ABDATABENCH: Evaluating AI Agents for Evidence-Grounded Antibody Record Construction

Anonymous ACL submission

Abstract

AI agents promise to turn scattered scientific evidence into AI-ready datasets, yet existing benchmarks examine isolated NLP skills or paper-level summarization and leave open whether agents can assemble complete, evidence-grounded records across multimodal and heterogeneous scientific sources. To address this gap, we introduce ABDATABENCH, an agent-agnostic benchmark for antibody functional data construction from scientific literature and patents. ABDATABENCH contains 44 expert-annotated, OCR-processed documents, including 27 papers and 17 patents, yielding 279 antibody-level records under a unified 22-field schema. We further develop ABCURATOR, a record-centric multi-agent framework that treats each antibody as the unit for planning, evidence aggregation, enrichment, and validation, coupled with a skill-accumulation workflow that distills expert feedback into reusable agent competencies across evaluation rounds. Across proprietary and open-source models, the strongest configurations achieve scores of 78.3 and 76.0, respectively. Analysis shows that current agents achieve high antibody recall but remain limited by over-generation, sequence attribution errors, and incomplete affinity grounding. These results position antibody record construction as a challenging long-horizon information-extraction task and establish ABDATABENCH as a benchmark framework for reliable AI-driven scientific data generation.

1 Introduction

AI agents are increasingly deployed to accelerate scientific discovery (Lu et al., 2024; Boiko et al., 2023), yet their impact hinges less on hypothesis generation than on a quieter bottleneck: turning dispersed scientific documents into reliable datasets for downstream AI. Therapeutic antibody research makes this bottleneck especially stark. Antibody structure prediction (Ruffolo et al., 2023), de novo design (Shanehsazzadeh et al., 2024), and affinity maturation with protein language models (Hie et al., 2024) all depend on supervision that pairs sequence with function, and recent biomolecular foundation models (Jumper et al., 2021; Lin et al., 2023) underscore how tightly scaling such models is bound to the size and fidelity of curated training data. Yet antibody evidence itself—variable region sequences, CDR loops, chain pairing (Malesys et al., 2024), isotype, target and epitope (Vita et al., 2025), binding kinetics, cross reactivity, stability, and in vivo efficacy—remains scattered across papers, patents, supplementary tables, figures, and external sequence or structure databases rather than consolidated into curated resources.

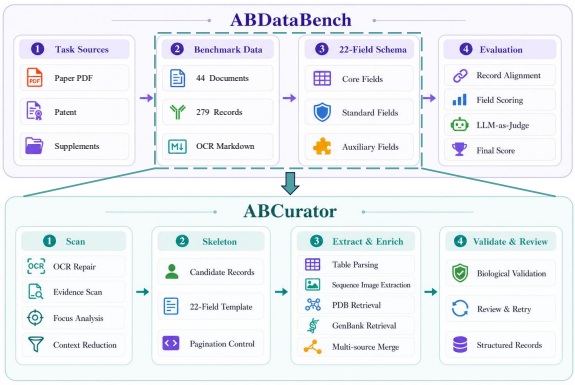

Figure 1: **Overview of ABDataBench and ABCurator.** ABDataBench is an agent-agnostic benchmark of 44 documents with 279 antibody records, and ABCurator is a multi-agent system for antibody data curation.

Antibody curation is a uniquely demanding setting. A single amino acid substitution can decide whether a record is usable for modeling (Starr et al., 2020), and a misattributed K_D (equilibrium dissociation constant) can attach the wrong function to an otherwise correct molecule. Functional identity is jointly defined by sequence, chain pairing (Malesys et al., 2024), target, epitope (Vita et al., 2025), affinity, reactivity, and efficacy, any of

which may appear under divergent aliases, assay-dependent units, or panels embedded in figures. Existing resources cover only parts of the space: SAbDab and Thera-SAbDab (Raybould et al., 2020) expose structural and therapeutic annotations, OAS (Olsen et al., 2022) aggregates repertoire sequences, and IMGT (Lefranc et al., 2009) standardizes nomenclature, but complete records that bind an antibody to its target, assay, measurement, and provenance are rarely curated at scale.

LLM- and agent-based extraction (Dagdelen et al., 2024; Lála et al., 2023; Singhal et al., 2023; Wu et al., 2023; Hong et al., 2023) offers a plausible path, yet antibody literature violates the assumptions of sentence-level information extraction (Yao et al., 2019): evidence spans text, tables, figures, supplements, and external databases; nomenclature drifts across aliases and construct variants; and numbers require unit normalization and assay-aware interpretation. Meeting this bar demands three shifts that no current benchmark or agent framework jointly provides (Table 1): organizing extraction around the *antibody record* as the atomic unit, fusing heterogeneous evidence under one schema while routing deterministic operations to reproducible tools (Gao et al., 2023), and closing a feedback loop with domain experts so that each curation round converts corrections into reusable agent competencies. We recast the task as **evidence-grounded antibody record construction** and instantiate these shifts as three coupled contributions:

- **Benchmark.** We release ABDATABENCH, the first benchmark that evaluates complete antibody records, pairing Hungarian record alignment with three-tier field weighting and hybrid rule plus LLM-as-Judge scoring that penalizes hallucination and omission.
- **Framework.** We develop ABCURATOR, a record-centric multi-agent framework that turns heterogeneous antibody literature into complete, evidence-grounded 22-field records, binding each antibody to its sequence, target, assay, and provenance while keeping closed-form operations on reproducible tools.
- **Skill-accumulation workflow.** We introduce a practical feedback loop in which domain experts classify recurring agent errors after each evaluation round and codify fixes as versioned skill modules loaded into agent prompts, enabling agent-and-scientist co-evolution.

2 Related Work

Antibody databases and functional data. Curated antibody databases serve different facets of the data landscape: SAbDab/Thera-SAbDab (Raybould et al., 2020) catalogs therapeutic antibody structures, OAS (Olsen et al., 2022) aggregates repertoire sequences from next-generation sequencing, AbYsis (Dunbar et al., 2014) integrates sequence-structure analysis, and IMGT (Lefranc et al., 2009) standardizes immunogenetics nomenclature. These resources primarily capture sequence or structural information. *Functional* data—binding kinetics, efficacy, cross-reactivity, thermal stability—are rarely curated systematically, creating a bottleneck for AI applications that require multi-objective antibody optimization. Unlike these curated repositories, ABDATABENCH targets the *construction* of complete antibody records from raw literature, evaluating whether AI agents can produce the structured entries that such databases ultimately host.

Scientific data-construction benchmarks. SciREX (Jain et al., 2020) and SciERC (Luan et al., 2018) evaluate relation extraction from scientific text but operate at the sentence or abstract level with flat schemas. BioRED (Luo et al., 2022) extends to document-level biomedical relations yet still targets isolated entity pairs rather than multi-field records. TDMS-IE (Hou et al., 2019) extracts task–dataset–metric tuples from NLP papers. More recent agent-oriented benchmarks—ScienceAgentBench (Chen et al., 2024), PaperBench (Starace et al., 2025), and InnovatorBench (Wu et al., 2025)—evaluate broader scientific reasoning but score task-level outputs (code artifacts, reproduction fidelity) rather than structured records with field-level provenance. None of these benchmarks jointly require multi-modal evidence fusion, tiered schema weighting, or hallucination/omission penalties. ABDATABENCH fills this gap by defining the antibody record as the atomic evaluation unit and scoring it through Hungarian alignment, deterministic checks, and calibrated LLM adjudication.

LLM-based scientific information extraction. LLMs extract structured data from scientific text with increasing accuracy (Dagdelen et al., 2024; Dunn et al., 2022). Domain-specific language models (BioBERT (Lee et al., 2020), BiomedBERT (Gu et al., 2022)) advance biomedical NLP but rely on predefined relation schemas and cannot han-

Table 1: **Comparison of scientific data-construction benchmarks across key evaluation dimensions.** *Task Unit* is the atomic object scored. Evidence modalities: T=text, Tab=tables, F=figures, S=supplements, DB=databases.

Benchmark	Task Unit	Evidence Modalities	Tiered Schema	Penalty
SciREX (Jain et al., 2020)	4-ary relation	T	×	×
SciERC (Luan et al., 2018)	Entity+Relation	T	×	×
BioRED (Luo et al., 2022)	Relation	T	×	×
TDMS-IE (Hou et al., 2019)	Tuple	T, Tab	×	×
ScienceAgentBench (Chen et al., 2024)	Task output	T, Tab	×	×
PaperQA/LitQA2 (Lála et al., 2023)	Answer	T	×	×
PaperBench (Starace et al., 2025)	Reproduction	T, Tab, F	×	×
InnovatorBench (Wu et al., 2025)	Artifact	T, Tab, F, S	×	×
ABDATABENCH (ours)	**Record**	**T, Tab, F, S, DB**	✓ (3 tiers)	✓

dle the flexible, multi-modal extraction required for antibody literature. Zero-shot and few-shot approaches (Wei et al., 2024) show promise for general extraction but struggle with precision-critical scientific data—exact sequence recovery, correct unit conversion, and entity-specific binding resolution. Our framework differs by combining LLM reasoning with deterministic tool-based operations, delegating closed-form checks to auditable code rather than relying entirely on model generation.

Multi-agent LLM systems. Multi-agent frameworks decompose complex tasks among specialized agents: AutoGen (Wu et al., 2023) enables multi-agent conversation, MetaGPT (Hong et al., 2023) applies role-based collaboration to software engineering, and autonomous agents plan chemical synthesis (Boiko et al., 2023) and generate research hypotheses (Lu et al., 2024). These systems target open-ended generation or software artifacts; ABCURATOR adapts the multi-agent paradigm to structured record construction, where correctness is defined by field-level alignment with expert annotations rather than task completion or novelty.

Document understanding for scientific papers. MinerU (Wang et al., 2024) and Nougat (Blecher et al., 2023) convert scientific PDFs to machine-readable formats. Our framework builds on MinerU's VLM-based OCR, which preserves tables, figure captions, and mathematical notation. However, document conversion alone does not solve the downstream challenge of binding extracted evidence to the correct antibody record—a gap that ABCURATOR addresses through its record-centric pipeline.

3 ABDATABENCH: Benchmark for Antibody Record Construction

ABDATABENCH evaluates AI agents' ability to construct structured, engineering-ready antibody records from primary scientific documents. Each task is derived from a paper or patent containing experimentally characterized antibodies, coupling the raw document with expert-annotated records and field-level evidence. This design captures the full curation workflow, from resolving antibody entities and locating dispersed measurements to producing records that can be directly consumed by downstream antibody engineering pipelines.

3.1 Overview of Benchmark

ABDATABENCH currently comprises 44 expert-annotated antibody curation tasks drawn from 27 research papers and 17 patent filings, yielding 279 densely annotated antibody-level records under a unified 22-field schema.

Document selection. Source documents were selected to maximize diversity along three axes: (1) therapeutic area, covering viral (SARS-CoV-2, HIV-1, influenza), oncology (HER2, EGFR, PD-L1), and parasitic (*P. falciparum*) targets; (2) document type, balancing peer-reviewed articles (27) with patent filings (17) that exhibit distinct layouts (claims, sequence listings, multi-panel figures); and (3) extraction difficulty, ranging from single-antibody papers with explicit tables to multi-antibody patents requiring cross-reference resolution across 50+ pages. Documents were sourced from PubMed, USPTO, and EPO, filtered to those published between 2018 and 2025 to reflect contemporary reporting conventions.

Annotation protocol. Each document was independently annotated by two domain experts (PhD-level immunologists or structural biologists with $\geq$3 years of antibody engineering experience). Annotators filled all 22 schema fields per antibody, recording provenance (source type, page/table/figure location, verbatim evidence quote) for every non-empty field. Disagreements were resolved through a third-party adjudication round. On a subset of 15 documents (100 records), inter-annotator

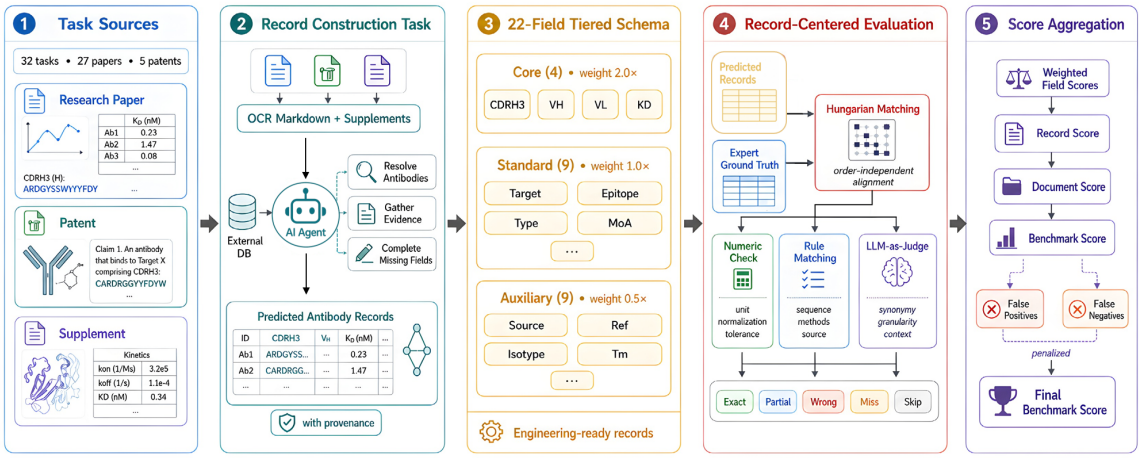

Figure 2: **ABDATABENCH benchmark workflow.** Papers, patents, and supplementary materials are processed by the extraction agent to produce structured antibody records under the 22-field tiered schema. Predicted records are then compared with expert records through order-independent record alignment, deterministic field checks, LLM adjudication, weighted aggregation, and false-positive and false-negative penalties to obtain final scores.

agreement measured by micro-averaged field-level Cohen's κ was 0.91 for categorical fields and 0.95 for numeric fields (within the 10% tolerance used in evaluation). The median annotation time was approximately 45 minutes per document.

Dataset statistics. The 44 documents yield 279 antibody records. Among these, 87% of records have at least one Core-tier field populated, 94% have at least one Standard-tier field, and 71% have at least one Auxiliary-tier field. Sequence fields (VH, VL, CDRH3) are present in 62% of records; binding affinity (K_D) is annotated for 48%.

3.2 Task Description

Each task provides an agent with the OCR-processed Markdown of a paper or patent, together with the associated supplementary materials as raw input. The agent must understand the document as a whole, identify all antibody names reported in the source, and integrate evidence from supplementary materials and external databases to infer and complete the corresponding sequence and experimental information. The agent must also provide provenance information for each extracted field, including the supporting source type, document location, evidence context, or external database identifier when available. The final output is a complete list of structured antibody information records.

Each complete antibody record consists of 22 standardized fields, which are grouped into **Core**, **Standard**, and **Auxiliary** tiers and assigned different weights according to their utility and importance for downstream antibody engineering (Table 3). Core fields primarily capture the most critical antibody sequence information and affinity metrics, which are the information most directly required for downstream structural modeling, affinity analysis, and antibody design. Standard fields describe the biological and experimental context of an antibody, providing the necessary basis for fully interpreting its function and the supporting experimental evidence. Auxiliary fields record extended information such as source, provenance, and target category, thereby improving data traceability and increasing the value of the schema for downstream integrative analyses. The complete field-by-field schema is provided in Appendix A.

3.3 Evaluation

Figure 2 summarizes the benchmark workflow. The evaluator compares model predictions with expert annotations through a record-centered protocol: align predicted and annotated antibody records, score each field, and aggregate into final scores.

Record alignment. Because extraction order is arbitrary and antibody names may vary, predicted records are aligned with expert annotations via the Hungarian algorithm (Kuhn, 1955) on a pairwise similarity score combining normalized antibody identity and contextual consistency.

Per-field scoring. Each matched antibody pair is evaluated over the 22-field schema. The evaluator assigns each field one of five verdicts: **exact** (full credit), **partial** (half credit), **wrong** or **miss** (zero credit), or **skip** (excluded). Scoring proceeds in three stages: (1) deterministic nu-

meric checks normalize values to common units and apply scalar tolerances for 7 quantitative fields (e.g., K_D, kon, koff, EC50); (2) rule-based string matching handles 6 fields covering sequences, experimental methods, antibody types, and references via edit-distance or set comparison; (3) the remaining 9 semantic fields (e.g., `Target_Name`, `Epitope`, `Mechanism_of_Action`) are adjudicated by an LLM-as-Judge (Zheng et al., 2024) (Gemini-3.1-Pro (Google DeepMind, 2026), not used as any extraction backbone) under field-specific rubrics. On a random sample of 120 field verdicts, the LLM judge achieves 91.7% agreement with expert annotations (Cohen's $\kappa = 0.87$). Complete per-field scoring rules and the judge prompt are provided in Appendix B and Appendix C.

Judge prompt calibration. We calibrated the LLM-as-Judge prompt to favor conservative positive judgments, requiring semantic equivalence to the expert annotation and preservation of field-specific biomedical constraints before assigning full credit. In an expert audit of positive LLM verdicts after calibration, all audited positive judgments were confirmed as true positives. This supports the use of the calibrated judge for semantic fields that require synonymy or granularity-aware adjudication, while deterministic checks remain the first-line evaluator for sequences, numerical values, units, antibody types, methods, and references.

Aggregation and penalty. Field verdicts are combined with three-tier schema weights ($w_f \in \{2.0, 1.0, 0.5\}$ for Core, Standard, Auxiliary) to produce a weighted record score. Document scores are averaged over all ground-truth antibodies and multiplied by a continuous penalty that penalizes unmatched annotations (λ_{FN}=0.05) more heavily than spurious predictions (λ_{FP}=0.01 per five extra records). The benchmark score is the mean of penalized document scores. The Hungarian similarity function combines normalized antibody-name identity and a five-field contextual consistency measure, rejecting assignments below threshold τ=0.5. Numeric fields use a two-level tolerance: $\leq$10% relative error for full credit and $\leq$3$\times$ fold difference for partial credit. Full scoring formulas are given in Appendix D.

4 ABCURATOR: A Record-Centric Multi-Agent Framework

General-purpose agents such as Claude Code (Anthropic, 2025) and Codex (OpenAI, 2025) provide flexible tool use for software engineering, whereas biomedical information-extraction systems typically produce named entities, relations, or paper-level summaries. Antibody record construction imposes a different output contract: each record must jointly bind an antibody to its complete variable-region sequences, quantitative binding measurements, experimental context, and field-level provenance, and must do so across noisy primary papers, patents, supplementary tables, and figures. We present ABCURATOR, a record-centric multi-agent framework that converts such heterogeneous evidence into structured records ready for downstream biomedical and engineering workflows.

4.1 Design Principles

Records are the unit of coordination. The atomic output of ABCURATOR is neither a document summary nor an entity mention but an evidence-bound record: one antibody, a target or condition, an assay, a value, and the literal evidence that supports each field. All planning, enrichment, and validation are performed at the granularity of this record, so that measurements specific to a given target, construct, or assay remain separable until they are committed to the correct antibody.

Reasoning where needed, determinism where possible. Only stages that require semantic judgment over noisy context are delegated to LLM or VLM agents. Regex scanning, database retrieval, table parsing, sequence recovery, and schema checking are implemented as pure Python tools whose outputs are deterministic and auditable, which both stabilizes the pipeline and bounds model cost.

Provenance-preserving merge. Evidence from heterogeneous sources is merged under record-level priority rules: every field carries a source label and quote, stronger sources overwrite weaker ones only when locally attributable to the same record, and unsupported fields remain empty.

Closed-loop validation. A rule-driven validator and an LLM reviewer form a bounded correction loop: rule failures produce targeted field-level instructions, the orchestrator replays extraction under these instructions, and the loop terminates on convergence or after a small retry budget.

4.2 Pipeline

Document conditioning. An OCR repair pass normalizes markdown syntax and table layout while preserving scientific tokens. A deterministic regex

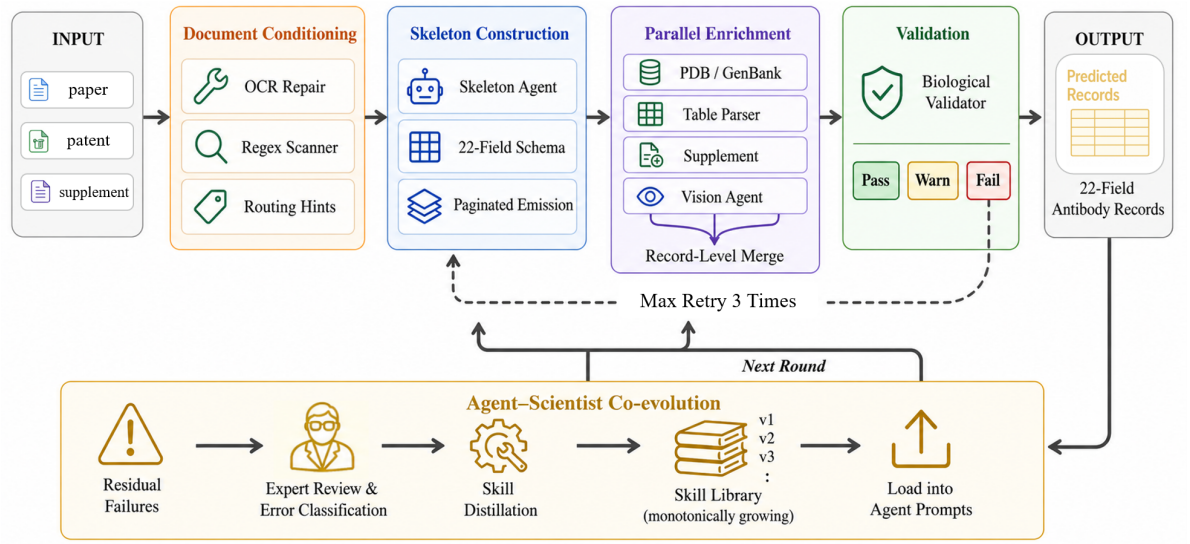

Figure 3: **ABCURATOR pipeline.** The Orchestrator converts OCR markdown and linked figures into antibody records through document conditioning, paginated skeleton construction, parallel evidence enrichment, biological validation, and bounded review. An outer co-evolution loop distills expert feedback on residual failures into versioned skill modules that are loaded into agent prompts at the next round, enabling monotonic capability growth.

scan locates antibody names, database accessions, CDR3 motifs, tables, and image references, emitting routing hints that activate downstream channels.

Skeleton construction. A Skeleton Agent produces candidate records via paginated calls, each emitting at most k records while tracking covered antibodies to eliminate duplication.

Parallel evidence enrichment. The skeleton is enriched through concurrent channels—database (PDB (Berman et al., 2000), GenBank (Sayers et al., 2022)), table, supplement, and vision—each targeting complementary fields, with results merged at the record level under source priority.

Biological validation. A rule-driven Validator checks amino-acid legality, sequence-length envelopes, CDR-H3 anchor consistency, and cross-record duplicates, emitting pass/warn/fail verdicts.

Review and bounded retry. On failure, a Reviewer Agent emits field-level corrections; the Orchestrator reruns extraction and validation until convergence or a retry budget.

Expert-guided skill evolution. Between rounds, domain experts review residual failures, classify error patterns, and formalize corrections into versioned skill modules that are loaded into agent prompts at the next cycle (Appendix E). This outer loop lets the pipeline absorb human expertise as durable, reusable competencies—each round compounds into measurable capability growth rather than requiring repeated manual intervention.

5 Experiments

5.1 Experimental Setup

We evaluate on the full 44-document, 279-record benchmark. Four proprietary models—Claude-4.6-Opus (Anthropic, 2026a), Claude-4.6-Sonnet (Anthropic, 2026b), GPT-5.5 (OpenAI, 2026), and Gemini-3.1-Pro (Google DeepMind, 2026)—and four open-source models—Qwen3.5-Plus (Qwen Team, 2026), DeepSeek-V4-Pro (DeepSeek AI, 2026), GLM-5.1 (Zhipu AI, 2026), and MiniMax-M2.7 (MiniMax, 2026)—are tested under identical conditions. All documents are pre-processed to Markdown via MinerU v2.7 (Wang et al., 2024) in VLM mode; the same artifacts are reused for every model so that Table 2 reflects model behavior rather than OCR differences. Missing outputs are scored as zero.

Table 2: **Model benchmark results on the ABDATABENCH.** Ab. Prec. and Ab. Rec. are antibody-record precision and recall. Seq. Hit and K_D Hit are hit rates for sequence and binding-affinity fields.

Model	Ab. Prec. ↑	Ab. Rec. ↑	Seq. Hit ↑	K_D Hit ↑	Score ↑
Proprietary Models					
Claude-4.6-Opus	**45.5**	86.4	**69.9**	**79.6**	**78.3**
Claude-4.6-Sonnet	36.8 $_{-8.7}$	**87.1** $_{+0.7}$	67.4 $_{-2.5}$	76.1 $_{-3.5}$	76.4 $_{-1.9}$
GPT-5.5	29.6 $_{-15.9}$	84.6 $_{-1.8}$	62.7 $_{-7.2}$	66.4 $_{-13.2}$	72.7 $_{-5.6}$
Gemini-3.1-Pro	29.4 $_{-16.1}$	75.3 $_{-11.1}$	41.0 $_{-28.9}$	44.2 $_{-35.4}$	63.1 $_{-15.2}$
Open-Source Models					
GLM-5.1	31.7	**97.8**	**66.8**	55.8	**76.0**
DeepSeek-V4-Pro	**49.6** $_{+17.9}$	84.9 $_{-12.9}$	64.2 $_{-2.6}$	**67.3** $_{+11.5}$	75.5 $_{-0.5}$
Qwen3.5-Plus	35.3 $_{+3.6}$	84.6 $_{-13.2}$	46.4 $_{-20.4}$	47.8 $_{-8.0}$	70.8 $_{-5.2}$
MiniMax-M2.7	32.5 $_{+0.8}$	80.3 $_{-17.5}$	41.8 $_{-25.0}$	34.5 $_{-21.3}$	59.5 $_{-16.5}$

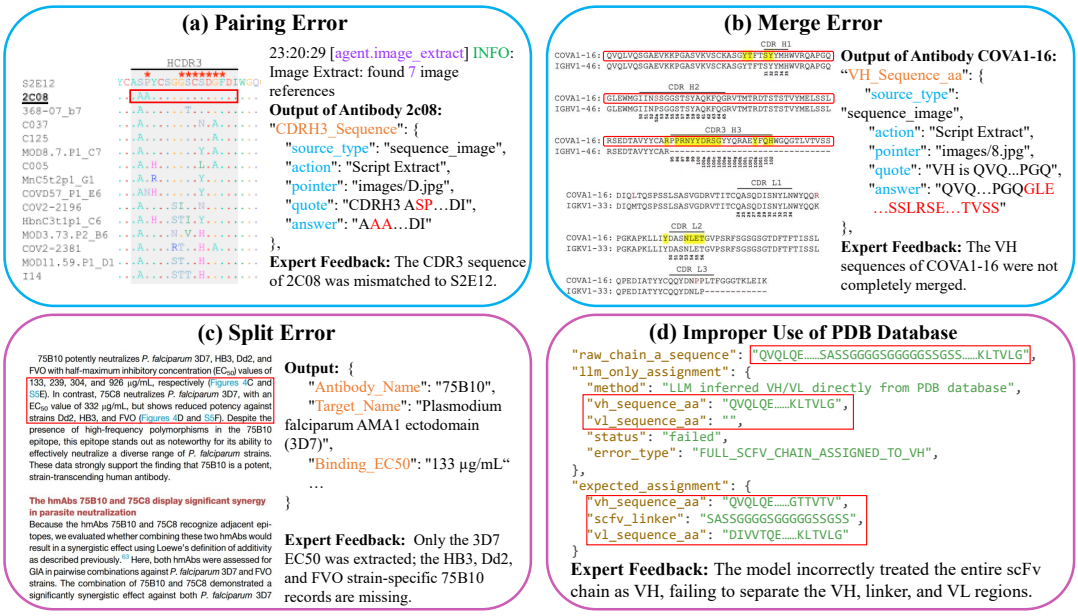

Figure 4: **Qualitative analysis of representative failure cases.** Four recurring error modes are shown: incorrect antibody–evidence pairing, incomplete sequence merging, failure to split strain-specific records, and improper assignment of full scFv chains from PDB data.

5.2 Model Benchmark Results

Table 2 shows that the strongest proprietary model, Claude-4.6-Opus, obtains the highest overall score (78.3), with the best proprietary antibody precision, sequence hit rate, and K_D hit rate. Claude-4.6-Sonnet follows at 76.4 with the highest proprietary antibody recall. GPT-5.5 reaches 72.7, while Gemini-3.1-Pro scores 63.1—the lowest among proprietary models—owing to substantially weaker sequence grounding (41.0 Seq. Hit) and affinity extraction (44.2 K_D Hit).

Among open-source models, GLM-5.1 achieves the highest overall score (76.0), driven by the best record recall (97.8) and sequence hit rate. DeepSeek-V4-Pro is close behind at 75.5 with the strongest open-source antibody precision (49.6) and K_D hit rate (67.3), indicating better control of over-generation and affinity grounding. Qwen3.5-Plus scores 70.8, while MiniMax-M2.7 reaches 59.5, mainly limited by weaker sequence and K_D recovery. All results use the full ABCURATOR skill configuration described in §4.

5.3 Qualitative Error Analysis

To interpret the quantitative gaps in Table 2, we manually reviewed representative failure cases with domain experts (Figure 4). The dominant failure modes are not generic formatting errors; they arise when a model must bind a value to the correct antibody, chain, strain, or external database record.

Pairing Error. The agent identifies the correct sequence panel but assigns the recovered CDRH3 to the wrong antibody row—e.g., attributing S2E12's CDR3 to 2C08—because multi-antibody panels lack explicit row-level anchors.

Merge Error. When a figure segments VH into separate FR/CDR blocks, partial merging yields an incomplete chain that omits one or more variable-region segments. For COVA1-16, only a subset of the alignment blocks were concatenated, directly reducing Seq. Hit for that record.

Split Error. A paper may report condition-specific values (e.g., EC50 per *P. falciparum* strain) that require separate antibody records per condition. For 75B10, the model emitted only the 3D7 record and omitted the HB3, Dd2, and FVO entries, lowering antibody recall.

Improper Use of PDB Database. A PDB chain encoding an scFv concatenates VH, linker, and VL without explicit domain boundaries. The model assigned the full chain to `vh_sequence_aa` and left `vl_sequence_aa` empty, producing a field-level error despite correct structural evidence.

5.4 Large-Scale Extraction Analysis

Beyond the curated benchmark, we test whether ABCURATOR scales to real antibody literature. We ran the pipeline over 48,584 antibody-related papers and patents, of which 39,586 produced valid outputs (81.5%). From these outputs, we randomly

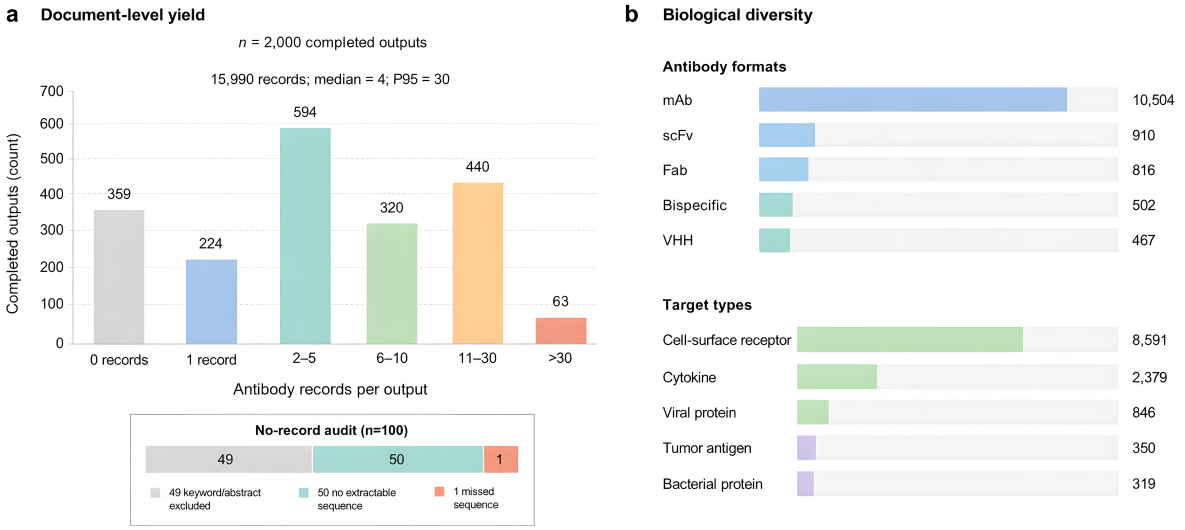

Figure 5: **Large-scale extraction profile.** (a) Document-level record yield and no-record audit in 2,000 sampled completed outputs. (b) Major antibody formats and representative therapeutic targets.

sampled 2,000 documents, yielding 15,990 antibody records. Overall, the sample shows three properties of real antibody data: record yield is long-tailed, engineering-critical fields are sparse, and the extracted records span diverse antibody formats and target biology.

Record yield. Among 2,000 sampled extraction outputs, 1,641 contained at least one antibody record, yielding 15,990 records in total, whereas 359 contained none (Figure 5a). Record counts were long-tailed, with a median of 4 records per positive output, a 95th percentile of 30, and a maximum of 115. In a manual audit of 100 no-record outputs, 49 were excluded at keyword or abstract screening, 50 lacked extractable sequence evidence after full-text and supplement review, and only one contained a sequence missed by ABCURATOR.

Biological diversity. The extracted corpus spans diverse antibody formats and therapeutic targets (Figure 5b), including common formats such as mAbs, scFvs, Fabs, VHHs, fusion proteins, and polyclonal antibodies, and representative targets such as EGFR, PD-1, HER2, VEGF, and SARS-CoV-2 Spike/RBD. A crude linear extrapolation to 98,315 crawled documents suggests approximately 6.4×10^5 antibody records, including roughly 1.3×10^5 high-quality records, with actual yield dependent on unseen document types.

6 Conclusion

We presented ABDATABENCH, the first benchmark for complete antibody record construction from scientific literature, and ABCURATOR, a record-centric multi-agent framework achieving 78.3 (proprietary) and 76.0 (open-source) on the 44-document, 279-record evaluation set. The benchmark pairs Hungarian record alignment with a three-tier weighted schema and hybrid rule plus LLM-as-Judge scoring, providing a rigorous evaluation protocol that penalizes both hallucination and omission. Residual errors concentrate in over-generation, sequence attribution, and affinity grounding, indicating that the core bottleneck is evidence binding rather than information recognition—a finding that points future work toward stronger entity-boundary reasoning and cross-modal provenance tracking. The skill-accumulation workflow demonstrates that expert feedback can be systematically distilled into reusable agent competencies, enabling monotonic capability growth across evaluation rounds without retraining. Beyond the curated benchmark, ABCURATOR produced 15,990 antibody records from a sample of 2,000 real-world documents, demonstrating that the same record-centric design scales to literature-wide curation and can serve as practical infrastructure for building large-scale antibody functional databases.

Limitations

No wet-lab validation. The benchmark measures extraction fidelity against expert annotations, not biological correctness. Whether agent-curated records support downstream tasks such as affinity-maturation training remains open.

Ethical Considerations

All source documents are publicly available scientific publications and patent filings; we redistribute only OCR-processed Markdown and structured annotations, not copyrighted full-text PDFs. ABDATABENCH contains no personally identifying information or individual-level clinical data. Annotations were performed by domain-expert collaborators on the research team. The benchmark does not introduce novel biological sequences beyond what is already publicly disclosed and does not create dual-use uplift; downstream users should still apply standard biosafety review when using extracted sequences to design new molecules.

Reproducibility and licensing. The benchmark, evaluation scripts, and extraction framework are released under the Apache 2.0 license. The Croissant 1.0 metadata file in the repository documents the dataset's structure and provides Responsible-AI fields covering data collection, preprocessing, annotation, use cases, biases, limitations, social impact, sensitive information, and a release/maintenance plan.

References

Anthropic. 2025. Claude code: An agentic coding tool. https://docs.anthropic.com/en/docs/claude-code. Accessed: 2025.

Anthropic. 2026a. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026.

Anthropic. 2026b. Claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026.

Helen M Berman, John Westbrook, Zukang Feng, and 1 others. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624:570–578.

Ziru Chen, Shijie Chen, Michael Ning, Qianheng Zhang, Boshi Wang, Xingyao Yu, and 1 others. 2024. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

DeepSeek AI. 2026. DeepSeek-V4. https://www.deepseek.com. Accessed: 2026.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, and 1 others. 2014. AbYsis: Integrated antibody sequence and structure—management, analysis, and prediction. *Journal of Molecular Biology*, 426:2216–2227.

Alexander Dunn, John Dagdelen, Nicholas Walker, and 1 others. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Google DeepMind. 2026. Gemini 3.1. https://deepmind.google/models/gemini/. Accessed: 2026.

Yu Gu, Robert Tinn, Hao Cheng, and 1 others. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1).

Brian L Hie, Varun R Shanker, Duo Xu, Theodora U J Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. 2024. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2):275–283.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *ACL*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *ACL*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna

Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Marie-Paule Lefranc, Véronique Giudicelli, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Fatima Bellahcene, Yan Wu, Eloi Gemrot, Xavier Brochet, Julia Lane, Lisandra Regnier, François Ehrenmann, Gérard Lefranc, and Pascale Duroux. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Research*, 37:D1006–D1012.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics*.

Simon Malesys, Rachel Torchet, Bertrand Saunier, and Nicolas Maillet. 2024. AntiBody Sequence Database. *NAR Genomics and Bioinformatics*, 6(4):lqae171.

MiniMax. 2026. MiniMax-M2.7. https://www.minimax.io. Accessed: 2026.

Tobias H Olsen, Fergus Boyles, Brennan S Sherborne, and 1 others. 2022. OAS: The observed antibody space database. *Cell Reports*, 34(6).

OpenAI. 2025. Codex cli. https://github.com/openai/codex. Accessed: 2025.

OpenAI. 2026. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026.

Qwen Team. 2026. Qwen3.5. https://qwenlm.github.io. Accessed: 2026.

Matthew IJ Raybould, Clare Marks, Alan P Lewis, and 1 others. 2020. Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Research*, 48:D383–D388.

Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. 2023. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14:2389.

Eric W Sayers, Evan E Bolton, J Rodney Brister, and 1 others. 2022. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50:D98–D106.

Amir Shanehsazzadeh, Sharrol Bachas, George Chow, and 1 others. 2024. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Giulio Starace, Aaron Jaech, and 1 others. 2025. PaperBench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.e20.

Randi Vita, Nina Blazeska, Daniel Marrama, IEDB Curation Team Members, Sebastian Duesing, Jason Bennett, Jason Greenbaum, Marcus De Almeida Mendes, Jarjapu Mahita, Daniel K. Wheeler, Jason R. Cantrell, James A. Overton, Darren A. Natale, Alessandro Sette, and Bjoern Peters. 2025. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Qiang Wei and 1 others. 2024. Zero-shot and few-shot learning for biomedical information extraction. *Briefings in Bioinformatics*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Xiangkun Hu, Pengrui Lu, Xiaojie Cai, Lyumanshan Ye, Wenhong Zhu, Yang Xiao, and Pengfei Liu. 2025. Innovatorbench: Evaluating agents' ability to conduct innovative LLM research. *arXiv preprint arXiv:2510.27598*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Zhipu AI. 2026. GLM-5.1. https://www.zhipuai.cn. Accessed: 2026.

A Full Schema of ABDATABENCH

Table 3: The ABDATABENCH schema comprising 22 scored fields grouped into three weight tiers reflecting downstream utility for computational antibody engineering.

Field	Type	Weight	Description
Core fields			
CDRH3_Sequence	Sequence	2.0	CDR-H3 amino acid sequence
vh_sequence_aa	Sequence	2.0	VH full variable-region sequence or IMGT germline assignment
vl_sequence_aa	Sequence	2.0	VL full variable-region sequence or IMGT germline assignment
Binding_Kinetics_KD	Numeric	2.0	Equilibrium dissociation constant (K_D)
Standard fields			
Target_Name	Text	1.0	Antigen name with variant information
Epitope	Text	1.0	Binding epitope or competition class
Antibody_Type	Categorical	1.0	Format classification (mAb, VHH, scFv, ...)
Mechanism_of_Action	Text	1.0	Functional mechanism (neutralization, ADCC, ...)
Experiment	Set	1.0	Experimental methods used for characterization
Binding_Kinetics_kon	Numeric	1.0	Association rate constant (k_{on})
Binding_Kinetics_koff	Numeric	1.0	Dissociation rate constant (k_{off})
Binding_EC50	Numeric	1.0	Half-maximal effective concentration
Structure	Text	1.0	Structural data (PDB ID, resolution, method)
Auxiliary fields			
source	Categorical	0.5	Species of origin (human, murine, rabbit, ...)
Reference_Source	Text	0.5	Publication DOI, PMID, or author–year citation
Target_Type	Categorical	0.5	Target molecular category
Antibody_Isotype	Categorical	0.5	Isotype and subclass (IgG1, IgG2a, ...)
Cross_Reactivity	Text	0.5	Cross-species or cross-variant reactivity profile
Quantitative_Metric	Numeric	0.5	Supplementary quantitative measurements (IC50, etc.)
In_Vivo_Half_Life	Numeric	0.5	Pharmacokinetic half-life
In_Vivo_Efficacy	Text	0.5	In vivo protection or therapeutic efficacy data
Thermal_Stability_Tm	Numeric	0.5	Melting temperature (T_m)

B Field Scoring Criteria

Table 4: Field-specific scoring criteria used in ABDATABENCH. Deterministic checks are applied before LLM adjudication whenever possible.

ID	Field	Tier	Scoring method	Scoring criteria
1	CDRH3_Sequence	Core	Sequence rule	Amino acid strings are normalized before comparison. Exact credit is given for identical sequences or complete containment; partial credit is given for small edit distance; larger sequence disagreement is wrong.
2	vh_sequence_aa	Core	Sequence rule	Full heavy chain variable sequences follow the same normalization, containment, and edit distance rules. Non sequence descriptions, including germline assignments, are judged by the LLM rubric.
3	vl_sequence_aa	Core	Sequence rule	Full light chain variable sequences follow the same normalization, containment, and edit distance rules. Non sequence descriptions, including germline assignments, are judged by the LLM rubric.
4	Binding_Kinetics_KD	Core	Numeric parser	Concentration units are normalized to a common scale. Scalar values within the strict numerical tolerance receive exact credit, values within the relaxed fold tolerance receive partial credit, and larger discrepancies are wrong. Ranges or multi value expressions are deferred to the LLM judge.
5	Target_Name	Standard	LLM rubric	Predictions are exact when the core antigen identity matches. Additional or missing strain, lineage, clade, or variant qualifiers are tolerated when the target identity is preserved; different targets are wrong.
6	Epitope	Standard	LLM rubric	Exact credit requires the same binding region, domain, motif, or epitope class. Coarser but relevant descriptions are partial; a different binding region is wrong.
7	Antibody_Type	Standard	Type rule	Antibody format tags are canonicalized. Compatible coarse and fine categories receive exact credit; incompatible core formats are wrong.
8	Mechanism_of_Action	Standard	LLM rubric	Exact credit is given when the prediction contains the annotated core mechanism. Mechanistically related but incomplete descriptions are partial; contradictory or unrelated mechanisms are wrong.
9	Experiment	Standard	Method set rule	Experimental methods are canonicalized into method sets. Exact credit is given when annotated methods are covered and extra methods are direct readouts; partial credit is given for overlap or peripheral extra methods; no overlap is wrong.
10	Binding_Kinetics_kon	Standard	Numeric parser	Scientific notation is parsed into scalar rates. The same strict and relaxed numerical tolerances used for kinetic fields determine exact, partial, and wrong outcomes; ambiguous expressions go to the LLM judge.
11	Binding_Kinetics_koff	Standard	Numeric parser	Scientific notation is parsed into scalar rates. The same strict and relaxed numerical tolerances used for kinetic fields determine exact, partial, and wrong outcomes; ambiguous expressions go to the LLM judge.
12	Binding_EC50	Standard	Numeric parser	Concentration units are normalized to a common scale. Scalar values are judged by strict and relaxed numerical tolerances; ranges or multi value expressions are deferred to the LLM judge.

Continued on next page

ID	Field	Tier	Scoring method	Scoring criteria
13	Structure	Standard	LLM rubric	Structural evidence is judged by consistency of deposited structure identifiers, experimental method, and resolution when available. Correct method with incomplete details is partial; conflicting structure evidence is wrong.
14	source	Auxiliary	LLM rubric	Source species or derivation is judged semantically. Exact credit requires the same source category; implicit or incomplete source descriptions are partial; conflicting source species are wrong.
15	Reference_Source	Auxiliary	Reference rule	Matching DOI receives exact credit. Author, year, and journal agreement also receives exact credit; incomplete bibliographic agreement is partial, and unrelated references are wrong.
16	Target_Type	Auxiliary	LLM rubric	Target molecular category is judged semantically. Compatible categories receive exact or partial credit depending on specificity; inconsistent categories are wrong.
17	Antibody_Isotype	Auxiliary	LLM rubric	Isotype and subclass are compared semantically. Compatible isotype information receives exact credit; missing subclass details are partial; conflicting isotypes are wrong.
18	Cross_Reactivity	Auxiliary	LLM rubric	Exact credit requires agreement on the relevant species, variants, or breadth profile. Partial coverage receives partial credit; contradictory reactivity profiles are wrong.
19	Quantitative_Metric	Auxiliary	LLM rubric	Supplementary quantitative measurements are judged by metric type, value, and condition. Partial coverage of multi part measurements receives partial credit; mismatched metrics or values are wrong.
20	In_Vivo_Half_Life	Auxiliary	LLM rubric	Half life values are compared semantically, including unit interpretation. Approximate or incomplete duration information is partial; inconsistent pharmacokinetic values are wrong.
21	In_Vivo_Efficacy	Auxiliary	LLM rubric	In vivo efficacy is judged by consistency of model, intervention, and outcome. Partial model or outcome coverage receives partial credit; conflicting efficacy claims are wrong.
22	Thermal_Stability_Tm	Auxiliary	Numeric parser	Melting temperatures are parsed as scalar temperature values. Scalar comparisons are judged by strict and relaxed numerical tolerances; ambiguous text is deferred to the LLM judge.

C LLM-as-Judge Prompt Design

The LLM judge is a constrained semantic adjudicator rather than a replacement for the deterministic evaluator. It is invoked only after empty value handling, numerical parsing, sequence comparison, experiment method canonicalization, antibody type matching, and reference matching fail to resolve a field. Each call compares one expert value with one predicted value for a fixed schema field. The system prompt specifies the scoring labels, tier priorities, numerical conventions, and field-specific semantic tolerances, while the user prompt provides the concrete field and value pair to be judged. The implementation uses zero temperature and enforces valid JSON labels and scores.

System Prompt Used by the Judge

You are a biomedical information extraction quality evaluator.
Your task is to compare the model prediction with the ground truth and assign a field score.
Labels. exact (1.0): semantically consistent and scientifically correct. Use exact when synonymous expressions have the same meaning; the prediction is more detailed than the ground truth but preserves the same core fact; the prediction is shorter but still preserves the core fact; numeric wording differs but the value is equivalent; empty markers such as n.d., N/A, ND, None, a hyphen marker, and the empty string are equivalent; the core germline gene segment is the same; or a sequence fully contains the expected core sequence.
partial (0.5): partly correct, with the right direction but important missing or shifted detail. Examples include partial numeric coverage, a correct mechanism missing key detail, or a reference missing author or article details.
wrong (0.0): contradictory or unrelated information, such as a different gene, sequence, target, epitope, mechanism, or value off by orders of magnitude.
Field Weights. Core fields include CDRH3_Sequence, vh_sequence_aa, vl_sequence_aa, and Binding_Kinetics_KD. They receive weight 2.0 and must be judged strictly, especially for sequences. Standard fields include Target_Name, Epitope, Antibody_Type, Mechanism_of_Action, Experiment, Binding_Kinetics_kon, Binding_Kinetics_koff, Binding_EC50, and Structure. They receive weight 1.0. Auxiliary fields include source, Reference_Source, Target_Type, Antibody_Isotype, Cross_Reactivity, Quantitative_Metric, In_Vivo_Half_Life, In_Vivo_Efficacy, and Thermal_Stability_Tm. They receive weight 0.5.
Binding Kinetics. Normalize concentration units for K_D and EC50 across pM, nM, uM, and M. A numeric difference of at most 10% is exact, a difference within three fold is partial, and a larger difference is wrong. Apply the same thresholds to kon, koff, and Tm when comparable.
Structure and Cross Reactivity. Structure text may include method, resolution, and PDB ID. The same method with similar resolution is exact, whereas the correct method with missing or different PDB detail is partial. For Cross_Reactivity, species and variant coverage matter, and broadly compatible statements may be partial when specific species or variants are omitted.
Important Notes. Compare semantics, not wording or language. More detailed predictions are usually not penalized when the core fact is correct. Epitope descriptions can be exact when they refer to the same region, domain, groove, motif, or key residue set at a different detail level. For information extraction, give high scores when the scientifically important fact is correct.

Lenient Defaults for Selected Fields. For Target_Name, if the core target matches, missing or extra clade, subtype, strain, lineage, or variant qualifiers can still be exact. For Epitope, descriptions of the same binding region, groove, domain, motif, or core residue set can be exact even at different detail levels. For Antibody_Type, compatible coarse and specific types can be exact. For Experiment, if the prediction covers the ground truth methods and extra methods are still direct readout, kinetics, or binding measurements, exact is acceptable; peripheral methods such as Western blot, crystallography, challenge models, sorting, gating, or sequencing should reduce to partial. For Mechanism_of_Action, if the prediction includes a core mechanism from the ground truth, compatible extra mechanisms usually do not penalize it unless the prediction contradicts or omits the ground truth mechanism.

User Prompt Template Used by the Judge

Judge the match for this field.
Field: {field_name} – {field_description}
Weight level: {weight_level}
Field specific rule: {field_guidance}
Ground truth: {gt_value}
Model prediction: {pred_value}
Return raw JSON only, without code fences:
{"label": "exact"/"partial"/"wrong", "score": 1.0/0.5/0.0, "reason": "one sentence English explanation"}

D Scoring Formulas

Record score. For each matched antibody pair $(a, \hat{a})$, the record score is

$$R(a) = \frac{\sum_{f=1}^{22} w_f \cdot v_{a,f}}{\sum_{f=1}^{22} w_f}, \tag{1}$$

where $w_f \in \{2.0, 1.0, 0.5\}$ is the tier weight (Core, Standard, Auxiliary) and $v_{a,f} \in \{1, 0.5, 0\}$ is the verdict credit for field f.

Document score with penalty. Let $|\mathcal{A}_d|$ be the number of ground-truth records in document d, n_{FP} the number of unmatched predictions, and n_{FN} the number of unmatched annotations. The penalized document score is

$$S_d = \frac{\sum_a R(a)}{|\mathcal{A}_d|} \times 100 \times \max\bigl(0,\ 1 - 0.01\lfloor n_{\text{FP}}/5 \rfloor - 0.05\, n_{\text{FN}}\bigr). \tag{2}$$

Omissions are penalized more heavily (λ_{FN}=0.05 per record vs. λ_{FP}=0.01 per five spurious records) because missed antibodies represent irretrievable evidence loss. The final benchmark score is $S = \frac{1}{|\mathcal{D}|} \sum_d S_d$.

Hungarian similarity function. The cost matrix entry for record alignment combines antibody-name similarity and contextual consistency:

$$\text{sim}(\hat{a}, a) = 0.85\, \bar{N}(\hat{a}, a) + 0.15\, C(\hat{a}, a), \tag{3}$$

where $\bar{N}$ is the normalized name similarity (longest common subsequence ratio, boosted to $\geq$0.98 on exact-name match), and $C \in [0, 1]$ is a weighted context score over five shared fields:

Context Field	Weight
Target_Name	0.45
Antibody_Type	0.25
Experiment	0.15
Structure	0.10
Reference_Source	0.05

Assignments with $\text{sim} < \tau$ (τ=0.5) are rejected; unmatched predictions become n_{FP} and unmatched annotations become n_{FN}.

Numeric tolerance. Deterministic numeric checks accept a prediction x_{pred} against ground truth x_{gt} under two tolerance levels:

$$\text{Exact } (v{=}1): \ \frac{|x_{\text{pred}} - x_{\text{gt}}|}{|x_{\text{gt}}|} \leq 0.1, \qquad \text{Partial } (v{=}0.5): \ \frac{1}{3} \leq \frac{x_{\text{pred}}}{x_{\text{gt}}} \leq 3. \tag{4}$$

Values outside both ranges receive v=0.

E Agent Skill and Prompt Modules

This appendix lists the reusable agent skills and prompt modules used to operationalize domain feedback in ABCURATOR. The skills define task-specific procedures for OCR repair, evidence grounding, sequence handling, and validation, while the prompts specify the extraction and review instructions passed to the corresponding agents.

E.1 Skill Modules

OCR Format Repair Skill

```
---
name: ocr-format-repair
description: Conservatively repair OCR-derived Markdown formatting while preserving every scientific token.
agent: ocr_repair

---
# When to Use
Use this skill before extraction when OCR Markdown contains broken table rows, broken headers, HTML wrappers,
duplicated spaces, or layout-induced line breaks.
# Inputs
Paper identifier, chunk index, total chunk count, and one OCR Markdown chunk.
# Procedure
Repair only formatting that can be fixed with tokens already present. Merge broken rows or headers, join OCR-split
lines, normalize whitespace, and repair obvious Markdown or HTML wrappers. For HTML tables, prefer header-only
cleanup and keep row labels, body cells, values, and units unchanged.
# Guardrails
Never change numeric values, amino-acid sequences, nucleotide sequences, residues, accessions, PDB identifiers, SEQ
ID numbers, antibody names, species names, scientific wording, or table body cells. If a repair requires scientific
inference, keep the original text.
# Output
Return the repaired Markdown chunk, whether it changed, and a short edit summary.
```

Evidence Reduction Skill

```
---
name: evidence-reduction
description: Filter long OCR chunks to preserve antibody, sequence, structure, table, image, and quantitative evidence.
agent: reducer

---
# When to Use
Use this skill before skeleton extraction when long documents exceed the reliable context budget.
# Inputs
Paper identifier, chunk index, total chunk count, and one paper chunk.
# Procedure
Keep original sentences, tables, figure legends, image Markdown, supplementary references, sequence blocks, accessions,
PDB identifiers, antibody names, clone names, binding values, epitope evidence, structure evidence, and assay-local
quantitative evidence. Preserve retained snippets in source order.
# Guardrails
Do not summarize, paraphrase, or rewrite scientific content. Remove references, author metadata, acknowledgments,
broad background, generic discussion, long protocols, instrument settings, and method details unless they directly
contain antibody-level evidence.
# Output
Return retained original snippets, evidence-type labels, and a short note. If the chunk has no extraction value, return
empty retained text.
```

Sequence Image Extraction Skill

```
---
name: sequence-image-extraction
description: Read antibody VH, VL, and CDRH3 sequences from sequence alignment or sequence block images.
agent: sequence_image_extract

---
# When to Use
```

Use this skill when sequence or alignment images may contain visible antibody heavy-chain, light-chain, or CDRH3 amino-acid evidence.
Inputs
One candidate image or adjacent image group, local source context, and optional seed antibody names.
Procedure
Transcribe visible amino-acid rows exactly. In alignments, retain gap characters and restore dot characters only from a visible reference row. Attribute sequences only to antibodies whose names and sequence rows are locally connected.
Guardrails
Do not infer invisible residues from background knowledge, germline labels, parent/WT labels, PDB identifiers, accessions, or prose context. Treat germline, consensus, WT, parent, reference, and control rows as references unless the image explicitly defines them as independent antibodies. Do not fill VH/VL from short CDR fragments or local tails.
Output
Return one record per supported antibody with visible VH, VL, and CDRH3 fields. Missing fields remain empty.

PDB Postprocess Backfill Skill

---
name: pdb-postprocess
description: Normalize PDB lookup summaries into mergeable antibody structure and sequence backfill fields.
agent: api_fetch

When to Use
Use this skill after external structure lookup when candidate PDB summaries and heuristic chain evidence need normalization before merging.
Inputs
Paper identifier, current antibody entry, candidate PDB summaries, and heuristic backfill evidence.
Procedure
Choose a structure only from the candidate PDB set. Fill VH and VL only from directly supported antibody variable-region chains. Fill CDRH3 only when supported by the selected VH sequence.
Guardrails
Do not add unsupported chains. Do not put linkers, tags, antigen sequences, constant regions, or noisy chain text into VH/VL. If attribution is uncertain, leave the field empty.
Output
Return one mergeable backfill object for the current antibody. Unsupported fields are empty strings.

E.2 Prompt Modules

System prompt for paper focus analysis

You are a pre-extraction analyzer for difficult immunology papers.
Your task is not to output final antibody JSON. Your task is to generate a short, executable paper-focus refinement for downstream extraction.
Context Overview
You will be given document excerpts, regex hints, and heuristic focus signals. Use only local evidence from these inputs. Do not introduce external knowledge.
Analysis Goals
Identify which antibodies are likely core study antibodies, which are likely comparator/control/reference-only antibodies, where records should be split, where sequences should be looked for first, and which local-binding risks are most important.
Special Directives
If supplementary tables, sequence listings, figure images, or tables are absent or truncated, explicitly state that affected fields should stay empty. If the paper contains FR/CDR row-order sequence blocks, instruct downstream extraction to read antibody names from FR1 order and bind later FR/CDR rows by that same order. If both VH and CDRH3 are present, instruct downstream validation to check that CDRH3 is derivable from the same VH.
Output Format
Return JSON only with llm_focus_summary, likely_core_antibodies, likely_reference_only_antibodies, record_split_alerts, sequence_source_priority, recommended_order, and local_binding_warnings. Use empty arrays or empty strings when evidence is insufficient.

System prompt for antibody skeleton extraction

You are a senior bioinformatics data architect for complex immunology literature.
Your task is to read document evidence and regex hints, then extract all antibodies with substantive data into benchmark-compatible flat records.

Article-Adaptive Protocol
Before extraction, decide whether the paper is table-led, supplement-led, figure-caption-led, sequence-led, structure-led, or mixed. Activate only the evidence modules supported by the current article. Regex hints and paper-focus analysis are for locating and prioritizing evidence; they never replace local source text.
Evidence-Unit Record Rule
A record represents an evidence unit: antibody name, target/object, condition/background, assay/readout, construct or format, and directly attributable value. Split records when target, condition, assay, construct, molecular format, strain, variant, cell background, or value attribution changes. Do not split for narrative summaries, generic cross-reactivity, or non-local differences.
Inclusion and Exclusion Policy
Include core study antibodies, recovered clones with clone-level sequence/isotype/target/binding data, image-supported study antibodies, and non-binding clones with sequence data. Exclude failed candidates, pure controls, commercial background antibodies, reference-only comparators, donor chains, unrecovered lineage nodes, MRCA or germline reversions, chain-swap constructs, public-clonotype search hits, and inferred-germline/control rows unless the paper reports independent antibody-level conclusions for that entity.
Anti-Broadcasting Policy
Experiment, Structure, In_Vivo_Efficacy, Epitope, and Mechanism_of_Action require local antibody-specific evidence. Do not copy a structure, method, mechanism, survival curve, epitope claim, or family-level conclusion to sibling records without a local name and condition anchor.
Sequence Policy
VH and VL fields require complete variable-region amino-acid sequences. Never fill them with germline names, V/J annotations, identity percentages, accessions, PDB IDs, dataset identifiers, grant numbers, catalog numbers, short tails, local CDR fragments, or natural-language placeholders. Resolve near-duplicate OCR conflicts by repeated support, full-chain consistency, SEQ ID evidence, and parent/variant chain evidence. For patent multi-column sequence tables, concatenate sequence columns row-wise. For FR/CDR row-order blocks, read antibody names from FR1 and bind later rows by the same order. CDRH3 must be explicit or derivable from the same VH.
Quantitative Policy
Exact local values with units override prose summaries. Fill K_D, kon, koff, EC50, IC50, quantitative metrics, and thermal stability only with precise values and units. Vague terms such as high affinity, strong binding, potent neutralization, nanomolar range, or broad reactivity do not replace exact table or panel values.
Output Format
Return strict JSON only. The top level is keyed by paper identifier. Each payload contains paper_id, title, optional category, and antibodies. Each antibody uses flat benchmark fields and string values. Missing, unsupported, unreadable, or non-unique fields are empty strings. Do not output nested value/quote objects, pointers, actions, confidence, germline helper fields, obsolete fields, Markdown fences, or explanation.

System prompt for figure VLM extraction

You are a biomedical figure triage and data-extraction expert.
Your task is to classify paper figures and extract visible antibody-related sequence, kinetics, quantitative, and efficacy data into mergeable JSON records.
Context Overview
You will receive one image or adjacent image group, local context, target antibody names, and missing fields. The image may contain sequence data, BLI/SPR kinetics, quantitative tables, in vivo efficacy plots, or no relevant data.
Triage Rules
First decide whether the image contains sequence data, kinetics data, quantitative table evidence, efficacy evidence, or no relevant extractable evidence. If the image is blurry or uncertain, treat it as not relevant.
Extraction Rules
Extract only data visible in the image. Do not guess values. Transcribe complete sequences without omission. Numeric values must include units. If one image contains multiple antibodies or multiple target-specific panels, output one record per antibody-data point. If the same antibody has different targets or different K_D values in different panels, output separate records.
Field Rules
Use Target_Name only when the target label is visible in the image. Use VH/VL fields only for complete amino-acid sequences, never germline gene names or annotations. Extract CDRH3 only when visible. Extract K_D, kon, koff, EC50, IC50, Tm, and efficacy values only with direct visual support. Do not output structure or mechanism information.
Efficacy Rules
Fill in vivo efficacy only when the same panel or legend locally binds the antibody name to the survival, protection, treatment, weight-loss, or viral-load value. Do not copy a group-level efficacy result to every antibody in a name list.
Output Format
Return only a JSON array. Each object may contain mAb, Target_Name, CDRH3, VH_sequence, VL_sequence, VDJ_Heavy, VJ_Light, K_D, kon, koff, EC50, IC50, Tm, and In_Vivo_Efficacy. Missing fields are empty strings.

F Detailed Experimental Results

Table 5: **Per-document ABDATABENCH scores for proprietary models.** Scores use the 44-document ground-truth order; absent predictions are scored as 0.

Task	Document	Opus-4.6	Sonnet-4.6	GPT-5.5	Gemini-3.1
1	Benjelloun et al. Jimmun. 2016	78.1	78.8	81.8	60.5
2	Chem Sci 2020	88.8	87.2	70.3	77.7
3	Colley et al. MABS 2018	89.2	81.6	90.2	90.2
4	Dacon et al. Science 2025	89.0	92.8	90.8	89.7
5	Hsiao et al. Nat Commun 2024	88.9	88.6	88.3	88.3
6	Ju et al. Cell 2025	84.8	54.0	66.7	76.5
7	Liu et al. Immunity 2020	72.5	76.7	92.1	95.3
8	Liu et al. Nat Commun 2019	32.2	81.7	82.8	28.9
9	Lowe et al. J Mol Biol 2011	91.9	85.9	91.3	91.6
10	Lu et al. J Biol Chem 2025	53.6	57.1	68.7	76.1
11	Malia et al. Proteins 2016	88.4	88.4	63.6	87.5
12	Nat Commun 2018	97.8	72.2	88.9	97.8
13	Nucleic Acids Res 2025	87.8	83.2	85.8	81.6
14	Olia et al. Fimmu. 2024	80.7	82.5	77.0	39.8
15	Patel et al. Cell Rep Med 2025	93.3	92.9	97.7	91.6
16	Peng et al. mAbs 2015	96.7	92.4	10.5	97.8
17	Pinto et al. Science 2021	71.7	72.2	68.0	75.8
18	PLoS Pathog 2021	86.8	88.2	60.0	86.2
19	Proteins 2005	98.4	92.4	92.4	98.4
20	Schmitz et al. Immunity 2021	86.0	88.9	80.6	60.8
21	Shi et al. Structure 2022	81.2	71.4	75.4	83.2
22	Sok et al. Immunity 2016	53.2	53.2	90.7	89.3
23	Song et al. Cell Reports 2025	79.8	81.0	71.8	61.1
24	Thai et al. Cell Rep 2023	78.3	80.0	81.5	13.7
25	Umotoy et al. Immunity 2019	79.4	79.0	70.1	75.8
26	US20080199481A1	0.0	0.0	0.0	21.5
27	US20180250395A1	89.0	84.0	84.1	33.0
28	WO2011004028A2	73.4	74.7	74.0	41.4
29	WO2016073890A1	78.7	74.1	43.6	43.6
30	WO2017021539A2	92.8	93.7	95.6	89.1
31	Zhou et al. Immunity 2018	85.2	85.2	79.3	64.3
32	Zhou et al. Science 2010	72.6	74.4	79.8	82.9
33	WO2022123040A1	90.2	92.0	59.0	92.0
34	US20160130366A1	91.9	96.6	90.5	27.4
35	US20250109195A1	90.4	89.9	81.0	47.4
36	WO2018068182A1	98.8	65.3	94.0	56.8
37	EP4050030A1	0.0	0.0	0.0	0.0
38	US11702478	73.8	84.7	95.7	23.6
39	EP3159358B1	87.0	85.6	26.8	0.0
40	WO2022049165A1	85.5	77.5	87.1	0.0
41	WO2022150740A1	89.6	84.8	87.2	55.0
42	WO2018048939A1	85.0	76.5	85.9	86.3
43	WO2019086573A1	95.3	94.6	98.9	97.8
44	US20250154275A1	37.8	24.6	0.0	0.0
Avg.	**Mean**	78.3	76.4	72.7	63.1

Table 6: **Per-document ABDATABENCH scores for open-source models.** Scores use the 44-document ground-truth order.

Task	Document	Qwen3.5	DeepSeek-V4	GLM-5.1	MiniMax-M2.7
1	Benjelloun et al. Jimmun. 2016	65.1	59.9	69.7	78.2
2	Chem Sci 2020	93.6	86.0	88.8	68.8
3	Colley et al. MABS 2018	75.0	41.3	88.0	38.2
4	Dacon et al. Science 2025	34.4	89.2	86.8	34.4
5	Hsiao et al. Nat Commun 2024	88.8	84.1	69.2	63.8
6	Ju et al. Cell 2025	85.7	80.7	72.9	58.3
7	Liu et al. Immunity 2020	94.9	96.7	70.2	96.5
8	Liu et al. Nat Commun 2019	85.6	83.2	91.1	89.6
9	Lowe et al. J Mol Biol 2011	82.8	90.8	84.4	58.9
10	Lu et al. J Biol Chem 2025	28.4	47.0	75.0	71.1

Task	Document	Qwen3.5	DeepSeek-V4	GLM-5.1	MiniMax-M2.7
11	Malia et al. Proteins 2016	95.1	83.7	76.0	54.6
12	Nat Commun 2018	92.3	97.8	87.9	49.0
13	Nucleic Acids Res 2025	87.1	79.8	82.9	37.2
14	Olia et al. Fimmu. 2024	37.9	81.0	81.4	37.9
15	Patel et al. Cell Rep Med 2025	85.6	93.5	90.3	80.6
16	Peng et al. mAbs 2015	88.8	96.7	19.5	0.0
17	Pinto et al. Science 2021	53.4	70.5	69.8	74.4
18	PLoS Pathog 2021	84.8	86.1	84.2	84.0
19	Proteins 2005	84.8	100.0	92.4	51.9
20	Schmitz et al. Immunity 2021	33.6	77.0	81.1	58.8
21	Shi et al. Structure 2022	90.1	90.5	59.6	49.5
22	Sok et al. Immunity 2016	88.5	97.6	96.1	90.5
23	Song et al. Cell Reports 2025	59.2	82.4	68.3	81.5
24	Thai et al. Cell Rep 2023	67.7	68.5	74.3	14.0
25	Umotoy et al. Immunity 2019	75.5	49.9	74.7	0.3
26	US20080199481A1	38.5	0.0	45.0	43.6
27	US20180250395A1	64.7	74.8	66.2	66.4
28	WO2011004028A2	63.3	76.8	63.4	58.7
29	WO2016073890A1	69.6	76.8	70.5	71.5
30	WO2017021539A2	86.8	85.6	96.2	23.8
31	Zhou et al. Immunity 2018	84.9	83.7	85.2	84.2
32	Zhou et al. Science 2010	75.1	74.4	71.8	69.0
33	WO2022123040A1	59.1	78.2	54.3	58.0
34	US20160130366A1	77.8	89.5	78.2	73.4
35	US20250109195A1	89.9	88.3	90.7	94.5
36	WO2018068182A1	99.0	69.8	97.6	97.0
37	EP4050030A1	0.0	0.0	79.2	1.1
38	US11702478	92.7	81.3	73.3	91.0
39	EP3159358B1	5.7	27.3	48.7	45.0
40	WO2022049165A1	83.3	84.8	78.4	78.8
41	WO2022150740A1	84.8	79.4	83.4	55.0
42	WO2018048939A1	87.5	85.3	85.9	88.2
43	WO2019086573A1	57.7	77.0	54.0	57.7
44	US20250154275A1	36.9	76.7	87.1	40.0
Avg.	**Mean**	70.8	75.5	76.0	59.5

G Abbreviations and Terminology

Abbreviation/Term	Meaning
Antibody structure & formats	
CDR	Complementarity-determining region
CDRH3	Heavy chain CDR3, the most critical loop for antigen binding
VH, VL	Variable heavy / light chain
FR	Framework region
scFv	Single-chain variable fragment
Fab	Fragment antigen-binding
VHH	Variable domain of heavy-chain antibody (nanobody)
mAb	Monoclonal antibody
Binding kinetics & measurements	
K_D	Equilibrium dissociation constant
k_{on}, k_{off}	Association / dissociation rate constants
EC50	Half-maximal effective concentration
IC50	Half-maximal inhibitory concentration
Tm	Thermal melting temperature
Experimental techniques	
BLI	Biolayer interferometry
SPR	Surface plasmon resonance
ADCC	Antibody-dependent cellular cytotoxicity
Databases & resources	
PDB	Protein Data Bank
GenBank	NIH genetic sequence database
IMGT	ImMunoGeneTics information system
SAbDab	Structural Antibody Database
Thera-SAbDab	Therapeutic Structural Antibody Database
OAS	Observed Antibody Space
Specific biological entities	
3D7, HB3, Dd2, FVO	Laboratory reference strains/isolates of *Plasmodium falciparum*
EGFR, PD-1, HER2, VEGF	Common therapeutic targets: epidermal growth factor receptor, programmed death 1, human epidermal growth factor receptor 2, vascular endothelial growth factor
SARS-CoV-2 Spike/RBD	Spike protein and its receptor-binding domain of SARS-CoV-2