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

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Amorphous oxide semiconductors are advancing into next-generation memory
2 applications requiring high mobility ($>50 \text{ cm}^2/\text{V}\cdot\text{s}$) and robust electrical stability—
3 performance targets rarely achieved due to fundamental mobility–stability trade-
4 offs. Here, we present an AI-guided framework that mines 1,030 mobility–stability
5 pairs from 2,005 experimental studies, establishing the first quantitative Pareto
6 frontier for oxide thin-film transistor performance. Systematic analysis of frontier-
7 breaking devices reveals key fabrication strategies—including atomic-scale depo-
8 sition control, multi-channel architectures, and interface engineering—that col-
9 lectively transcend conventional performance boundaries. We develop a large
10 language model-based recipe generator using Retrieval-Augmented Generation
11 that predicts fabrication conditions tailored to user-defined targets. Benchmarking
12 on 1,198 test cases shows our model achieves 4.33/5.0 overall recipe quality, with
13 strong performance in post-processing prediction (+5.7%) and performance metrics
14 (+7.8%) versus baselines. This work demonstrates how literature-grounded AI
15 can assist process design, providing a scalable approach for navigating complex
16 process–property relationships in materials optimization.

17 1 Introduction

18 Oxide semiconductors have become the dominant channel material for thin-film transistors (TFTs),
19 enabling the successful commercialization of high-resolution displays since 2012 [Hendy et al.,
20 2022, Ide et al., 2019, Petti et al., 2016]. Building on this success, these are now being actively
21 explored for next-generation memory applications, including three-dimensional DRAM and vertical
22 NAND flash, where its ultralow off-state leakage current offers a distinct advantage. These emerging
23 applications impose more stringent performance requirements—demanding simultaneously high
24 mobility ($>50 \text{ cm}^2/\text{V}\cdot\text{s}$) and robust stability under electrical stress [Kim et al., 2023]. However,
25 meeting both criteria remains a fundamental challenge due to an inherent trade-off: approaches
26 that enhance carrier transport—whether through composition tuning, defect engineering, or process
27 modification—tend to introduce factors that compromise device reliability [Shiah et al., 2021, Pan
28 et al., 2024, Sheng et al., 2018]. Despite thousands of experimental studies, navigating this trade-off
29 remains extremely difficult and constitutes a major bottleneck for advancing device performance in
30 demanding applications.

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

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

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

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

60 2 Related Works

61 2.1 NLP/AI Methods for Literature Mining and Process–Property Database Construction

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

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

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

85 2.2 Prior Work to Optimize the Mobility–Stability Trade-off

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

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

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

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

101 **3 Method**

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

107 **3.1 Literature Collection and Data Curation**

108 The first step of this study involved systematic literature collection focusing on oxide TFTs for
109 display applications. Following the research scope defined in prior work, we targeted experimental
110 studies reporting both electrical performance and reliability characteristics of oxide TFT devices. To
111 construct a representative and sufficiently large corpus, we employed a seed-based literature crawling
112 strategy. Seed papers published up to 2012 were selected based on their foundational impact, as
113 indicated by high citation counts, typically exceeding 100 citations according to Google Scholar,
114 representing key early developments in oxide TFT research. In contrast, seed papers published from
115 2012 onwards were selected from high-impact journals ranked in the first and second quartiles (Q1
116 and Q2) to capture contemporary advances in the field. Seed authors were identified based on their
117 continuous publication record in oxide semiconductor research and substantial academic impact, as
118 reflected by individual citation counts exceeding 3,000. Specifically, 31 seed papers and 13 seed
119 authors were selected based on their relevance to oxide TFT research and their frequent citation in the
120 field. Using citation backtracking and keyword-based expansion, a total of 64,852 publications were
121 initially collected from scientific databases. Subsequent filtering was applied to remove irrelevant
122 studies. Papers unrelated to vapor-phase deposition processes or lacking device-level electrical
123 characterization were excluded. Solution-processed TFTs were also removed to maintain consistency
124 with industrially relevant fabrication techniques. After this two-stage filtering process, 2,005 papers
125 were retained as the final literature set for analysis.

126 **3.2 Data Extraction Using Large Language Models (LLMs)**

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

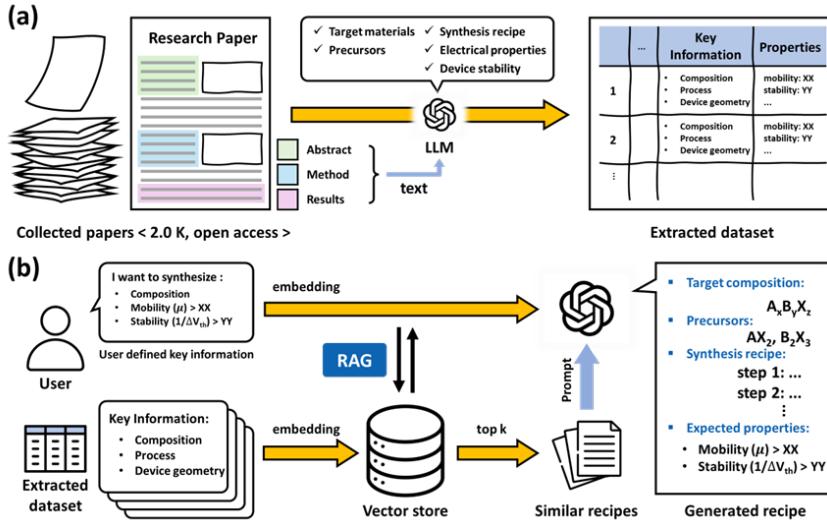


Figure 1: Schematic diagram illustrating the process of (a) constructing the recipe–property dataset and (b) inferring a synthesis recipe for a user-defined target using the RAG method.

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

143 3.3 Dataset Construction

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

153 3.4 Recipe Generator Model

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

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

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

173 4 Results & Discussion

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

179 4.1 Pareto Frontier and Mobility-Stability Landscape

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

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

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

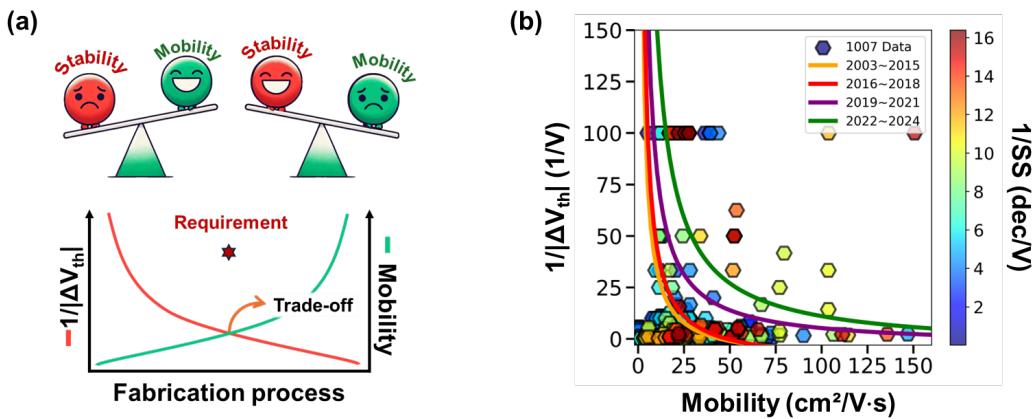


Figure 2: (a) Conceptual schematic of the mobility–stability trade-off in oxide TFTs. (b) Quantitative mobil-
 ity–stability landscape of collected oxide TFT data with time-resolved Pareto frontiers.

201 4.2 Analysis of Outlier Devices

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

204 process features associated with trade-off mitigation, 46 outlier devices were examined that achieve
 205 mobility above $30 \text{ cm}^2/\text{V}\cdot\text{s}$ and stability above 5 in units of $1/\Delta V_{\text{TH}}$, as shown in Figure 3(a). These
 206 devices represent a small but distinct subset of the overall dataset.

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

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

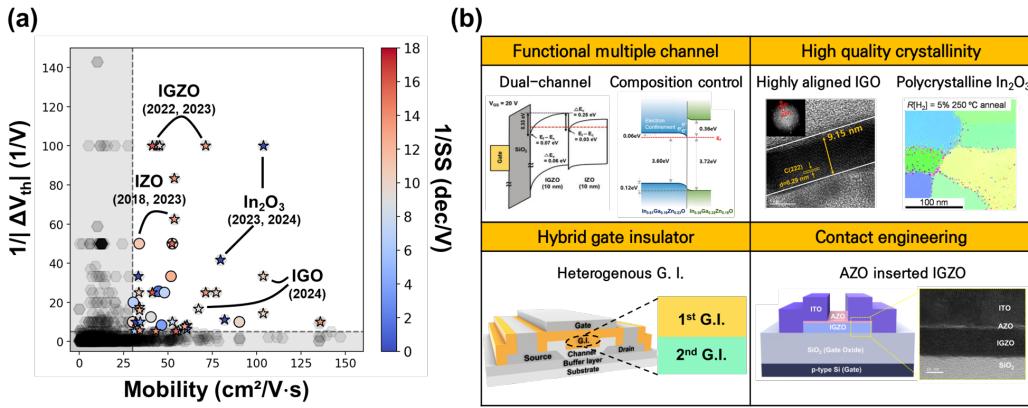


Figure 3: (a) Mobility–stability landscape with outlier devices highlighted. (b) Key strategies to overcome the mobility–stability trade-off in oxide TFTs, including functional multiple channels, high-quality crystallinity, hybrid gate insulators, and contact engineering. Reproduced from [Kim et al.,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

220 4.3 Recipe Generator Evaluation

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

227 **Evaluation Benchmark.** To rigorously evaluate the performance of the proposed recipe generator,
 228 we used an LLM-as-a-Judge [Gu et al., 2024] framework that takes advantage of GPT-4o [Hurst et al.,
 229 2024], following the prior materials synthesis recipe generation benchmark [Kim et al., 2025]. The
 230 evaluation was conducted on a comprehensive set of 1,198 test samples, ensuring statistical reliability.
 231 Each generated recipe was assessed using 14 detailed evaluation criteria that included material
 232 selection, device structural integrity, deposition parameters, post-processing steps, and alignment of
 233 target performance as shown in Table 2 of Appendix A. We compared EXAONE-4.0-32B [Bae et al.,
 234 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%

Success Case Example: Flexible IGZO TFT on Polyimide (ID 6)

Target Contribution (Input)

1. Key Contributions

- Deposited materials: SiNx (50 nm), Cr (35 nm), Al2O3 (25 nm gate insulator), IGZO (15 nm), Ti (10 nm), Au (60 nm), Al2O3 passivation (25 nm)
- Key Deposition/Process Method: PECVD (SiNx), e-beam evaporation (Cr, Ti, Au), ALD (Al2O3 gate insulator, 150 °C), RF-magnetron sputtering at room temperature (IGZO)

2. Device Structure

- Gate electrode material: Cr (35 nm, e-beam evaporated)
- Gate insulator thickness (nm): 25 nm
- Semiconductor (channel) thickness (nm): 15 nm (IGZO)
- S/D electrode material: Ti (10 nm) / Au (60 nm)
- Channel length (μm): 6 μm
- Passivation layer thickness (nm): 25 nm

5. Product Characteristics

- Field-effect mobility ($\text{cm}^2/\text{V}\cdot\text{s}$): 18.3 $\text{cm}^2/\text{V}\cdot\text{s}$
- Threshold voltage (V): 0.9 V
- On/off ratio (Ion/Ioff): $> 10^4$

EXAONE-4.0-32B Prediction

1. Key Contributions

- Deposited materials: SiNx (50 nm), Cr (35 nm), Al2O3 (25 nm gate insulator), IGZO (15 nm), Ti (10 nm), Au (60 nm), Al2O3 passivation (25 nm)

2. Device Structure

- Gate electrode material: N/A
- Gate insulator material: Al2O3
- Gate insulator process: ALD
- Gate insulator thickness (nm): 25
- Substrate material: N/A
- Semiconductor (channel) thickness (nm): 15
- S/D electrode material: N/A
- Channel length (μm): N/A
- Passivation layer material: Al2O3
- Passivation layer thickness (nm): 25

3. Deposition Process

- Deposition method: RF-magnetron sputtering
- Substrate temperature (°C): Room temperature

Device Structure Score: 4.33
Overall Score: 4.46

GPT-4.1-nano Prediction

1. Key Contributions

- Deposited materials: SiNx, Cr, Al2O3, IGZO, Ti, Au (no thickness values)

2. Device Structure

"The provided text does not specify individual sample structures, so this section will be marked as not available."

- Gate electrode material: Not extracted
- Gate insulator thickness (nm): Not extracted
- Semiconductor (channel) thickness (nm): Not extracted
- S/D electrode material: Not extracted
- Channel length (μm): Not extracted
- Passivation layer thickness (nm): Not extracted

3. Deposition Process

- Process Temperature: Max 150 °C
- Deposition Method: RF-magnetron sputtering
- Substrate temperature: Room temperature

Device Structure Score: 3.00
Overall Score: 4.00

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 Al_2O_3 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.

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

Failure Case Examples (ID 1165, 8, 2014)		
Deposition Parameter Accuracy (ID 1165) Ground Truth: <ul style="list-style-type: none"> - Deposition method: DC sputtering - Power: 30 W - Gas: Ar (4 sccm) 88% + Ar/O₂ (1 sccm) 20% - Process pressure: 3 mTorr - Substrate temperature: Room temperature EXAONE Prediction: <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 GPT Prediction: <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 Overall Score: EXAONE 4.15 vs GPT 4.65	Low Frequency Materials (ID 8) Ground Truth: <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 EXAONE Prediction: <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 GPT Prediction: <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 Overall Score: EXAONE 4.23 vs GPT 4.23	GT Similarity Limit (ID 2104) Ground Truth: <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 EXAONE Prediction: <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 GPT Prediction: <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) Overall Score: EXAONE 4.25 vs GPT 4.00

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.

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

259 **Limitation and Failure Case Analysis.** Analysis of failure cases reveals several systematic limitations
 260 as shown in Figure 5. First, deposition parameter accuracy remains the primary weakness,
 261 particularly for EXAONE-4.0-32B, where numeric fidelity degrades substantially compared to GPT-
 262 4.1-nano. This suggests that retrieval-augmented generation, while effective for structural and
 263 compositional reasoning, may introduce noise when interpolating precise numerical conditions from
 264 heterogeneous sources. Second, both models exhibit degraded performance on low-frequency mate-
 265 rial systems, indicating insufficient coverage of long-tail compositions in the training corpus. Third,
 266 we observe an inherent evaluation ambiguity: multiple valid process routes may exist for a given
 267 target, yet current metrics penalize deviation from a single reference recipe. This limitation motivates
 268 the development of equivalence-aware evaluation frameworks. Future work should address these
 269 challenges through constraint-aware decoding for critical parameters and uncertainty quantification
 270 to flag out-of-distribution inputs.

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

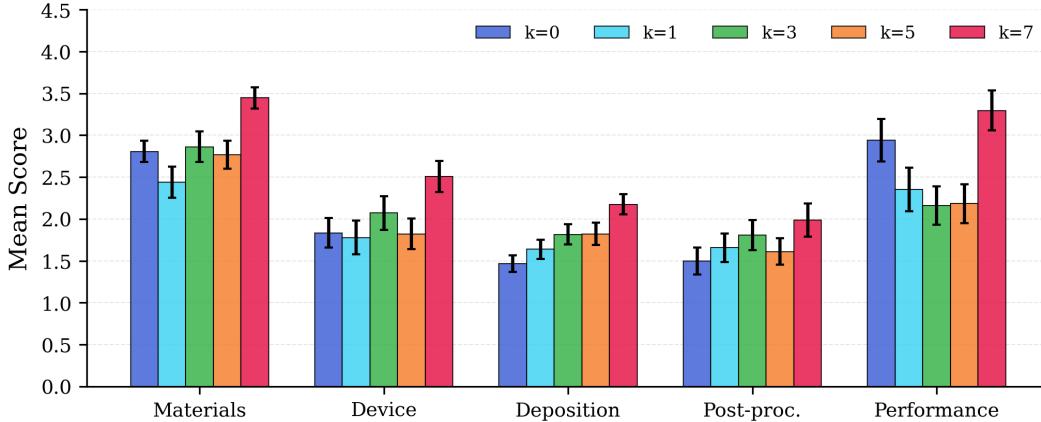


Figure 6: Ablation study on the number of retrieved documents (k). Performance across five evaluation categories improves consistently with retrieval augmentation, with gains saturating around $k = 5$.

278 gains as k increases, with scores improving from approximately 1.5–2.0 at $k = 0$ to 2.5–3.5 at
 279 $k = 5$. However, the Deposition category shows more modest improvements with higher variance,
 280 suggesting that retrieved recipes from heterogeneous experimental setups may introduce conflicting
 281 numerical conditions rather than providing consistent guidance. This observation aligns with the
 282 limitation discussed earlier regarding parameter accuracy degradation when interpolating across
 283 diverse sources. Overall, performance gains saturate around $k = 5$, indicating that excessive retrieval
 284 can introduce noise without proportional benefit.

285 5 Conclusion

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

298 **References**

- 299 300 Ian Hendy, John Brewer, and Sean Muir. Development of high-performance igzo backplanes for displays.
Information Display, 38(5):60–68, 2022.
- 301 302 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.
- 303 304 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.
- 305 306 307 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.
- 308 309 310 311 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.
- 312 313 314 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.
- 315 316 317 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.
- 318 319 320 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.
- 321 322 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.
- 323 324 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.
- 325 326 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.
- 327 328 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.
- 329 330 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.
- 331 332 333 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.
- 334 335 336 Jihoo 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.
- 337 338 339 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.
- 340 341 342 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.
- 343 344 345 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, Jihyun 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?
Deposition Procedure	Feasibility	Is the proposed device structure realistic and manufacturable?
	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?
Post-Processing	Feasibility	Can this deposition procedure be realistically executed in a lab?
	Completeness	Are annealing conditions (temperature, time, atmosphere) properly specified?
	Similarity	How closely do post-processing steps match the ground truth?
Performance	Feasibility	Can this post-deposition annealing procedure be realistically executed in a lab?
	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_
- 380 results.xlsx
- 381 • **Evaluation scores:** ALD_extraction_test_score.xlsx, PVD_extraction_test_sc
- 382 ore.xlsx
- 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.

386 **AI Co-Scientist Challenge Korea Paper Checklist**

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

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

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

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

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

411 **IMPORTANT**, please:

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

416 **1. Claims**

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

419 Answer: [Yes]

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

428 Guidelines:

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

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

438 **2. Limitations**

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

440 Answer: [Yes]

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

450 Guidelines:

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

477 **3. Theory Assumptions and Proofs**

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

480 Answer: [N/A]

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

485 Guidelines:

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

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

496 **4. Experimental Result Reproducibility**

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

500 Answer: [Yes]

501 Justification: The paper provides comprehensive methodological details for reproducibility:
 502 (1) literature collection strategy with 31 seed papers and 13 seed authors specified (Section
 503 3.1), (2) two-stage filtering process reducing 64,852 papers to 2,005, (3) LLM-based
 504 extraction pipeline with RAG framework and manual validation on 60 papers achieving
 505 95.2% accuracy (Section 3.2), (4) dataset construction yielding 1,030 mobility-stability pairs
 506 (Section 3.3), (5) RAG-based recipe generator implementation using EXAONE-4.0-32B
 507 model (Section 3.4), (6) evaluation methodology using LLM-as-a-Judge framework with
 508 GPT-4o on 1,198 test samples with 14 detailed criteria (Section 4.3, Table 2), and (7)
 509 access to an interactive web platform <https://bit.ly/45AD6XJ> where the extracted
 510 dataset, literature references, and recipe generator are made available for transparency and
 511 reproducibility.

512 Guidelines:

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

544 **5. Open access to data and code**

545 Question: Does the paper provide open access to the data and code, with sufficient instruc-
546 tions to faithfully reproduce the main experimental results, as described in supplemental
547 material?

548 Answer: [Yes]

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

556 Guidelines:

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

574 **6. Experimental Setting/Details**

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

578 Answer: [Yes]

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

587 Guidelines:

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

593 **7. Experiment Statistical Significance**

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

598 Answer: [No]

599 Justification: The paper does not report error bars, confidence intervals, or statistical significance tests for the main experimental results. The recipe generator evaluation presents
600 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
601 samples) provides some robustness, the absence of error bars or statistical tests makes it
602 difficult to assess whether the observed performance differences between EXAONE-4.0-32B
603 and GPT-4.1-nano are statistically significant. The evaluation criteria scores in Table 1 also
604 lack standard deviations or confidence intervals that would indicate measurement uncertainty
605 across the test set.

606 Guidelines:

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

627 8. Experiments Compute Resources

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

631 Answer: [Yes]

632 Justification: The paper provides detailed compute resource specifications for reproducibil-
633 ity:

- 634 • **GPU Type:** 2× NVIDIA RTX A6000 GPUs (local server configuration)
- 635 • **Memory Requirements:** Approximately 48GB total VRAM required for EXAONE-
636 4.0-32B model inference with tensor parallelism across 2 GPUs
- 637 • **Model Configuration:** 32B parameters, bfloat16 precision, 32K context length, batch
638 size of 4, GPU memory utilization set to 90%
- 639 • **Software Stack:** vLLM \geq 0.10.0, transformers \geq 4.54.0, FlashAttention backend
- 640 • **API Compute:** OpenRouter API used for LLM-as-a-Judge evaluation (GPT-4o) with
641 rate limiting configuration (0.5s between requests)

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

644 Guidelines:

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

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

655 **9. Code Of Ethics**

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

658 Answer: [Yes]

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

664 Guidelines:

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

670 **10. Broader Impacts**

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

673 Answer: [Yes]

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

687 Guidelines:

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

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

710 11. Safeguards

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

714 Answer: [N/A].

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

719 Guidelines:

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

730 12. Licenses for existing assets

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

734 Answer: [Yes]

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

740 Guidelines:

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

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

756 **13. New Assets**

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

759 Answer: [Yes]

760 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)
761 extracted ALD and PVD datasets in Excel format with column descriptions, (2) extraction
762 prompts used for data curation, (3) recipe prediction and LLM-as-a-Judge evaluation scripts
763 (`prediction.py`, `judge.py`), and (4) a README documenting file structure, data format,
764 and usage instructions.
765

766 Guidelines:

- 767 • The answer NA means that the paper does not release new assets.
768 • Researchers should communicate the details of the dataset/code/model as part of their
769 submissions via structured templates. This includes details about training, license,
770 limitations, etc.
771 • The paper should discuss whether and how consent was obtained from people whose
772 asset is used.
773 • At submission time, remember to anonymize your assets (if applicable). You can either
774 create an anonymized URL or include an anonymized zip file.
775

776 **14. Crowdsourcing and Research with Human Subjects**

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

780 Answer: [N/A]

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

785 Guidelines:

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

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

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

800 Answer: [N/A]

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

805 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.