RISK-CONTROLLED BIOMEDICAL CLAIM VERIFICATION OVER HETEROGENEOUS EVIDENCE SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Verifying a biomedical claim is adjudication across an ecosystem rather than lookup in a corpus. One assertion that a compound inhibits a target is recorded at once in abstracts, curated pharmacology records, interaction graphs and solved structures, each with its own provenance and update cadence, and these records disagree for structural rather than accidental reasons: a curated entry may assert an interaction that a later abstract negates under a different assay. Existing resources assume one corpus with one gold rationale, so they cannot express such conflict, and the verifiers built on them return unauditable labels alongside confidences that do not bound error. We introduce BioMSVerify, $10{,}206$ claims over four types with gold evidence spanning at least two source modalities, 813 with conflicting sources, and EVIDICT, a verifier fusing a generative energy over typed evidence, a reliability-weighted and direction-preserving contradiction score, and a bidirectional process reward model multiplying each step's grounding by its value-to-go, then turning the fused score into a commit-or-abstain decision by group-conditional split conformal prediction. Three measurements motivate the design. Closed-book PubMedBERT reaches 57.31 macro-F_1 overall and 71.66 on multi-hop drug–target–pathway–phenotype claims, so accuracy cannot certify grounding. Given an identical evidence pool, a frontier reasoning LLM reaches 81.34 while a 110M retrieval-grounded encoder reaches 87.82, so scale is not the bottleneck. And confidence is not risk: committing on temperature-scaled scores leaves a realized error of 0.219 on conflicting-source claims, twice a 0.10 target, whereas the conformal procedure keeps realized risk at or below α for every claim type and on the conflict subset at three target levels. EVIDICT improves macro-F_1 by 8.14 points on conflicting-source claims over the strongest baseline, is the most robust tested system to five counterfactual perturbation families, and costs $1/80$ as much at inference as GPT-4o.

1 INTRODUCTION

Biomedical knowledge does not live in one place. A statement such as "compound X inhibits kinase Y" may be asserted by a curated pharmacology record, implied by an interaction-graph edge, qualified by an abstract that saw the effect in one cell line, and constrained by a co-crystal structure. These sources are not redundant copies of one fact: they come from different communities, assays and times, and disagree for informative reasons. A curated biochemical assay and an abstract reporting no effect in cells can both be correct. A verifier that averages such a pair away as noise is wrong in exactly the cases a practitioner most needs help with.

Scientific claim verification has so far been built on the opposite assumption. Benchmarks from FEVER to SciFact (Thorne et al., 2018; Wadden et al., 2020; Sarrouti et al., 2021; Kotonya & Toni, 2020; Saakyan et al., 2021) pair each claim with one textual corpus and one gold rationale, and even the recent multimodal SciVer (Wang et al., 2025) draws its tables and figures from a single paper. None can express that two sources about the same claim conflict, so none can measure whether a verifier handles it. We make three observations about what follows from this gap.

Accuracy does not certify grounding. A closed-book PubMedBERT with no access to evidence reaches 71.66 macro-F_1 on our multi-hop drug–target–pathway–phenotype (DTPP) claims (Table 3), because parametric knowledge and surface regularities suffice to reject many implausible pathway

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

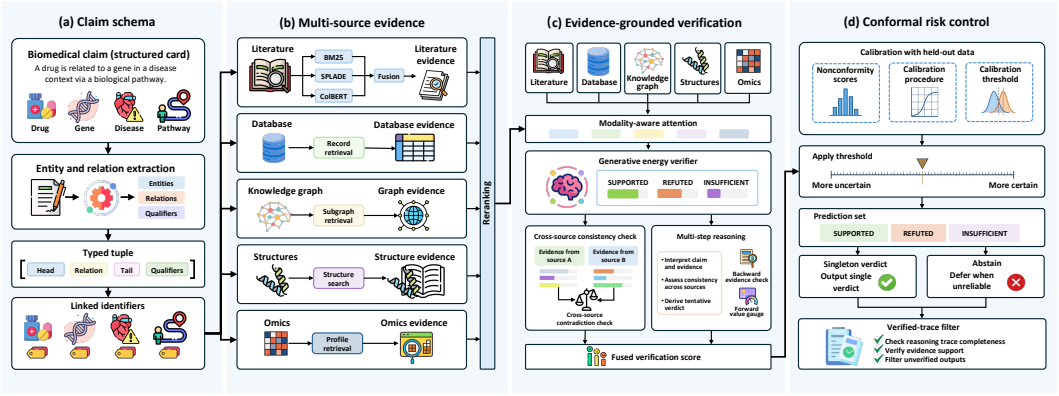


Figure 1: **EVIDICT.** (a) A claim is normalized to a typed tuple and linked to ontology identifiers. (b) Evidence is retrieved per source and merged by dev-tuned rank fusion; the omics channel is supported by the framework but unused on BioMSVerify, whose gold evidence spans text, curated records, graphs and structures. (c) Typed encoders feed three label-conditional scores: a generative energy, a reliability-weighted and direction-preserving cross-source contradiction score, and a bidirectional process reward whose backward term checks each step against evidence and whose forward term gauges its value toward the verdict. (d) A group-conditional split-conformal layer returns a single verdict when the prediction set is a singleton and abstains otherwise; the returned trace keeps only steps that pass the PRM's grounding check. Details in Section A.3.

chains. A label is therefore only as trustworthy as the evidence it is attached to, which is why we score evidence selection separately and credit it only when the cited source modality matches the gold pointer. *Scale is not the bottleneck.* Given an identical, controlled evidence pool, the strongest frontier reasoning model reaches 81.34 on BioMSVerify while a 110M retrieval-grounded encoder reaches 87.82 (Table 2); what limits the large models is how they read typed, structured evidence, not how much they know. *Confidence is not risk.* Temperature scaling makes our own scores well calibrated on average, yet committing on them leaves a realized error of 0.219 on conflicting-source claims (Table 5); a calibrated confidence says how often the model is right on average, not how often it will be wrong on the claims that matter.

We contribute BioMSVerify, 10,206 claims whose gold evidence spans at least two modalities, 813 with conflicting sources, and EVIDICT (Figure 1), a verifier with three label-conditional scores and a commit-or-abstain output under a finite-sample, group-conditional coverage guarantee. Across five public benchmarks and BioMSVerify, EVIDICT is best on five of six datasets; its margin is modest on single-relation claims (+1.15 on drug–target) and large where the evidence conflicts (+8.14). We are explicit about where it falls short: on SciVer, whose evidence is tables and figures (Section 5.1); on a benchmark artifact that lets closed-book models solve many pathway and mechanism claims (Section 5.2); and in a guarantee that covers test-time-visible groups rather than the gold conflict label (Section 5.4).

2 RELATED WORK

Scientific and biomedical claim verification. FEVER established retrieve-then-verify over Wikipedia (Thorne et al., 2018); SciFact, HealthVer, PubHealth and COVID-Fact moved the task to scientific and health claims over abstracts or articles (Wadden et al., 2020; Sarrouti et al., 2021; Kotonya & Toni, 2020; Saakyan et al., 2021), and SciVer to tables and figures (Wang et al., 2025). Strong verifiers include full-document NLI models such as MultiVerS (Wadden et al., 2022) built on biomedical encoders (Beltagy et al., 2019; Gu et al., 2021; Yasunaga et al., 2022), program-guided decomposition (Pan et al., 2023), and retrieval-augmented generation with reflection or graph structure (Lewis et al., 2020; Izacard et al., 2023; Asai et al., 2024; Edge et al., 2024; Gutiérrez et al., 2024). All of them assume a single evidence modality and a single consistent rationale. BioMSVerify differs in that disagreement between sources is part of the ground truth.

Table 1: **BioMSVerify composition.** Conf. counts conflict claims across all splits (they are contained in Train/Dev/Test, not added to them). Avg. #Ev./#Mod.: gold evidence items and distinct modalities per claim.

Claim type	Train	Dev	Test	Total	Conf.	Avg. #Ev.	Avg. #Mod.
Drug–target (DT)	2,590	589	712	3,891	291	3.8	2.3
Gene–disease–pathway (GDP)	1,630	371	448	2,449	204	4.1	2.4
Drug–target–pathway–phenotype (DTPP)	1,404	319	386	2,109	186	5.6	2.9
Molecular explanation (ME)	1,170	265	322	1,757	132	4.0	2.1
All	**6,794**	**1,544**	**1,868**	**10,206**	**813**	4.3	2.4

Verifiers and process rewards. Generative verifiers cast reward modeling as next-token prediction (Zhang et al., 2025), which connects naturally to energy-based learning (LeCun et al., 2006) and to the implicit reward of DPO (Rafailov et al., 2023). Process reward models score intermediate reasoning steps (Lightman et al., 2024; Khalifa et al., 2025) and can be trained from Monte-Carlo rollouts without step labels (Wang et al., 2024; Luo et al., 2024). Our PRM differs by factoring each step into a backward, evidence-grounding term and a forward, value-to-go term, so a step can be penalized for being ungrounded even when it leads to the right verdict, which keeps the learned score auditable at the level of individual steps.

Calibration, selective prediction and conformal methods. Post-hoc calibration (Guo et al., 2017) improves average agreement between confidence and accuracy but bounds no error rate; selective classification trades coverage for risk (Geifman & El-Yaniv, 2017). Split conformal prediction and conformal risk control give distribution-free, finite-sample guarantees (Vovk et al., 2005; Angelopoulos & Bates, 2023; Angelopoulos et al., 2024), and group-conditional (Mondrian) variants give the guarantee within pre-specified groups (Vovk, 2012; Cherian et al., 2024). Conformal methods have been applied to language-model factuality (Mohri & Hashimoto, 2024; Quach et al., 2024); we apply them to a verdict whose groups are defined by claim type and by the source profile of the retrieved evidence.

3 The BioMSVerify Benchmark

BioMSVerify covers four claim types: drug–target (DT), gene–disease–pathway (GDP), drug–target–pathway–phenotype (DTPP), and molecular explanation (ME). SUPPORTED claims are sampled from curated relations in DrugBank, Open Targets and Reactome (Knox et al., 2024; Buniello et al., 2025; Milacic et al., 2024). REFUTED claims are generated by the five counterfactual operators of Section 4.5 (affinity perturbation yields REFUTED only when assay conditions are comparable, and INSUFFICIENT otherwise), and further INSUFFICIENT claims by qualifier corruption that removes the comparability of conditions. Every generated claim is validated by domain annotators (Krippendorff's $\alpha_{\mathrm{K}} = 0.81$ on the validated subset (Krippendorff, 1989)). Labels are balanced ($3{,}402$ / $3{,}405$ / $3{,}399$). Each claim carries gold evidence pointers spanning at least two of four modalities (text, curated database, knowledge graph, molecular structure), and 813 claims are marked *conflict*: their gold sources disagree on the relation or its direction. The conflict subset is stratified across the train/dev/test splits in proportion to split size, so it is a subset of every split rather than a separate held-out set, and all "Conflict" results refer to its test portion. Table 1 gives the composition; statistics for all six benchmarks and construction details are in Section A.5.

4 Methodology: Evidict

4.1 Problem formulation and overview

Given a claim c, the verifier outputs a label $y \in \mathcal{Y} = \{\text{SUPPORTED}, \text{REFUTED}, \text{INSUFFICIENT}\}$ or ABSTAIN, together with a selected evidence set $E = \{e_i\}_{i=1}^{m}$ drawn from heterogeneous sources and a reasoning trace. We impose three requirements: the label must be supported by retrieved evidence rather than parametric knowledge alone; each step of a multi-step trace must be checkable against a concrete evidence item; and, within each group of claims the system can identify at test time,

the probability of committing to a wrong label must be at most a user-specified level α. EVIDICT computes three *label-conditional* scores and fuses them,

$$s_y(c, E) \;=\; -\mathcal{E}_\theta(c, E, y) \;+\; \lambda_c\, \mathcal{C}_\phi(c, E, y) \;+\; \lambda_p\, \mathcal{P}_\psi(c, E, y), \tag{1}$$

where each term is larger when the evidence better supports label y; $\tilde{p}(y \mid c, E) = \mathrm{softmax}_y\, s_y$ is the fused probability passed to the conformal layer (Section 4.6).

Normalization, retrieval and encoding. Claims are normalized to typed tuples $z(c) = \{(h, r, t, q)\}$ with qualifiers q (organism, cell type, assay, mutation, residue, condition) by an instruction-tuned biomedical LLM (Labrak et al., 2024), with entity linking by SapBERT (Liu et al., 2021) to DrugBank/ChEMBL, UniProt, UMLS, Reactome/KEGG/GO and PDB identifiers. Text is retrieved by BM25, SPLADE and ColBERTv2 over MedCPT embeddings (Robertson & Zaragoza, 2009; Formal et al., 2021; Santhanam et al., 2022; Jin et al., 2023), fused by weighted reciprocal rank fusion (Cormack et al., 2009) with dev-tuned weights; curated records $e_i^{\mathrm{db}} = (s_i, r_i, o_i, a_i, \rho_i)$ carry assay, organism, confidence and provenance; knowledge-graph evidence is retrieved as query-focused subgraphs (Edge et al., 2024); and structural evidence comes from PDB, PDBbind and BioLiP, with AlphaFold3 complexes and their pLDDT/PAE confidences substituted when no experimental structure exists (Berman et al., 2000; Liu et al., 2017; Yang et al., 2012; Abramson et al., 2024). We tested general-domain neural reranking and found that it *demotes* the structured evidence this task needs (Section A.10), so the final system keeps the top eight items per source by fused rank. Text and linearized records are encoded by PubMedBERT, subgraphs by an R-GCN, proteins by ESM-2, and complexes by EquiformerV2 with Uni-Mol ligand features (Gu et al., 2021; Schlichtkrull et al., 2018; Lin et al., 2023; Liao et al., 2024; Zhou et al., 2023); the pooled structural embedding is invariant to global rigid motions. Item embeddings are projected into the LLM's input space as soft tokens, alongside the linearized evidence text.

4.2 GENERATIVE ENERGY

The energy of a claim, evidence set and label is read from the backbone's likelihood of a verdict token, $\mathcal{E}_\theta(c, E, y) = -\log p_{\mathrm{LM}}(\langle y \rangle \mid c, E)$, which casts verification as next-token prediction (Zhang et al., 2025); renormalizing over the three labels gives $p_\theta(y \mid c, E) \propto \exp(-\mathcal{E}_\theta)$. Training combines the supervised loss $\mathcal{L}_{\mathrm{cls}} = -\log p_\theta(y^* \mid c, E)$, an InfoNCE term $\mathcal{L}_{\mathrm{nce}}$ over in-batch and retrieved hard-negative evidence sets (Gutmann & Hyvärinen, 2010; Oord et al., 2018), and a margin term requiring a true claim to receive lower SUPPORTED-energy than its counterfactual c^- under the same evidence, $\mathcal{L}_{\mathrm{rank}} = \max\{0,\, \gamma + \mathcal{E}_\theta(c, E, \textsc{Supported}) - \mathcal{E}_\theta(c^-, E, \textsc{Supported})\}$.

4.3 RELIABILITY-WEIGHTED, DIRECTION-PRESERVING CONTRADICTION SCORE

Each evidence item receives a stance toward the claim, $\pi_i(y) = p_\phi(\mathrm{stance}_i = y \mid e_i, c)$, from a DeBERTaV3 entailment model (He et al., 2021; Bowman et al., 2015; Williams et al., 2018) with ambiguous items escalated to an LLM judge (Zheng et al., 2023); stances are read against the normalized relation, so an item reporting activation is a REFUTED stance toward an inhibition claim rather than a generic "association". Every pair also receives a relation $r_{ij} \in \{\textsc{Entail}, \textsc{Contradict}, \textsc{Neutral}\}$. With item reliabilities $w_i = \sigma(\mathbf{a}_i^\top \mathbf{v})$ and pair weights $w_{ij} = \sigma(\mathbf{a}_{ij}^\top \mathbf{v})$ computed from source type, curation confidence, recency, evidence level and the agreement of organism, assay and disease context,

$$\begin{aligned} \mathcal{C}_\phi(c, E, y) &= \log \frac{\sum_i w_i\, \pi_i(y)}{\sum_i w_i} \;-\; \kappa(c, E)\, \mathbf{1}[y \neq \textsc{Insufficient}], \\ \kappa(c, E) &= -\sum_{i<j} w_{ij} \log\big(1 - p_\phi(r_{ij} = \textsc{Contradict})\big). \end{aligned} \tag{2}$$

The first term is a reliability-weighted vote; the second is a non-negative conflict mass that penalizes committing to either polar label when reliable sources contradict each other, and leaves INSUFFICIENT untouched. A conflict between a high-confidence curated record and a low-confidence abstract therefore moves little mass, while a conflict between two reliable sources moves the decision toward INSUFFICIENT and, through the conformal layer, toward abstention.

4.4 BIDIRECTIONAL PROCESS REWARD MODEL

For each candidate verdict y the backbone samples N reasoning traces $\tau_y = (s_1, \ldots, s_L)$ ending in y (Wei et al., 2022); single-relation claims yield one-step traces. Each step is scored by a backward grounding term and a forward value term,

$$P_\psi(s_\ell) = \underbrace{p_\psi^{\text{back}}(g_\ell = 1 \mid s_\ell, E)}_{\text{evidence grounding}} \cdot \underbrace{v_\psi^{\text{fwd}}(s_{\le \ell}; y)}_{\text{value-to-go toward } y}, \qquad \mathcal{P}_\psi(c, E, y) = \max_{\tau_y} \frac{1}{L} \sum_{\ell=1}^{L} \log P_\psi(s_\ell), \quad (3)$$

so a step is penalized for being ungrounded even when it leads to the right verdict, and for leading nowhere even when it is grounded. Grounding labels come from aligning normalized step triples to retrieved records; value labels are Monte-Carlo estimates of the probability that rollouts from the step reach the gold verdict (Wang et al., 2024; Luo et al., 2024). $\mathcal{L}_{\text{prm}}$ is step-level cross-entropy on g_ℓ plus a regression loss on v^{fwd}. The trace returned to the user keeps only steps whose grounding p_ψ^{back} exceeds 0.5, so every retained step points to a supporting item; this filter is a heuristic and carries no separate guarantee.

4.5 COUNTERFACTUAL NEGATIVES, PREFERENCE OPTIMIZATION AND OBJECTIVE

Five operators perturb a supported claim c into c^-: *entity substitution* within the same ontology neighborhood; *relation inversion* (ACTIVATES↔INHIBITS); *pathway substitution* to a plausible but unsupported pathway; *residue/interaction corruption*; and *affinity perturbation* ($|pK_{\text{claim}} - pK_{\text{evidence}}| > \epsilon$, labeled REFUTED when conditions are comparable and INSUFFICIENT otherwise). For each context $x \in \{(c, E), (c^-, E)\}$ we form a preference pair with the correct verdict y_w as chosen and the other claim's verdict y_l as rejected, and apply DPO (Rafailov et al., 2023) with the *same* context in numerator and denominator, $\mathcal{L}_{\text{dpo}} = -\log \sigma\big(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big)$; its implicit reward $\beta \log \pi_\theta / \pi_{\text{ref}}$ acts as a negative energy over the same verdict tokens that define $\mathcal{E}_\theta$. The full objective is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{nce}}\mathcal{L}_{\text{nce}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{dpo}}\mathcal{L}_{\text{dpo}} + \lambda_{\text{contr}}\mathcal{L}_{\text{contr}} + \lambda_{\text{prm}}\mathcal{L}_{\text{prm}}$, with $\mathcal{L}_{\text{contr}}$ the cross-entropy of stances and pair relations, and the backbone is adapted with QLoRA (Hu et al., 2021; Dettmers et al., 2023).

4.6 GROUP-CONDITIONAL CONFORMAL SELECTIVE VERIFICATION

With nonconformity score $\nu(c, E, y) = 1 - \tilde{p}(y \mid c, E)$ and a calibration set partitioned into groups $g \in \mathcal{G}$, let $\hat{q}_g$ be the $\lceil (n_g + 1)(1 - \alpha) \rceil$-th smallest score among the n_g calibration claims of group g (with $\hat{q}_g = \infty$ if that index exceeds n_g), and define $\mathcal{C}(c, E) = \{y : \nu(c, E, y) \le \hat{q}_{g(c,E)}\}$. EVIDICT commits to the single label when $|\mathcal{C}| = 1$ and abstains otherwise. Groups are the 4 claim types crossed with 3 *source profiles* (single-modality-dominant, concordant multi-source, and flagged-conflict), where the profile is computed from the retrieved evidence and the contradiction detector (Section A.2), never from the gold label.

Proposition 1 (Group-conditional coverage)**.** *If calibration and test claims are exchangeable within each group and $g(c, E)$ does not depend on y, then for every $g \in \mathcal{G}$, $\Pr[y^* \in \mathcal{C}(c, E) \mid g(c, E) = g] \ge 1 - \alpha$. Since* EVIDICT *can commit to a wrong label only by returning a singleton that excludes y^*, the probability that it commits to a wrong label on a claim of group g is also at most α. Both bounds hold for any union of groups.*

The proof is the standard exchangeability argument applied within each group (Vovk, 2012) (Section A.1); the procedure is split conformal prediction, equivalently conformal risk control (Angelopoulos et al., 2024) with the miscoverage loss. Two limits of the guarantee matter for how our results should be read. It holds in expectation over the calibration draw, so a realized risk slightly above α on one test set would not contradict it. And it covers the groups the system can *see*: the gold "Conflict" subset of BioMSVerify is not a calibration group, so the realized risk we report on it is an empirical check rather than a consequence of Proposition 1. ECE (Guo et al., 2017) and the Brier score (Glenn et al., 1950) are reported as diagnostics only.

Table 2: **Main results** (macro-F_1, %, mean over 5 seeds). "+R": identical multi-source evidence pool for every row. Representative rows per family; all 24 rows are in Table 8. **Best** / second over all 24 systems; *: EVIDICT significantly better than the best baseline in that column ($p < 0.01$).

Family	Model	SciFact	HealthVer	PubHealth	COVID-Fact	SciVer	BioMSVerify	Avg.
Trivial	Majority class	24.10	22.80	19.40	38.20	21.70	17.10	23.88
Closed-book	PubMedBERT	45.10	47.20	53.40	59.10	35.20	57.31	49.55
Encoder (+R)	PubMedBERT	76.90	73.50	68.20	74.10	54.30	87.20	72.37
	BioLinkBERT	77.80	74.60	68.90	74.80	55.10	87.82	73.17
Specialized	MultiVerS	80.10	76.30	70.20	77.50	56.40	84.63	74.19
	DeBERTaV3 pipeline	78.40	77.10	71.40	78.20	57.90	85.41	74.74
RAG	Self-RAG	76.80	74.20	69.40	76.10	59.20	83.16	73.14
	GraphRAG	75.40	73.10	68.60	74.90	58.10	84.27	72.40
LLM (+R)	GPT-4o	79.60	77.20	71.80	79.10	66.80	76.06	75.09
	Claude 4.7	81.30	78.90	73.20	80.40	71.50	78.41	77.29
	Gemini 3.1 Pro	80.70	78.10	72.60	79.80	**76.40**	77.31	77.49
Reasoning (+R)	o3	82.40	79.80	73.90	81.20	74.60	80.27	78.70
	DeepSeek-V4	83.10	80.50	74.10	81.70	75.20	81.34	79.32
Ours	**EVIDICT**	**84.70***	**82.30***	**75.40***	**83.20***	74.80	**89.21***	**81.60***

5 EXPERIMENTS

Datasets and baselines. We evaluate on SciFact, HealthVer, PubHealth, COVID-Fact, SciVer (zero-shot, test only) and BioMSVerify (statistics in Table 7), reporting macro-F_1 over each benchmark's native label set. Baselines span seven families (Section A.7): trivial predictors; closed-book encoders (Beltagy et al., 2019; Gu et al., 2021; Yasunaga et al., 2022); the same encoders fine-tuned on claim plus retrieved evidence; specialized verifiers (MultiVerS, a DeBERTaV3 NLI pipeline, ProgramFC); RAG systems (RAG, Atlas, Self-RAG, GraphRAG, HippoRAG); instruction-tuned LLMs (BioMistral-7B prompted without fine-tuning, Qwen2.5-72B, GPT-4o, Claude 4.7, Gemini 3.1 Pro (Hui et al., 2024; Achiam et al., 2023; Anthropic, 2026; Google DeepMind, 2026)); and reasoning LLMs (o3, DeepSeek-V4 (OpenAI, 2025; Xu et al., 2026)). Three controls keep the comparison fair: every retrieval-based method receives the *identical* multi-source evidence pool used by EVIDICT ("+R"), so gains are attributable to the verifier; every LLM uses chain-of-thought, five-sample self-consistency and Chain-of-Verification (Wei et al., 2022; Wang et al., 2022; Dhuliawala et al., 2024); and every method's decision and INSUFFICIENT thresholds are tuned on dev.

Protocol and metrics. Each number is the mean over five runs with seeds $\{13, 21, 42, 87, 2026\}$. For trained models the seed controls initialization, data order, negative sampling, trace sampling, and a random halving of the 1,544 dev claims into a 772-claim selection half (model selection, fusion and retrieval weights) and a 772-claim calibration half used only by the conformal layer, which keeps calibration exchangeable with test. For API models five passes permute evidence order, with self-consistency samples drawn at temperature 0.7. Significance (*, $p < 0.01$) is a two-sided paired bootstrap over test claims (10,000 resamples) against the strongest baseline in each column. Ev-F_1 is item-level F_1 against gold evidence pointers, credited only when the source modality also matches. Cov. is the fraction of test claims whose prediction set contains the gold label; Risk $= 1-$Cov., which upper-bounds the rate of wrong commitments; the nominal level is $\alpha = 0.10$ unless stated. Implementation details and the 278.2M trainable-parameter breakdown are in Section A.4.

5.1 MAIN RESULTS

EVIDICT is best on five of six datasets and on the average (Table 2); seed variability is small (± 0.22 on BioMSVerify, Table 9). Three patterns are more informative than the ranking. First, on the four text-only public benchmarks the gain over the strongest baseline, DeepSeek-V4, is steady but modest ($+1.3$ to $+1.8$), which is what we would expect when two of EVIDICT's three scores have little to act on: with a single textual source there is little cross-modal conflict to detect and little structure to ground. Second, SciVer is a clean negative result: its evidence is tables and figures, which none of our encoders reads, and Gemini 3.1 Pro leads (76.40 vs. 74.80). Third, on BioMSVerify the ordering of families inverts. Every frontier LLM falls below every fine-tuned encoder given the same pool (best LLM 81.34 vs. worst encoder 85.09), and prompted BioMistral-7B, our own backbone, reaches

Table 3: **BioMSVerify by claim type** (macro-F_1, %) and evidence-selection F_1. *Conflict*: test-split conflict claims. *Overall* is computed on the pooled test split, not averaged over columns. Ev-F_1 requires the cited modality to match the gold pointer. Δ is taken over the strongest other row in each column; *: $p < 0.01$, paired bootstrap. Overall scores of all 24 systems are in Table 8.

Family	Model	DT	GDP	DTPP	ME	Conflict	Overall	Ev-F_1
Closed-book	PubMedBERT	38.27	71.14	71.66	69.86	36.41	57.31	–
Encoder (+R)	SciBERT	85.64	82.56	81.74	87.74	65.54	85.09	41.20
	PubMedBERT	88.61	83.46	84.84	87.70	62.11	87.20	43.60
	BioLinkBERT	89.59	86.10	82.12	89.31	60.28	87.82	44.80
Specialized	MultiVerS	86.20	83.10	80.40	86.90	58.40	84.63	57.80
	ProgramFC	82.60	81.40	83.60	80.80	61.70	82.18	55.30
RAG	Self-RAG	83.80	82.40	81.90	83.80	59.80	83.16	61.40
	GraphRAG	84.40	85.20	84.70	82.90	63.20	84.27	63.70
LLM (+R)	GPT-4o	78.70	71.80	73.20	77.60	64.90	76.06	58.20
	Claude 4.7	80.40	74.60	76.10	80.20	67.30	78.41	60.50
Reasoning (+R)	DeepSeek-V4	82.70	78.40	80.60	82.10	69.80	81.34	62.10
Ours	**EVIDICT**	**90.74**	**87.51**	**87.46**	**89.94**	**77.94**	**89.21**	**68.40**
Δ *over best baseline*		$+1.15^*$	$+1.41^*$	$+2.62^*$	$+0.63$	$+8.14^*$	$+1.39^*$	$+4.70^*$

only 62.38. The large models are not short of knowledge; they are poor at reading linearized curated records and structural contacts as evidence, which task-specific fine-tuning on the same linearized evidence fixes even at 110M parameters.

5.2 WHERE THE GAIN COMES FROM ON BIOMSVERIFY

The margin is modest on DT and ME ($+1.15$, $+0.63$), larger on the multi-hop DTPP type ($+2.62$), and decisive on Conflict ($+8.14$). Two things about the baselines deserve emphasis. The encoders that are second-best on label F_1 are *worst* at citing evidence (Ev-F_1 41–45): they reach the right label from the wrong items, which is exactly the failure a label-only benchmark cannot see. And the conflict subset reorders the field: the best encoder, BioLinkBERT, falls to 60.28 there, the lowest of the three encoders and below every LLM in the table, while the reasoning LLM becomes the strongest baseline (69.80). Long-form reasoning helps when sources must be weighed against each other, even though the same models trail the fine-tuned encoders on the clean claim types.

The closed-book row is a warning about the benchmark itself, and we state it as such. Without any evidence PubMedBERT scores 38.27 on DT but around 70 on GDP, DTPP and ME. Drug–target counterfactuals substitute an entity from the same ontology neighborhood, which parametric knowledge cannot reject; pathway and mechanism counterfactuals are often implausible on priors alone. Label F_1 on those three types therefore partly measures prior knowledge, which is why we report Ev-F_1 alongside it and why the Conflict column, where the closed-book model drops to 36.41, is the more reliable signal of evidence use (Section A.6).

5.3 ABLATIONS

Each component's effect concentrates where its design says it should be (Table 4). The contradiction score is almost invisible overall (-0.45) and produces the largest single-subset drop in the table on Conflict (-11.63); without it EVIDICT falls to 66.31, *below* the reasoning LLM, so the conflict advantage comes from this term, not a stronger backbone. The structural encoder is worth -2.53 overall but -8.88 on ME. Because this variant also removes AlphaFold3 complexes from the pool, it sees less structural evidence than the BioLinkBERT row of Table 3 (89.31 on ME), so the two are not directly comparable; what the row shows is that EVIDICT's advantage on molecular claims rests on this channel. The PRM is the largest overall contributor (-5.84), and its value-to-go half alone is worth 1.87 points overall and 6.09 on Conflict, which supports factoring grounding from progress: on conflict claims, a step that is grounded in one source but does not move toward a verdict consistent with the others is common, and only the forward term penalizes it. Counterfactual DPO buys little clean accuracy (-0.89) but 4.11 points of robustness on average (Section 5.5). Finally, the

Table 4: **Ablation on BioMSVerify** (macro-F_1, %, mean$_{\pm\text{std}}$ over 5 seeds; each variant retrained). Δ is relative to the full model on Overall. †The conformal layer does not change the argmax label, so macro-F_1 is unchanged by construction; its effect is on realized risk ($0.109 \rightarrow 0.083$, Table 5).

Configuration	Overall	Δ	ME	Conflict
Full model	**89.21**$_{\pm 0.22}$	0.00	**89.94**$_{\pm 0.44}$	**77.94**$_{\pm 0.57}$
w/o generative energy (plain classifier head)	87.56$_{\pm 0.31}$	-1.65	88.62$_{\pm 0.49}$	73.88$_{\pm 0.71}$
w/o process reward model	83.37$_{\pm 0.44}$	-5.84	86.41$_{\pm 0.58}$	67.20$_{\pm 0.94}$
backward-only PRM (no value-to-go)	87.34$_{\pm 0.29}$	-1.87	88.75$_{\pm 0.47}$	71.85$_{\pm 0.76}$
w/o structural encoder + AlphaFold3	86.68$_{\pm 0.35}$	-2.53	81.06$_{\pm 0.82}$	76.42$_{\pm 0.66}$
w/o contradiction score	88.76$_{\pm 0.26}$	-0.45	89.61$_{\pm 0.45}$	66.31$_{\pm 0.88}$
w/o counterfactual DPO	88.32$_{\pm 0.33}$	-0.89	89.10$_{\pm 0.51}$	75.60$_{\pm 0.69}$
BM25 only (no fusion)	88.54$_{\pm 0.28}$	-0.67	89.28$_{\pm 0.46}$	75.10$_{\pm 0.64}$
+ monoT5 reranking	87.92$_{\pm 0.30}$	-1.29	88.83$_{\pm 0.48}$	74.63$_{\pm 0.72}$
w/o conformal layer (temperature scaling)	89.21$_{\pm 0.22}$	$\pm 0.00^{\dagger}$	89.94$_{\pm 0.44}$	77.94$_{\pm 0.57}$

Table 5: **Calibration and selective verification on BioMSVerify** ($\alpha = 0.10$). Cov.: fraction of test claims whose set contains the gold label; Risk $= 1-$Cov. Temperature scaling commits on every claim ($|\mathcal{C}| = 1$), so its Risk is $1-$accuracy. LLM rows use uncalibrated sets from each pass's verbalized label distribution (labels with probability ≥ 0.25), while their macro-F_1 uses the self-consistency vote, so their coverage can fall below their F_1. ‡ECE/Brier of the conformal row are computed on committed claims and are not directly comparable with the other rows.

Method	ECE$\downarrow$	Brier$\downarrow$	Cov.$\uparrow$	$\lvert\mathcal{C}\rvert\downarrow$	Risk$\downarrow$	RC-AUC$\downarrow$
GPT-4o (+R)	0.182$_{\pm 0.009}$	0.311$_{\pm 0.011}$	0.764$_{\pm 0.010}$	1.41$_{\pm 0.04}$	0.236	0.121$_{\pm 0.006}$
DeepSeek-V4 (+R)	0.143$_{\pm 0.008}$	0.268$_{\pm 0.010}$	0.798$_{\pm 0.009}$	1.29$_{\pm 0.03}$	0.202	0.098$_{\pm 0.005}$
EVIDICT, temperature scaling	0.048$_{\pm 0.004}$	0.183$_{\pm 0.006}$	0.891$_{\pm 0.006}$	1.00$_{\pm 0.00}$	0.109	0.038$_{\pm 0.003}$
EVIDICT, conformal	$0.029^{\ddagger}{}_{\pm 0.003}$	$0.177^{\ddagger}{}_{\pm 0.005}$	**0.917**$_{\pm 0.005}$	1.06$_{\pm 0.02}$	**0.083**	**0.035**$_{\pm 0.002}$

retrieval rows are a negative result we want plain: a general-domain reranker *hurts* (-1.29), because it demotes curated and structural records that look nothing like its training passages (Section A.10).

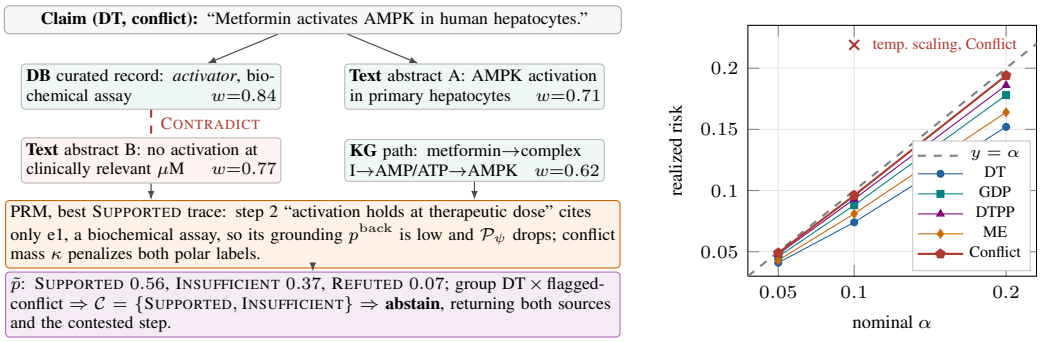
5.4 RISK CONTROL

Only the conformal variant attains Risk $\leq \alpha$ (Table 5), at the cost of abstaining on 5.4% of claims (mean set size 1.06). Temperature scaling is the right comparison to isolate the mechanism: it shares the model and the probabilities, has the lowest ECE among the rows without conformal calibration, and still realizes 0.109 overall. Its failure is not uniform, which is the point of conditioning on groups. Committing on every claim is within target on DT (0.093) but not on GDP or DTPP (0.125), and on Conflict the error is 0.219, more than twice the target (Figure 2b; Table 10). The conformal layer holds every claim type and the conflict subset at or below α at $\alpha \in \{0.05, 0.10, 0.20\}$ by abstaining where the risk is: 18.6% on Conflict at $\alpha = 0.10$ against 3.9% on DT.

Two features of these numbers should temper how they are read. The guarantee is conservative where it is not needed: DT abstains on 3.9% of claims although committing on all of them would already meet the target, which is the price of a per-group finite-sample correction on a calibration half of 772 claims split twelve ways. And the conflict result is empirical rather than guaranteed. The formal statement covers the *flagged-conflict* group, which is defined by the detector, and it is the overlap of that group with the gold conflict label that makes the conflict subset behave. Where the two diverge, as for a conflict the detector misses, the claim is calibrated with a non-conflict group and inherits its smaller threshold (Section A.2; Section A.13 works through one such claim).

5.5 ROBUSTNESS AND FURTHER RESULTS

Robustness to counterfactual perturbations. Table 6 applies each of the five operators to the test claims. EVIDICT scores highest under every operator (93.96 on average against 72.92 for DeepSeek-V4), and the gap is widest where reading a structured record is unavoidable. Relation inversion and



(a) Illustrative walk-through of a conflict claim (schematic values). (b) Realized risk vs. nominal α (Table 10).

Figure 2: **Risk control in practice.** (a) On a claim whose curated record and a later abstract disagree under different assays, the contradiction score and the poorly grounded PRM step move mass toward INSUFFICIENT, and the flagged-conflict calibration group's larger threshold returns a two-label set, so EVIDICT abstains and returns the contested evidence rather than a confident verdict. The example is schematic and its numbers are for exposition. (b) Realized risk stays at or below $y = \alpha$ for every claim type and for the conflict subset at all three levels; committing on temperature-scaled scores gives 0.219 on Conflict at $\alpha = 0.10$ (cross).

Table 6: **Robustness on BioMSVerify.** Each slice pairs every applicable supported test claim with its perturbed version, so both polar labels are present; cells are macro-F_1 (%). The operators are the same families used to build training negatives.

Method	EntSub	RelInv	PathSub	ResCorr	AffPert	Avg.
GPT-4o (+R)	71.60	58.30	62.40	81.70	54.80	65.76
DeepSeek-V4 (+R)	78.10	66.90	69.20	87.00	63.40	72.92
EVIDICT w/o DPO	93.49	90.44	76.92	97.86	90.52	89.85
EVIDICT	**98.82**$_{\pm 0.24}$	**93.38**$_{\pm 0.37}$	**85.71**$_{\pm 0.61}$	**99.24**$_{\pm 0.18}$	**92.67**$_{\pm 0.42}$	**93.96**$_{\pm 0.31}$

affinity perturbation, which require a direction or a magnitude, cost GPT-4o 13–17 points relative to entity substitution, while EVIDICT loses 5–6. Counterfactual DPO is responsible for 4.11 points on average and 8.79 on pathway substitution, against 0.89 on clean claims, so a component that looks marginal in Table 4 carries robustness. These are not scores against novel attacks: the operators are the families used to build training negatives, and pathway substitution, where the substitute often stays plausible from text alone, is EVIDICT's weakest slice (85.71).

Further results. *Retrieval* (Section A.10): dev-tuned fusion beats every single retriever on every modality (Recall@10 63.9 vs. 59.7 for BM25), whereas untuned equal-weight fusion is *worse* than BM25 (56.0), because the learned retrievers are nearly blind to structural evidence (8.7, 2.7). *Efficiency* (Section A.12): the full test split costs \$0.51 against \$41.20 for GPT-4o, at 0.56 s per claim on one A100. *Variability* (Table 9): five seeds span 88.94–89.47 on BioMSVerify.

6 CONCLUSION AND OUTLOOK

We argued that biomedical claim verification is adjudication across sources that can legitimately disagree, and built BioMSVerify and EVIDICT around that view. The components that matter most are the ones a single-corpus benchmark cannot reward: a contradiction score worth 11.63 points on conflicting claims but 0.45 overall, a structural channel worth 8.88 points on molecular explanation claims, and a conformal layer that changes no label yet holds realized risk at or below α for every claim type and for the conflict subset. Two findings reach beyond our system: frontier LLMs given the same evidence pool trail 110M encoders fine-tuned on it, and a well-calibrated score still commits wrongly on 21.9% of conflict claims. Improving the contradiction detector's recall, prior-matched counterfactuals, and encoders for tables and figures are the natural next steps. When sources disagree, the useful output is a verdict or an abstention with the contested evidence attached.

AI USE STATEMENT

This work uses AI models in two ways. First, API-based AI model calls are part of the proposed verification methodology and are used within the system pipeline and evaluation protocol, as described in Section 4 and Section 5. Second, AI tools were used for language polishing, including improving grammar, clarity, concision, and consistency of wording.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. BioMSVerify is derived from public biomedical resources, listed in Section A.3 and Section A.5, and the work involves no new human-subject data, private medical records, or personally identifiable information. Because biomedical verification systems may inform scientific, clinical, or pharmacovigilance decisions, outputs should be treated as decision support rather than medical advice, with expert review and provenance inspection, especially when sources conflict or the system abstains.

REPRODUCIBILITY STATEMENT

We submit the code as part of the supplementary materials. Benchmark composition, claim types, and the conflict definition are given in Section 3 and Section A.5; the method, retrieval sources, encoders, training objective, and conformal procedure in Section 4 and Section A.3; the proof of Proposition 1 in Section A.1; and hyperparameters, trainable-parameter counts, and training cost in Section A.4. The evaluation protocol, including the five seeds, the dev split into selection and calibration halves, the significance test, and the Ev-F_1 and risk definitions, is given in Section 5. Full results for all 24 systems, seed variability, group-conditional risk, retrieval diagnostics, and efficiency measurements are in Section A.8 to Section A.12.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anthropic. Claude opus 4.7 system card. Technical report, Anthropic, April 2026. URL `https://www.anthropic.com/claude-opus-4-7-system-card`.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, et al. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1), 2022.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3615–3620, 2019.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 632–642, 2015.

Annalisa Buniello, Daniel Suveges, Carlos Cruz-Castillo, Manuel Bernal Llinares, Helena Cornu, Irene Lopez, Kirill Tsukanov, Juan María Roldán-Romero, Chintan Mehta, Luca Fumis, et al. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic acids research*, 53(D1):D1467–D1475, 2025.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2009.

Allan Peter Davis, Thomas C Wiegers, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database (ctd): update 2023. *Nucleic acids research*, 51(D1):D1257–D1262, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2288–2292, 2021.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33: 1970–1981, 2020.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Google DeepMind. Gemini 3.1 pro model card. Technical report, Google DeepMind, February 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *arXiv preprint arXiv:2504.16828*, 2025.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, 2020.

Klaus Krippendorff. Content analysis. In Erik Barnouw, George Gerbner, Wilbur Schramm, Theodore L. Worth, and Larry Gross (eds.), *International Encyclopedia of Communication*, volume 1, pp. 403–407. Oxford University Press, New York, 1989.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pp. 5848–5864, 2024.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, volume 2024, pp. 39282–39309, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4228–4238, 2021.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2): 302–309, 2017.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678, 2024.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025. URL `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Luu Anh Tuan, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 6981–7004, 2023.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2116–2129, 2021.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 3499–3512, 2021.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2019.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 7534–7550, 2020.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the association for computational linguistics: NAACL 2022*, pp. 61–76, 2022.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. Sciver: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8562–8579, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8003–8016, 2022.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, volume 2025, pp. 12476–12505, 2025.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023.

A Appendix

A.1 Proof of Proposition 1

Fix a group g with calibration scores $\nu_1, \ldots, \nu_{n_g}$ and a test claim with $g(c, E) = g$ and true-label score $\nu_{n_g+1} = \nu(c, E, y^*)$. Because $g(\cdot)$ does not depend on y, conditioning on g preserves exchangeability of the $n_g + 1$ scores, so the rank of ν_{n_g+1} is uniform on $\{1, \ldots, n_g + 1\}$ (ties broken at random). Hence $\Pr[\nu_{n_g+1} \le \hat{q}_g \mid g] \ge \lceil (n_g + 1)(1 - \alpha) \rceil / (n_g + 1) \ge 1 - \alpha$, and $y^* \in \mathcal{C}$ is exactly this event. If the index exceeds n_g, $\hat{q}_g = \infty$ and coverage is 1. A wrong commitment requires $|\mathcal{C}| = 1$ and $y^* \notin \mathcal{C}$, so its probability is at most the miscoverage. For a union $U = \cup_k g_k$, $\Pr[y^* \notin \mathcal{C} \mid U] = \sum_k \Pr[g_k \mid U] \Pr[y^* \notin \mathcal{C} \mid g_k] \le \alpha$. □

The bound is marginal over the calibration draw. With n_g calibration points the realized coverage of a fixed calibration set is Beta-distributed around $1 - \alpha$ with spread $O(n_g^{-1/2})$, which is why small groups show both conservative sets and occasional near-α realized risk.

A.2 SOURCE PROFILES

The source profile of (c, E) is computed after retrieval and before the verdict: *flagged-conflict* if the conflict mass $\kappa(c, E)$ of Equation (2) exceeds a threshold tuned on the selection half; otherwise *concordant multi-source* if the retained evidence spans at least two modalities with total reliability above a second threshold; otherwise *single-modality-dominant*. Neither threshold uses the gold label of the test claim. The gold conflict label enters only as an evaluation subset. A conflict that the detector fails to flag is therefore calibrated with a non-conflict group whose threshold is smaller; that is the mechanism by which conflict-level risk can exceed α although the guarantee holds for every group, and it is why the detector's recall is the operative quantity for deployments that care about conflicts specifically.

A.3 METHOD DETAILS

Retrieval sources. Curated records come from DrugBank, ChEMBL, Open Targets, DisGeNET, CTD, UniProt, STRING, Reactome, KEGG and GO (Knox et al., 2024; Zdrazil et al., 2024; Buniello et al., 2025; Piñero et al., 2020; Davis et al., 2023; Bateman et al., 2022; Szklarczyk et al., 2023; Milacic et al., 2024; Kanehisa & Goto, 2000; Ashburner et al., 2000); diseases are linked through UMLS (Bodenreider, 2004). Structural evidence uses PDB, PDBbind with the CASF protocol, and BioLiP (Berman et al., 2000; Wang et al., 2004; Liu et al., 2017; Su et al., 2019; Yang et al., 2012). The late-interaction score is $\sum_{u \in c} \max_{v \in p} f(u)^\top f(v)$; fusion is $\mathrm{RRF}(p) = \sum_r \omega_r / (k_{\mathrm{rrf}} + \mathrm{rank}_r(p))$ over BM25, SPLADE and ColBERTv2 with $k_{\mathrm{rrf}} = 60$ and dev-tuned weights $\omega_{\mathrm{BM25}} : \omega_{\mathrm{SPLADE}} : \omega_{\mathrm{ColBERT}} = 1.8 : 1.2 : 1$. **Encoders.** Item embeddings are pooled by modality-aware attention $\mathbf{u}(c, E) = \sum_i \eta_i \mathbf{h}_i$, $\eta_i \propto \exp(\mathbf{q}_c^\top W_{m(i)} \mathbf{h}_i)$, and each $\mathbf{h}_i$ is also projected to one soft token. EquiformerV2 is equivariant; its pooled scalar channels are invariant to rigid motions of the complex. Protein structural context uses geometric pretraining features (Zhang et al., 2022); earlier equivariant networks (Satorras et al., 2021; Fuchs et al., 2020) were not competitive in preliminary runs. **Inference.** For each y the backbone samples $N = 8$ traces at temperature 0.7, and $\mathcal{P}_\psi(c, E, y)$ takes the best-scoring one. The label is $\arg\max_y s_y$. With a committed verdict, the returned trace is the best-scoring trace for that label; on abstention, EVIDICT returns the best-scoring trace for each label in $\mathcal{C}$, which exposes the contested step.

A.4 IMPLEMENTATION

Text and linearized records use PubMedBERT-base (110M); KG evidence a 2-layer R-GCN (12M); protein sequences frozen ESM-2 650M; complexes EquiformerV2 (31M, $L_{\max} = 6$) with Uni-Mol ligand features (48M). The generative energy head and the PRM are two LoRA adapters ($r = 16$, $\alpha_{\mathrm{LoRA}} = 32$, dropout 0.05) over a 4-bit NF4 BioMistral-7B (QLoRA). Trainable parameters: 201.0M in the encoders plus 77.2M in the two adapters, 278.2M in total. AdamW, lr 2×10^{-5}, cosine schedule with 6% warmup, batch 32, 6 epochs, gradient clipping 1.0, bf16. Loss weights $\lambda_{\mathrm{nce}} = 0.3$, $\lambda_{\mathrm{rank}} = 0.5$, $\lambda_{\mathrm{dpo}} = 0.2$, $\lambda_{\mathrm{contr}} = 0.4$, $\lambda_{\mathrm{prm}} = 0.6$, margin $\gamma = 0.5$, DPO $\beta = 0.1$; fusion weights $\lambda_c = 0.35$, $\lambda_p = 0.45$ tuned on the selection half. Retrieval keeps top-50 per source before fusion and top-8 after. One training run takes 14.2 GPU-hours in total on $8\times$A100 80GB (about 1.8 hours wall-clock).

A.5 DATASETS AND BENCHMARK CONSTRUCTION

Construction. SUPPORTED claims instantiate curated relations as natural-language templates paraphrased by an LLM and checked by annotators. REFUTED claims apply one of the five operators of Section 4.5 to a supported claim (affinity perturbation under non-comparable conditions yields INSUFFICIENT instead); further INSUFFICIENT claims corrupt a qualifier (organism, cell type, assay) so that no retrieved record addresses the stated condition. Gold evidence pointers are the records from which the claim was built plus annotator-added items; a claim is *conflict* when two gold pointers from different sources disagree on the relation or its direction under the claim's qualifiers. Splits are made at the level of the underlying relation, so no relation appears in two splits in different surface forms.

Table 7: **Dataset statistics.** Sources: T text, D curated database, G knowledge graph, S molecular structure, Tab tables, F figures. SciVer is evaluated zero-shot. Macro-F_1 is computed over each benchmark's native label set.

Dataset	#Train	#Dev	#Test	#Conf.	Sources
SciFact	809	300	300	–	T
HealthVer	10,591	1,884	1,855	–	T
PubHealth	9,466	1,183	1,183	–	T
COVID-Fact	3,263	400	423	–	T
SciVer	–	–	3,000	–	T, Tab, F
BioMSVerify	6,794	1,544	1,868	813	T, D, G, S

A.6 Known benchmark artifacts

The closed-book row of Table 3 is the clearest diagnostic of what BioMSVerify measures. On DT, closed-book PubMedBERT is near chance (38.27), because entity substitution within an ontology neighborhood produces counterfactuals that are as plausible a priori as the originals. On GDP, DTPP and ME it reaches about 70 to 72: pathway substitution and mechanism corruption often produce statements that a model with biomedical priors can reject without evidence. We therefore regard label F_1 on those three types as an upper bound on evidence-grounded ability, and recommend that future work on BioMSVerify report Ev-F_1 and the conflict subset alongside it. Making counterfactuals prior-matched, by sampling substitutes conditioned on a closed-book model's confidence, is the fix we would apply in a second release.

A.7 Baselines

Trivial: majority class and a lexical-overlap classifier. *Closed-book*: SciBERT, PubMedBERT and BioLinkBERT fine-tuned on claim→label (Beltagy et al., 2019; Gu et al., 2021; Yasunaga et al., 2022). *Encoder (+R)*: the same encoders fine-tuned on the claim concatenated with the linearized evidence pool. *Specialized*: MultiVerS (Wadden et al., 2022), a DeBERTaV3-large NLI retrieve-then-verify pipeline (He et al., 2021), and ProgramFC (Pan et al., 2023). *RAG*: RAG, Atlas, Self-RAG, GraphRAG and HippoRAG (Lewis et al., 2020; Izacard et al., 2023; Asai et al., 2024; Edge et al., 2024; Gutiérrez et al., 2024), each restricted to the controlled pool. *LLM (+R)* and *Reasoning (+R)*: BioMistral-7B (prompted, not fine-tuned, to isolate the effect of our training), Qwen2.5-72B, GPT-4o, Claude 4.7, Gemini 3.1 Pro, o3 and DeepSeek-V4, each with chain-of-thought, five-sample self-consistency and Chain-of-Verification; closed-source model versions and access dates are recorded with the released prompts.

A.8 Full main results and seed variability

Table 8: **Main results, all** 24 **systems** (macro-F_1, %, mean over 5 seeds). Table 2 is a subset.

Family	Model	SciFact	HealthVer	PubHealth	COVID-Fact	SciVer	BioMSVerify	Avg.
Trivial	Majority class	24.10	22.80	19.40	38.20	21.70	17.10	23.88
	Lexical overlap	31.60	29.40	24.80	45.30	26.20	17.24	29.09
Closed-book	SciBERT	42.30	44.60	51.70	57.40	33.80	53.42	47.20
	PubMedBERT	45.10	47.20	53.40	59.10	35.20	57.31	49.55
	BioLinkBERT	43.80	45.90	52.60	58.30	34.50	53.80	48.15
Encoder (+R)	SciBERT	74.60	71.80	66.90	72.40	52.80	85.09	70.60
	PubMedBERT	76.90	73.50	68.20	74.10	54.30	87.20	72.37
	BioLinkBERT	77.80	74.60	68.90	74.80	55.10	87.82	73.17
Specialized	MultiVerS	80.10	76.30	70.20	77.50	56.40	84.63	74.19
	DeBERTaV3 pipeline	78.40	77.10	71.40	78.20	57.90	85.41	74.74
	ProgramFC	76.20	73.90	69.70	75.80	58.60	82.18	72.73
RAG	RAG	69.80	68.40	63.10	70.20	51.30	77.84	66.77
	Atlas	72.50	70.90	65.80	72.60	53.70	79.52	69.17
	Self-RAG	76.80	74.20	69.40	76.10	59.20	83.16	73.14
	GraphRAG	75.40	73.10	68.60	74.90	58.10	84.27	72.40
	HippoRAG	76.10	73.80	68.90	75.40	58.80	83.75	72.79
LLM (+R)	BioMistral-7B (prompted)	63.70	62.10	55.40	64.80	42.60	62.38	58.50
	Qwen2.5-72B	78.20	75.60	70.10	77.30	63.40	74.16	73.13
	GPT-4o	79.60	77.20	71.80	79.10	66.80	76.06	75.09
	Claude 4.7	81.30	78.90	73.20	80.40	71.50	78.41	77.29
	Gemini 3.1 Pro	80.70	78.10	72.60	79.80	**76.40**	77.31	77.49
Reasoning (+R)	o3	82.40	79.80	73.90	81.20	74.60	80.27	78.70
	DeepSeek-V4	83.10	80.50	74.10	81.70	75.20	81.34	79.32
Ours	**EVIDICT**	**84.70***	**82.30***	**75.40***	**83.20***	74.80	**89.21***	**81.60***

Table 9: **Seed-level variability** (macro-F_1, %, mean$_{\pm\text{std}}$). BioLinkBERT (+R) and EVIDICT vary over five training seeds; DeepSeek-V4 over five API passes with permuted evidence order.

Model	SciFact	HealthVer	PubHealth	COVID-Fact	SciVer	BioMSVerify
BioLinkBERT (+R)	$77.80_{\pm 0.54}$	$74.60_{\pm 0.47}$	$68.90_{\pm 0.39}$	$74.80_{\pm 0.62}$	$55.10_{\pm 0.71}$	$87.82_{\pm 0.35}$
DeepSeek-V4 (+R)	$83.10_{\pm 0.41}$	$80.50_{\pm 0.36}$	$74.10_{\pm 0.44}$	$81.70_{\pm 0.52}$	$75.20_{\pm 0.58}$	$81.34_{\pm 0.49}$
EVIDICT	$\mathbf{84.70}_{\pm 0.31}$	$\mathbf{82.30}_{\pm 0.28}$	$\mathbf{75.40}_{\pm 0.33}$	$\mathbf{83.20}_{\pm 0.40}$	$74.80_{\pm 0.46}$	$\mathbf{89.21}_{\pm 0.22}$
Per-seed BioMSVerify overall for EVIDICT*:*	13: 89.02	21: 88.94	42: 89.47	87: 89.33	2026: 89.29	(mean 89.21, sd 0.22)

A.9 GROUP-CONDITIONAL RISK CONTROL

Table 10: **Group-conditional conformal results on BioMSVerify.** Risk = 1−Cov.; Abst.: abstention rate. *All* is the test-weighted aggregate of the four claim types, which partition the test split; Conflict overlaps them. The last row commits on every claim with temperature-scaled scores.

α	Quantity	DT	GDP	DTPP	ME	Conflict	All
0.05	Risk	0.041	0.046	0.048	0.043	0.049	0.044
	Abst.	0.086	0.121	0.157	0.104	0.312	0.112
	$\lvert\mathcal{C}\rvert$	1.11	1.16	1.21	1.13	1.38	1.15
0.10	Risk	0.074	0.088	0.093	0.081	0.096	0.083
	Abst.	0.039	0.058	0.081	0.047	0.186	0.054
	$\lvert\mathcal{C}\rvert$	1.04	1.07	1.10	1.05	1.23	1.06
0.20	Risk	0.152	0.178	0.186	0.164	0.194	0.167
	Abst.	0.008	0.014	0.022	0.011	0.067	0.013
	$\lvert\mathcal{C}\rvert$	1.01	1.01	1.02	1.01	1.07	1.01
Temp. scaling (always commit), Risk		0.093	0.125	0.125	0.101	0.219	0.109

Three regularities. Abstention on Conflict is roughly three to five times the test-split aggregate at every α (0.312 vs. 0.112, 0.186 vs. 0.054, 0.067 vs. 0.013), which is where the flagged-conflict groups' larger thresholds act. At $\alpha = 0.05$ realized risk on Conflict (0.049) is closest to its nominal level of any cell, consistent with conflict claims being the hardest and the smallest calibration groups having the least slack. And the always-commit row separates groups the model already handles (DT, within target without abstention) from groups it does not; the conformal layer spends abstentions on DT anyway, which is the conservatism of a per-group finite-sample correction rather than a need of the data.

A.10 RETRIEVAL ANALYSIS

Table 11: **Evidence Recall@10 (%) on BioMSVerify by modality.** Fusion weights are BM25:SPLADE:ColBERTv2; 1:1:1 is the untuned equal-weight default. Avg. is the unweighted mean over modalities.

Retriever	Text	Database	KG	Structure	Avg.
BM25	94.7	40.0	38.7	65.3	59.7
SPLADE	63.3	44.7	37.3	8.7	38.5
ColBERTv2 (MedCPT)	57.3	40.1	18.3	2.7	29.6
Hybrid RRF (1:1:1, untuned)	90.7	42.8	37.3	53.3	56.0
+ monoT5 reranking	87.5	36.8	29.3	48.0	50.4
Ours (tuned RRF, 1.8:1.2:1)	**95.1**	**48.6**	**43.2**	**68.7**	**63.9**

Tuned fusion beats every single retriever on every modality, with the largest gains over the best single retriever on KG (+4.5) and curated records (+3.9). Two negative results matter more. Untuned equal-weight fusion is worse than BM25 alone (56.0 vs. 59.7): the learned retrievers are nearly blind to structural evidence, whose identifiers and residue strings resemble nothing in their training data, and equal weighting lets them dilute BM25's exact-match strength; the tuned weights restore BM25's lead. And a general-domain reranker lowers recall on every modality (56.0 → 50.4), most on KG (−8.0), curated records (−6.0) and structure (−5.3), consistent with its −1.29 end-task cost in Table 4. Dense text retrieval is not the right tool for identifier-heavy biomedical records, and a system that fuses it naively pays for it.

A.11 ROBUSTNESS TO COUNTERFACTUAL PERTURBATIONS

Full-slice results appear in Table 6 in the main text.

Relation inversion and affinity perturbation are the hardest operators for the LLMs (GPT-4o 58.30 and 54.80): both require reading a direction or a magnitude out of a structured record rather than matching text, which is the same weakness that places every LLM below every fine-tuned encoder on BioMSVerify. DPO contributes 4.11 points on average and 8.79 on pathway substitution, far more than its 0.89 on clean claims. EVIDICT's weakest slice is also pathway substitution (85.71), where the substituted pathway is often still plausible given text alone. The scores are high partly because these operator families are the ones used to create training negatives; the table shows that EVIDICT learned to resist them, not that it resists perturbations it has never seen.

A.12 EFFICIENCY

Table 12: **Efficiency on the BioMSVerify test split** ($1{,}868$ claims). Latency is end-to-end per claim including retrieval, on one A100 80GB for EVIDICT and through public APIs (batch size 1, median of 3 passes) for the LLMs; the settings are not hardware-matched. Cost is inference only, in USD for the full split.

Method	Train. params	Peak mem.	Lat. (s) ↓	Cost ↓	F_1 ↑
GPT-4o (+R)	–	–	7.90	$41.20	76.06
DeepSeek-V4 (+R)	–	–	24.30	$12.80	81.34
EVIDICT	278.2M	31.4 GB	**0.56**	**$0.51**	**89.21**

The full split costs $\$0.51$ for EVIDICT, about $1/80$ of GPT-4o and $1/25$ of DeepSeek-V4; $1{,}868 \times 0.56$ s is 0.29 A100-hours, which is consistent with that cost at typical on-demand pricing (about $\$1.8$ per GPU-hour). The comparison excludes the one-time 14.2 GPU-hours of training, which at the same price is roughly $\$25$ and is recovered within a single test-split-sized workload against GPT-4o. Latency ratios ($14\times$ and $43\times$) mix local batched inference with rate-limited APIs and should be read as a deployment observation rather than a controlled comparison.

A.13 CASE STUDY: A CONFLICT THE DETECTOR MISSES

Figure 2a shows the intended behavior on a flagged conflict: EVIDICT abstains and returns the contested evidence. Here we describe a drug–target conflict claim on which the same system commits to a wrong label. The case isolates the limit of Proposition 1 discussed in Section 5.4 and Section A.2. The run uses $\alpha = 0.10$. On its calibration draw, the threshold is $\hat{q}_g = 0.55$ for the DT×concordant multi-source group and 0.78 for the DT×flagged-conflict group, and a claim is flagged when $\kappa(c, E) > 0.40$.

Claim and evidence. The claim is "Resveratrol directly activates SIRT1", with gold label INSUFFICIENT and the conflict mark. Retrieval returns three relevant items. A curated record lists resveratrol as a SIRT1 activator on the basis of a fluorophore-tagged peptide assay ($w = 0.81$). Abstract A reports activation in the same assay ($w = 0.72$). Abstract B reports no activation with native, untagged substrates and attributes the earlier effect to the fluorophore ($w = 0.75$). The sources disagree under assay conditions that are not comparable, so the gold label is INSUFFICIENT under the construction rule of Section 3.

What EVIDICT does. The contradiction model labels both the (record, B) pair and the (A, B) pair as CONTRADICT with $p_\phi \approx 0.9$. Because each of these pairs spans different assays, its pair weight is small ($w_{ij} \approx 0.05$). The two pairs therefore contribute about $2 \times 0.05 \times \log 10 \approx 0.23$ to the conflict mass of Equation (2), and the concordant (record, A) pair adds about 0.02. The resulting $\kappa = 0.25$ is below the flagging threshold, so the claim is calibrated with the DT×concordant group. As a single-relation claim it yields a one-step trace (Section 4.4). The best SUPPORTED step cites the curated record and abstract A and is well grounded ($p^{\text{back}} = 0.86$). Because the step never cites abstract B, neither the backward term nor the forward term of the PRM penalizes it. The fused probabilities are 0.71 for SUPPORTED, 0.24 for INSUFFICIENT and 0.05 for REFUTED. Since $\nu(\text{INSUFFICIENT}) = 0.76 > 0.55$, the prediction set is $\mathcal{C} = \{\text{SUPPORTED}\}$, and EVIDICT commits to a wrong label.

What the case shows. The error lies in the group assignment, not in the conformal layer. Had the claim been flagged, the threshold of 0.78 would have admitted INSUFFICIENT ($0.76 \leq 0.78$). The

set would then have been {SUPPORTED, INSUFFICIENT}, and EVIDICT would have abstained as in Figure 2a. Proposition 1 still holds for the group the claim was assigned to, but the gold conflict is invisible to calibration. The miss has a specific cause. Down-weighting contradictions between assay-mismatched sources keeps EVIDICT from abstaining when two sources are both correct under different conditions. The same down-weighting lets through conflicts in which one of the assays is itself an artifact. Misses of this kind are one reason the realized risk on Conflict (0.096) sits closer to α than on DT (0.074) (Table 10). The returned trace keeps the assay qualifier of every cited item, so a curator can still see the discrepancy. Raising the detector's recall on assay-mediated conflicts is the most direct fix (Section 6).