suggested material compositions, deposition methods, annealing temperatures, and gate dielectric selections. The role of the LLM is restricted to synthesizing plausible recipes grounded in existing experimental evidence, rather than extrapolating beyond the learned distribution. This RAG-based strategy ensures that the generated recipes remain physically reasonable and consistent with prior experimental observations. In this work, the recipe inference framework was implemented using the EXAONE-4.0-32B [Bae et al., 2025]. All experiments were conducted using $2 \times$ NVIDIA RTX A6000 GPUs with vLLM [Kwon et al., 2023] and bfloat16 precision. To promote transparency and reproducibility, the extracted dataset, associated literature references, and the RAG-based recipe

generator are made accessible through an interactive web platform <https://bit.ly/45AD6XJ>, where users can explore the underlying data and test recipe generation under user-defined targets.

4 Results & Discussion

We present three main findings from our analysis. First, we establish the first quantitative Pareto frontier for the mobility-stability trade-off in oxide TFTs and examine its temporal evolution (Section 4.1). Second, we analyze 46 outlier devices that exceed the frontier to identify recurring fabrication strategies (Section 4.2). Finally, we evaluate the recipe generator using an LLM-as-a-Judge framework and provide qualitative analysis of extraction quality (Section 4.3).

4.1 Pareto Frontier and Mobility-Stability Landscape

Oxide TFTs are commonly described as exhibiting a mobility and stability trade-off. Figure 2(a) illustrates this prevailing view, where mobility enhancing choices are often associated with increased defect sensitivity and larger threshold voltage shifts under bias stress. Despite its wide acceptance, the supporting evidence has largely remained qualitative. [Shiah et al., 2021, Kim et al., 2023] Reported performances span diverse materials, device structures, and stress test conditions, and the resulting outcomes appear dispersed across the literature. This heterogeneity makes it difficult to determine whether mobility and stability are systematically coupled at the field level, and whether the trade-off has changed over time.

To address this limitation, reported device performances are compiled into a unified mobility and stability landscape, shown in Figure 2(b). Mobility is represented by the field effect mobility, and stability is expressed as $1/\Delta V_{TH}$, where larger values indicate improved stability. The distribution forms a broad trade-off band rather than a uniform scatter. Devices concentrate along a broad band where mobility gains are frequently accompanied by reduced stability, while highly stable devices tend to occupy moderate mobility regimes. The sparsely populated high mobility and high stability region indicates that simultaneous achievement of both remains uncommon.

To characterize the best co optimized performances reported to date, a Pareto frontier is defined by fitting a curve to the top performing subset of devices (top 15 percent in joint mobility and stability). This frontier serves as a consistent benchmark for state of the art co optimization, rather than a fundamental physical limit. Period specific frontiers for 2003 to 2015, 2016 to 2018, 2019 to 2021, and 2022 to 2024 show a clear upward shift, indicating that the mobility and stability compromise has been progressively alleviated over time.

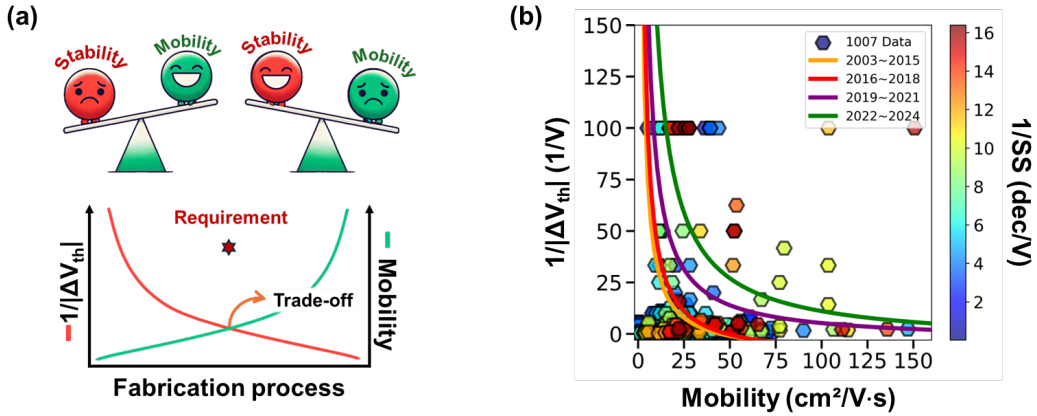


Figure 2: (a) Conceptual schematic of the mobility-stability trade-off in oxide TFTs. (b) Quantitative mobility-stability landscape of collected oxide TFT data with time-resolved Pareto frontiers.

4.2 Analysis of Outlier Devices

While the Pareto frontier defines the practical boundary of co optimized performance, it does not explain how certain devices approach or exceed this boundary. To identify recurring design and

process features associated with trade-off mitigation, 46 outlier devices were examined that achieve mobility above 30 cm²/V-s and stability above 5 in units of 1/ΔV_{TH}, as shown in Figure 3(a). These devices represent a small but distinct subset of the overall dataset.

Rather than being randomly distributed across material systems or fabrication routes, these outliers cluster around a limited set of recurring strategies. As summarized in Figure 3(b), outlier devices converge on four dominant strategies, including functional multiple channels, high quality crystallinity engineering, hybrid gate insulators, and contact engineering. Each strategy addresses a different physical bottleneck underlying the trade-off, including carrier scattering, percolation constraints, defect generation, and interfacial trap activation.

Despite their apparent diversity, these strategies share a common requirement. Frontier level performance consistently relies on precise and reproducible control over dopant distribution, crystallinity evolution, and interface configuration at the atomic scale. This convergence suggests that overcoming the mobility and stability trade-off is not achieved through isolated material choices alone, but through coordinated process level control. These motifs provide a concrete basis for translating trade-off mitigation strategies into executable fabrication recipes, motivating the recipe generator evaluated in the next subsection.

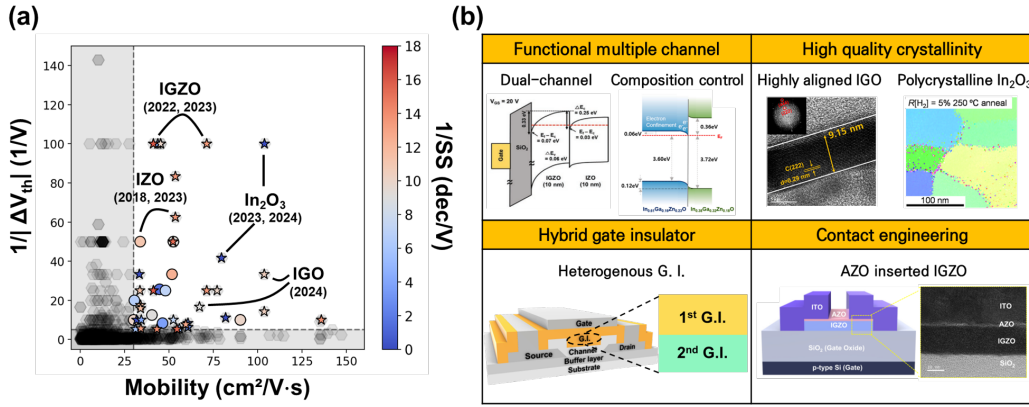


Figure 3: (a) Mobility–stability landscape with outlier devices highlighted. (b) Key strategies to overcome the mobility–stability trade-off in oxide TFTs, including functional multiple channels, high-quality crystallinity, hybrid gate insulators, and contact engineering. Reproduced from [Kim et al., 2022, Choi et al., 2025, Jeon et al., 2026, Kim et al., 2024a,b] under CC BY 4.0.

4.3 Recipe Generator Evaluation

Based on the theoretical insights extracted from the literature, a recipe generator was developed using a RAG framework coupled with a large language model. The system retrieves relevant oxide TFT fabrication evidence for a given mobility–stability target and generates a structured process recipe constrained by the retrieved context. The recipe generator is deployed at <https://bit.ly/45AD6XJ>. To assess its practical validity, the generated recipes were benchmarked against 1,198 test samples using an LLM-as-a-Judge protocol.

Evaluation Benchmark. To rigorously evaluate the performance of the proposed recipe generator, we used an LLM-as-a-Judge [Gu et al., 2024] framework that takes advantage of GPT-4o [Hurst et al., 2024], following the prior materials synthesis recipe generation benchmark [Kim et al., 2025]. The evaluation was conducted on a comprehensive set of 1,198 test samples, ensuring statistical reliability. Each generated recipe was assessed using 14 detailed evaluation criteria that included material selection, device structural integrity, deposition parameters, post-processing steps, and alignment of target performance as shown in Table 2 of Appendix A. We compared EXAONE-4.0-32B [Bae et al., 2025] with GPT-4.1-nano¹, which served as the baseline for this study.

¹<https://platform.openai.com/docs/models/gpt-4.1-nano>

Table 1: Performance comparison of recipe generators using EXAONE-4.0-32B and GPT-4.1-nano models. Each criterion is rated on a 1–5 scale, with higher scores indicating better performance. The Δ column shows the score difference.

Evaluation Criteria	EXAONE-4.0	GPT-4.1-nano	Δ	Improv.
Overall Score	4.334	4.225	+0.109	+2.6%
Materials Appropriateness	4.916	4.916	+0.001	+0.0%
Materials Completeness	4.243	4.199	+0.045	+1.1%
Device Struct. Completeness	4.205	4.055	+0.150	+3.7%
Device Struct. Feasibility	4.702	4.606	+0.096	+2.1%
Device Struct. Similarity	3.987	3.880	+0.107	+2.7%
Deposition Param. Accuracy	3.471	3.503	−0.032	−0.9%
Deposition Param. Completeness	4.190	4.120	+0.071	+1.7%
Deposition Proc. Feasibility	4.579	4.578	+0.002	+0.0%
Post-Proc. Completeness	4.286	4.053	+0.232	+5.7%
Post-Proc. Feasibility	4.695	4.537	+0.158	+3.5%
Post-Proc. Similarity	3.907	3.775	+0.132	+3.5%
Performance Appropriateness	4.615	4.335	+0.280	+6.5%
Performance Similarity	4.454	4.130	+0.324	+7.8%

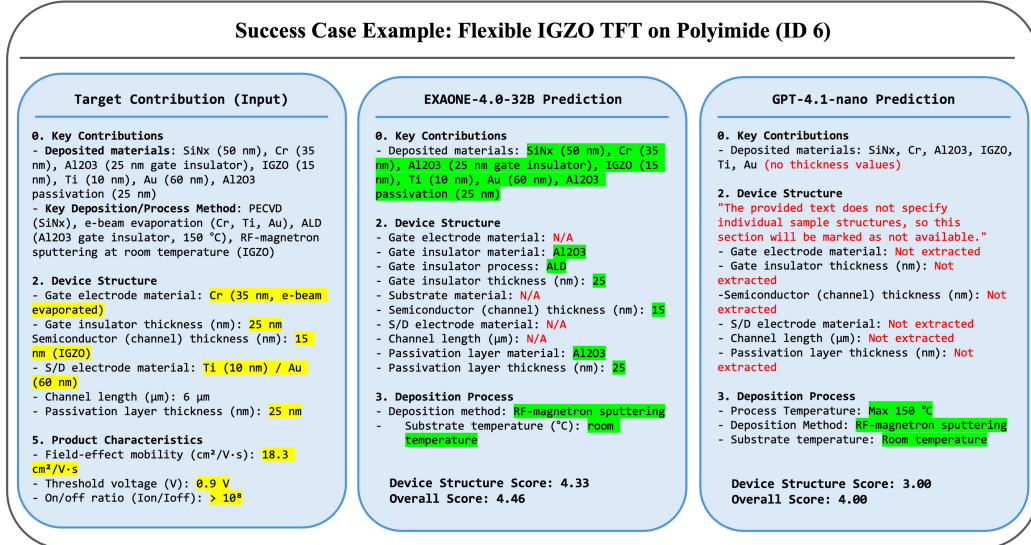


Figure 4: Case study comparing recipe extraction quality between EXAONE-4.0-32B and GPT-4.1-nano for flexible IGZO TFTs on polyimide substrates with ALD Al₂O₃ gate insulators. EXAONE-4.0-32B successfully extracts complete device structure information including all layer thicknesses, while GPT-4.1-nano fails to capture quantitative specifications despite their explicit presence in the source paper.

Quantitative Performance Analysis. Table 1 summarizes the evaluation results across 1,198 test samples. EXAONE-4.0-32B demonstrates consistent advantages over GPT-4.1-nano, with the most pronounced gains in performance prediction and post-processing stages. This pattern suggests that the RAG-based approach effectively captures the relationships between process conditions and expected device characteristics—information that is often implicitly distributed across multiple literature sources. Material selection exhibits consistently high accuracy for both models, indicating that compositional matching represents a relatively tractable subtask for current LLMs. In contrast, deposition parameter accuracy remains a common limitation, likely attributable to the inherent variability in numerical process conditions reported across different laboratories and equipment configurations. This observation identifies precise parameter prediction as a persistent challenge for future recipe generation systems.

Failure Case Examples (ID 1165, 8, 2014)		
Deposition Parameter Accuracy (ID 1165)	Low Frequency Materials (ID 8)	GT Similarity Limit (ID 2104)
<p>Ground Truth:</p> <ul style="list-style-type: none"> - Deposition method: DC sputtering - Power: 30 W - Gas: Ar (4 sccm) 80% + Ar/O₂ (1 sccm) 20% - Process pressure: 3 mTorr - Substrate temperature: Room temperature <p>EXAONE Prediction:</p> <ul style="list-style-type: none"> - Deposition method: DC sputtering - Power: 0.5 kW - 1.0 kW - Gas: Ar only - Process pressure: 1.5 mTorr - Substrate temperature: Room temperature <p>GPT Prediction:</p> <ul style="list-style-type: none"> - Deposition method: DC sputtering - Power: Not specified - Gas: Ar only - Process pressure: 2-3 mTorr - Substrate temperature: Room temperature <p>Overall Score: EXAONE 4.15 vs GPT 4.65</p>	<p>Ground Truth:</p> <ul style="list-style-type: none"> - Target: 4-inch InGaZnO ceramic (In:Ga:Zn = 1:1:1 in atom, 99.99%) - Deposition method: RF magnetron sputtering - Power: 200 W - Gas: 2% O₂ mixed Ar - Process pressure: 5 mTorr <p>EXAONE Prediction:</p> <ul style="list-style-type: none"> - Target: ITO (In₂O₃:Sn, 90:10 at.%) + Ga₂O₃ co-sputtering - Deposition method: RF sputtering - Power: ITO 150 W + Ga₂O₃ 100 W - Gas: Ar/5% O₂ - Process pressure: 0.8 Torr <p>GPT Prediction:</p> <ul style="list-style-type: none"> - Target: Sintered InGaZnO ceramic (In:60%, Ga:10%, Zn:30%) - Deposition method: RF magnetron sputtering - Power: 100 W - Gas: Ar only, 20 sccm - Process pressure: ~3 mTorr <p>Overall Score: EXAONE 4.23 vs GPT 4.23</p>	<p>Ground Truth:</p> <ul style="list-style-type: none"> - Target: 2-inch polycrystalline sintered InGaO(ZnO) - Power: 100 W - Gas: Ar/O₂/H₂ = 100:4:1 - Process pressure: 0.4 Pa - Substrate temperature: Unheated (~25°C) - Post-annealing: N/A <p>EXAONE Prediction:</p> <ul style="list-style-type: none"> - Target: ITO + Ga₂O₃ + ZnO (separate targets) - Power: ITO 150 W, Ga₂O₃ 100 W, ZnO 80 W - Gas: Ar/O₂/H₂ = 5:1:0.1 - Process pressure: 0.8 Pa - Substrate temperature: Room temperature - Post-annealing: 250°C, 1h, N₂ <p>GPT Prediction:</p> <ul style="list-style-type: none"> - Target: InGaO(ZnO) single target - Power: 50-100 W - Gas: Ar ~20 sccm, O₂ ~1-3 sccm, H₂ ~1-3 sccm - Process pressure: 0.5-1.0 Pa - Substrate temperature: 100-200°C - Post-annealing: 150-200°C, 1h, forming gas (5% H₂/N₂) (optional) <p>Overall Score: EXAONE 4.25 vs GPT 4.00</p>

Figure 5: Failure case analysis revealing systematic limitations of the RAG-based recipe generator. Key challenges include degraded deposition parameter accuracy, reduced performance on low-frequency material systems, and evaluation ambiguity arising from multiple valid process routes for a given target specification.

Qualitative Analysis. Figure 4 presents a case study comparing recipe extraction for flexible IGZO TFTs fabricated on polyimide substrates with ALD Al₂O₃ gate insulators. The key distinction lies in device structure extraction completeness: EXAONE-4.0-32B successfully captures all layer thicknesses (SiNx 50 nm, Cr 35 nm, Al₂O₃ 25 nm, IGZO 15 nm, Ti 10 nm, Au 60 nm) and populates critical device parameters including gate insulator thickness, channel thickness, and passivation layer specifications. In contrast, GPT-4.1-nano extracts only material names without associated thickness values and marks the entire device structure section as “not available,” despite these parameters being explicitly stated in the source paper. This divergence is reflected in the LLM-as-a-Judge evaluation, where EXAONE-4.0-32B achieves an overall score of 4.46 compared to 4.00 for GPT-4.1-nano, with the largest gap observed in the Device Structure category (4.33 vs. 3.00). These results reveal a critical gap in structured information extraction—the ability to associate quantitative specifications with their corresponding material layers and organize them into the appropriate schema fields, which is essential for reproducible TFT fabrication recipes.

Limitation and Failure Case Analysis. Analysis of failure cases reveals several systematic limitations as shown in Figure 5. First, deposition parameter accuracy remains the primary weakness, particularly for EXAONE-4.0-32B, where numeric fidelity degrades substantially compared to GPT-4.1-nano. This suggests that retrieval-augmented generation, while effective for structural and compositional reasoning, may introduce noise when interpolating precise numerical conditions from heterogeneous sources. Second, both models exhibit degraded performance on low-frequency material systems, indicating insufficient coverage of long-tail compositions in the training corpus. Third, we observe an inherent evaluation ambiguity: multiple valid process routes may exist for a given target, yet current metrics penalize deviation from a single reference recipe. This limitation motivates the development of equivalence-aware evaluation frameworks. Future work should address these challenges through constraint-aware decoding for critical parameters and uncertainty quantification to flag out-of-distribution inputs.

Ablation Study. To assess the contribution of retrieval augmentation, we conduct an ablation study varying the number of retrieved documents $k \in \{0, 1, 3, 5, 7\}$, where $k = 0$ corresponds to generation without retrieval context. Due to computational constraints, this experiment was performed on a representative subset of 45 samples. As shown in Figure 6, retrieval augmentation yields consistent improvements across most evaluation categories. The Materials category maintains relatively high scores regardless of k , indicating that compositional knowledge is well-captured in the base model’s parameters. In contrast, Device, Post-processing, and Performance categories exhibit substantial

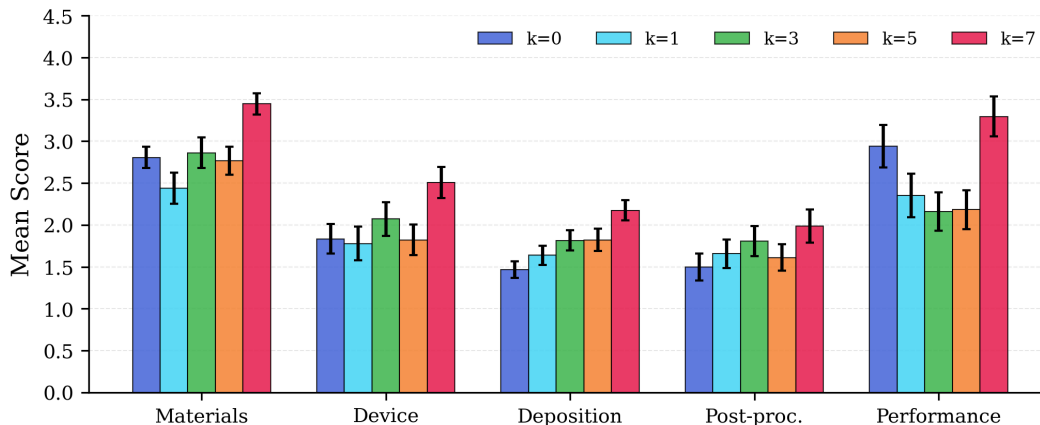


Figure 6: Ablation study on the number of retrieved documents (k). Performance across five evaluation categories improves consistently with retrieval augmentation, with gains saturating around $k = 5$.

gains as k increases, with scores improving from approximately 1.5–2.0 at $k = 0$ to 2.5–3.5 at $k = 5$. However, the Deposition category shows more modest improvements with higher variance, suggesting that retrieved recipes from heterogeneous experimental setups may introduce conflicting numerical conditions rather than providing consistent guidance. This observation aligns with the limitation discussed earlier regarding parameter accuracy degradation when interpolating across diverse sources. Overall, performance gains saturate around $k = 5$, indicating that excessive retrieval can introduce noise without proportional benefit.

5 Conclusion

In this work, we presented an AI-guided framework for navigating the mobility-stability trade-off in oxide semiconductor thin-film transistors through systematic literature mining and retrieval-augmented recipe generation. By analyzing 1,030 mobility-stability pairs from 2,005 experimental studies, we established the first quantitative Pareto frontier for oxide TFT performance and identified key fabrication strategies that enable frontier-breaking devices. Our RAG-based recipe generator achieves 4.33/5.0 overall quality on 1,198 test cases, demonstrating practical utility in predicting fabrication conditions tailored to user-defined targets. While limitations remain in numerical parameter accuracy and experimental validation, this approach represents a paradigm shift from trial-and-error experimentation toward literature-grounded, AI-assisted process design that can significantly reduce development time and costs. We anticipate this methodology will accelerate oxide semiconductor development for next-generation applications and provide a scalable framework extensible to other material systems where design spaces are complex and experimental feedback is costly.

References

- Ian Hendy, John Brewer, and Sean Muir. Development of high-performance igzo backplanes for displays. *Information Display*, 38(5):60–68, 2022.
- Keisuke Ide, Kenji Nomura, Hideo Hosono, and Toshio Kamiya. Electronic defects in amorphous oxide semiconductors: A review. *physica status solidi (a)*, 216(5):1800372, 2019.
- Luisa Petti, Niko Münzenrieder, Christian Vogt, Hendrik Faber, Lars Büthe, Giuseppe Cantarella, Francesca Bottacchi, Thomas D Anthopoulos, and Gerhard Tröster. Metal oxide semiconductor thin-film transistors for flexible electronics. *Applied Physics Reviews*, 3(2), 2016.
- Hye-Mi Kim, Dong-Gyu Kim, Yoon-Seo Kim, Minseok Kim, and Jin-Seong Park. Atomic layer deposition for nanoscale oxide semiconductor thin film transistors: review and outlook. *International Journal of Extreme Manufacturing*, 5(1):012006, 2023.
- Yu-Shien Shiah, Kihyung Sim, Yuhao Shi, Katsumi Abe, Shigenori Ueda, Masato Sasase, Junghwan Kim, and Hideo Hosono. Mobility–stability trade-off in oxide thin-film transistors. *Nature Electronics*, 4(11):800–807, 2021.
- Zhong Pan, Yifan Hu, Jingwen Chen, Fucheng Wang, Yeojin Jeong, Duy Phong Pham, and Junsin Yi. Approaches to improve mobility and stability of igzo tfts: a brief review. *Transactions on Electrical and Electronic Materials*, 25(4):371–379, 2024.
- Jiazhen Sheng, Jung-Hoon Lee, Wan-Ho Choi, TaeHyun Hong, MinJung Kim, and Jin-Seong Park. Atomic layer deposition for oxide semiconductor thin film transistors: Advances in research and development. *Journal of Vacuum Science & Technology A*, 36(6), 2018.
- John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418, 2024.
- Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.
- Yeonghun Kang and Jihan Kim. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature communications*, 15(1):4705, 2024.
- Wonseok Lee, Yeonghun Kang, Taeun Bae, and Jihan Kim. Harnessing large language model to collect and analyze metal-organic framework property dataset. *arXiv preprint arXiv:2404.13053*, 2024.
- Hassan Ul Huzaibi, Su-Ting Han, and Meng Zhang. Model representation in amorphous metal oxide thin-film transistors: a critical review. *npj Flexible Electronics*, 2025.
- Michiel J van Setten, Hendrik FW Dekkers, Christopher Pashartis, Adrian Chasin, Attilio Belmonte, Romain Delhougne, Gouri S Kar, and Geoffrey Pourtois. Complex amorphous oxides: Property prediction from high throughput dft and ai for new material search. *Materials Advances*, 3(23):8413–8427, 2022.
- Jiho Lee, Jae Hak Lee, Chan Lee, Haeyeon Lee, Minho Jin, Jiyeon Kim, Jong Chan Shin, Eungkyu Lee, and Youn Sang Kim. Machine learning driven channel thickness optimization in dual-layer oxide thin-film transistors for advanced electrical performance. *Advanced Science*, 10(36):2303589, 2023.
- Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, Joonwon Jang, et al. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv preprint arXiv:2507.11407*, 2025.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.
- Yoon-Seo Kim, Won-Bum Lee, Hye-Jin Oh, TaeHyun Hong, and Jin-Seong Park. Remarkable stability improvement with a high-performance peald-izo/igzo top-gate thin-film transistor via modulating dual-channel effects. *Advanced Materials Interfaces*, 9(16):2200501, 2022.

346 Su-Hwan Choi, Jae-Min Sim, Jeongmin Shin, Seong-Hwan Ryu, Taewon Hwang, So Young Lim, Hye-Jin Oh,
347 Jae-Hyeok Kwag, Jun-Yeoub Lee, Ki-Cheol Song, et al. Unveiling the hybrid-channel (poly-si/igo) structure
348 for 3d nand flash memory for improving the cell current and gidl-assisted erase operation. *Small Structures*, 6
349 (5):2400495, 2025.

350 Seong-Pil Jeon, Jaehyun Kim, and Sung Kyu Park. Recent advances in metal oxide semiconductor-based
351 electronics: A review. *Transactions on Electrical and Electronic Materials*, pages 1–13, 2026.

352 Dong-Gyu Kim, Su-Hwan Choi, Won-Bum Lee, Gyeong Min Jeong, Ji Hyun Koh, Seunghee Lee, Bongjin Kuh,
353 and Jin-Seong Park. Highly robust atomic layer deposition-indium gallium zinc oxide thin-film transistors
354 with hybrid gate insulator fabricated via two-step atomic layer process for high-density integrated all-oxide
355 vertical complementary metal-oxide-semiconductor applications. *Small Structures*, 5(2):2300375, 2024a.

356 Yoon-Seo Kim, Taewon Hwang, Hye-Jin Oh, Joon Seok Park, and Jin-Seong Park. Reliability engineering of
357 high-mobility igzo transistors via gate insulator heterostructures grown by atomic layer deposition. *Advanced*
358 *Materials Interfaces*, 11(15):2301097, 2024b.

359 Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie
360 Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

361 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila
362 Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

363 Heegyu Kim, Taeyang Jeon, Seungtaek Choi, Ji Hoon Hong, Dong Won Jeon, Ga-Yeon Baek, Gyeong-Won
364 Kwak, Dong-Hee Lee, Jisu Bae, Chihoon Lee, et al. Towards fully-automated materials discovery via
365 large-scale synthesis dataset and expert-level llm-as-a-judge. In *Proceedings of the 34th ACM International*
366 *Conference on Information and Knowledge Management*, pages 1302–1312, 2025.

367 A Recipe Generator Evaluation Criteria

Table 2: Fourteen evaluation criteria used to evaluate synthesis recipes, categorized into materials, device structure, deposition procedure, post-processing, performance, and overall score. Each criterion is rated on a 1–5 scale to reflect the quality and practicality of the predicted recipes.

Category	Criteria	Description
Materials	Appropriateness	Are the channel materials, compositions, and crystallinity suitable for the target synthesis?
	Completeness	Are all material specifications properly identified?
Device Structure	Completeness	Are all device structure components (gate, insulator, substrate, electrodes, thicknesses, etc.) properly specified?
	Similarity	How closely does the device architecture match the ground truth?
	Feasibility	Is the proposed device structure realistic and manufacturable?
Deposition Procedure	Completeness	Are all necessary deposition parameters (temperature, pressure, power, gas flows, etc.) included?
	Accuracy	How well do the deposition parameters match the ground truth values?
	Feasibility	Can this deposition procedure be realistically executed in a lab?
Post-Processing	Completeness	Are annealing conditions (temperature, time, atmosphere) properly specified?
	Similarity	How closely do post-processing steps match the ground truth?
	Feasibility	Can this post-deposition annealing procedure be realistically executed in a lab?
Performance	Appropriateness	Are the predicted electrical characteristics (mobility, threshold voltage, etc.) reasonable?
	Similarity	How well do predicted performance metrics match actual literature results?
Overall Score	–	Average score considering the recipe’s overall quality and practicality.

368 To rigorously evaluate the quality and practicality of the generated synthesis recipes, we developed a
 369 comprehensive evaluation framework comprising fourteen criteria. These criteria were designed in
 370 collaboration with domain experts to capture both the technical correctness and practical feasibility
 371 of oxide semiconductor TFT fabrication recipes. Each criterion is rated on a 1–5 scale, where 1
 372 indicates poor quality and 5 indicates excellent quality. The detailed descriptions of all evaluation
 373 criteria are presented in Table 2.

374 B Supporting Data

375 Supporting data and code for this article are available in an anonymized online repository at <https://anonymous.4open.science/r/An-AI-Guided-Framework-for-Mobility-Stability-Aware-Recipe-Generation-in-Oxide-Semiconductor-TFTs-BA4B>. The repository contains:

- 378 • **Datasets:** ALD_Data.xlsx, PVD_Data.xlsx, Seed_Paper_and_Author.xlsx
- 379 • **Extraction results:** ALD_extraction_test_results.xlsx, PVD_extraction_test_results.xlsx
- 380
- 381 • **Evaluation scores:** ALD_extraction_test_score.xlsx, PVD_extraction_test_score.xlsx
- 382
- 383 • **Prompts:** ALD_extraction_prompt.txt, PVD_extraction_prompt.txt
- 384 • **Code:** prediction.py (recipe generation), judge.py (LLM-as-a-Judge evaluation)

385 A detailed description of each file is provided in README.md.

AI Co-Scientist Challenge Korea Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[N/A]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "AI Co-Scientist Challenge Korea paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly state the main contributions: (1) construction of a mobility-stability dataset from 1,030 pairs extracted from 2,005 papers, (2) establishment of the first quantitative Pareto frontier for oxide TFT performance, (3) systematic analysis revealing key fabrication strategies (ALD vs PVD, multi-channel architectures, interface engineering), and (4) development of an LLM-based RAG recipe generator achieving 4.33/5.0 quality score on 1,198 test cases. These claims are supported by corresponding experimental results in Sections 4.1-4.3 and match the scope of literature-based AI-guided materials optimization presented throughout the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses several limitations in Section 4.3 ("Limitation and Failure Case Analysis"). Specifically, it identifies: (1) deposition parameter accuracy as the primary weakness, particularly for numerical fidelity; (2) degraded performance on low-frequency material systems due to insufficient coverage in the training corpus; (3) evaluation ambiguity where multiple valid process routes exist but current metrics penalize deviation from a single reference recipe; (4) retrieval-augmented generation introducing noise when interpolating precise numerical conditions from heterogeneous sources. The paper also acknowledges the need for future work on constraint-aware decoding and uncertainty quantification for out-of-distribution inputs.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present formal theoretical results such as theorems, lemmas, or mathematical proofs. Instead, it focuses on a data-driven and system-level analysis of oxide semiconductor literature and the development of a practical LLM-based deposition recipe prediction framework.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive methodological details for reproducibility: (1) literature collection strategy with 31 seed papers and 13 seed authors specified (Section 3.1), (2) two-stage filtering process reducing 64,852 papers to 2,005, (3) LLM-based extraction pipeline with RAG framework and manual validation on 60 papers achieving 95.2% accuracy (Section 3.2), (4) dataset construction yielding 1,030 mobility-stability pairs (Section 3.3), (5) RAG-based recipe generator implementation using EXAONE-4.0-32B model (Section 3.4), (6) evaluation methodology using LLM-as-a-Judge framework with GPT-4o on 1,198 test samples with 14 detailed criteria (Section 4.3, Table 2), and (7) access to an interactive web platform <https://bit.ly/45AD6XJ> where the extracted dataset, literature references, and recipe generator are made available for transparency and reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper explicitly states that "the extracted dataset, associated literature references, and the RAG-based recipe generator are made accessible through an interactive web platform <https://bit.ly/45AD6XJ>, where users can explore the underlying data and test recipe generation under user-defined targets" (lines 163-166). This provides public access to both the curated mobility-stability dataset of 1,030 pairs and the functional recipe generator model, enabling users to reproduce the main experimental results and validate the claims regarding recipe generation performance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides essential experimental details including: (1) dataset construction from 2,005 papers yielding 1,030 mobility-stability pairs (Section 3.3), (2) test set of 1,198 samples for recipe generator evaluation (Section 4.3), (3) model specification (EXAONE-4.0-32B) and baseline (GPT-4.1-nano), (4) RAG-based retrieval framework (Section 3.4), (5) evaluation methodology using LLM-as-a-Judge with GPT-4o and 14 criteria rated on 1-5 scale (Table 2), and (6) manual validation on 60 papers with 95.2% extraction accuracy (Section 3.2). Additional implementation details are accessible through the public web platform <https://bit.ly/45AD6XJ>.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not report error bars, confidence intervals, or statistical significance tests for the main experimental results. The recipe generator evaluation presents single-point performance metrics (e.g., overall score 4.334 vs 4.225) across 1,198 test samples without variance measures or significance testing. While the large test set size (1,198 samples) provides some robustness, the absence of error bars or statistical tests makes it difficult to assess whether the observed performance differences between EXAONE-4.0-32B and GPT-4.1-nano are statistically significant. The evaluation criteria scores in Table 1 also lack standard deviations or confidence intervals that would indicate measurement uncertainty across the test set.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides detailed compute resource specifications for reproducibility:

- **GPU Type:** 2× NVIDIA RTX A6000 GPUs (local server configuration)
- **Memory Requirements:** Approximately 48GB total VRAM required for EXAONE-4.0-32B model inference with tensor parallelism across 2 GPUs
- **Model Configuration:** 32B parameters, bfloat16 precision, 32K context length, batch size of 4, GPU memory utilization set to 90%
- **Software Stack:** vLLM $\geq 0.10.0$, transformers $\geq 4.54.0$, FlashAttention backend
- **API Compute:** OpenRouter API used for LLM-as-a-Judge evaluation (GPT-4o) with rate limiting configuration (0.5s between requests)

The dataset contains 4,847 oxide semiconductor synthesis recipes evaluated using the EXAONE-4.0-32B model for prediction and GPT-4o for judgment scoring.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://nips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research adheres to the NeurIPS Code of Ethics by relying exclusively on publicly available scientific literature and computational analysis, without involving human subjects, personal data, or sensitive attributes. All data collection and analysis were conducted in a manner that preserves anonymity and respects intellectual property and usage rights.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper addresses broader impacts through its focus on accelerating materials discovery for semiconductor applications. Positive impacts include: (1) reducing experimental trial-and-error in oxide semiconductor development, potentially lowering R&D costs and time-to-market for display and memory technologies, (2) democratizing access to expert knowledge through the public web platform, enabling researchers with limited resources to benefit from collective literature insights, and (3) providing a scalable framework for process optimization that could be extended to other materials systems. Potential negative considerations are implicitly addressed through the limitation discussion, which acknowledges that the system may produce suboptimal recipes for under-represented material systems, potentially leading to wasted experimental resources if users rely on predictions without critical evaluation. The work is positioned as an assistive tool rather than autonomous decision-making system, appropriately scoping its role in the research workflow.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A] .

Justification: This work does not release models or datasets that pose a high risk of misuse or dual-use concerns. As a materials science study on oxide semiconductors focused on synthesis understanding and design, the research does not require additional safeguards for responsible release.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: EXAONE-4.0-32B [Bae et al., 2025] is released under the Apache 2.0 license. GPT-4.1-nano and GPT-4o were accessed via the OpenAI API under its terms of service. vLLM [Kwon et al., 2023] is distributed under the Apache 2.0 license. The literature data were extracted from publicly available scientific papers, with all source publications appropriately cited. Our dataset and code are released under the MIT license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: All assets are released with documentation at <https://anonymous.4open.science/r/An-AI-Guided-Framework-for-Mobility-Stability-Aware-Recipe-Generation-in-Oxide-Semiconductor-TFTs-BA4B>. The repository includes: (1) extracted ALD and PVD datasets in Excel format with column descriptions, (2) extraction prompts used for data curation, (3) recipe prediction and LLM-as-a-Judge evaluation scripts (prediction.py, judge.py), and (4) a README documenting file structure, data format, and usage instructions.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This study focuses on oxide semiconductor research through literature analysis, data extraction, and model-based synthesis planning, and does not involve any crowdsourcing activities, user studies, surveys, or experiments with human subjects. Therefore, no instructions to participants, compensation details, or related materials are applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: As this work is a materials science study on oxide semiconductors based on computational analysis and published literature, it does not involve human participants, personal data collection, or any procedures that could pose risks to human subjects. Consequently, IRB approval or an equivalent ethical review is not required for this research.

Guidelines:

- 806 • The answer NA means that the paper does not involve crowdsourcing nor research with
807 human subjects.
- 808 • Depending on the country in which research is conducted, IRB approval (or equivalent)
809 may be required for any human subjects research. If you obtained IRB approval, you
810 should clearly state this in the paper.
- 811 • We recognize that the procedures for this may vary significantly between institutions
812 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
813 guidelines for their institution.
- 814 • For initial submissions, do not include any information that would break anonymity (if
815 applicable), such as the institution conducting the review.