XXXII Symposium on Bioinformatics and
Computer-Aided Drug Discovery (BCADD-2026)

Retrieval-augmented LLM systems for biomedical question answering and prediction interpretation

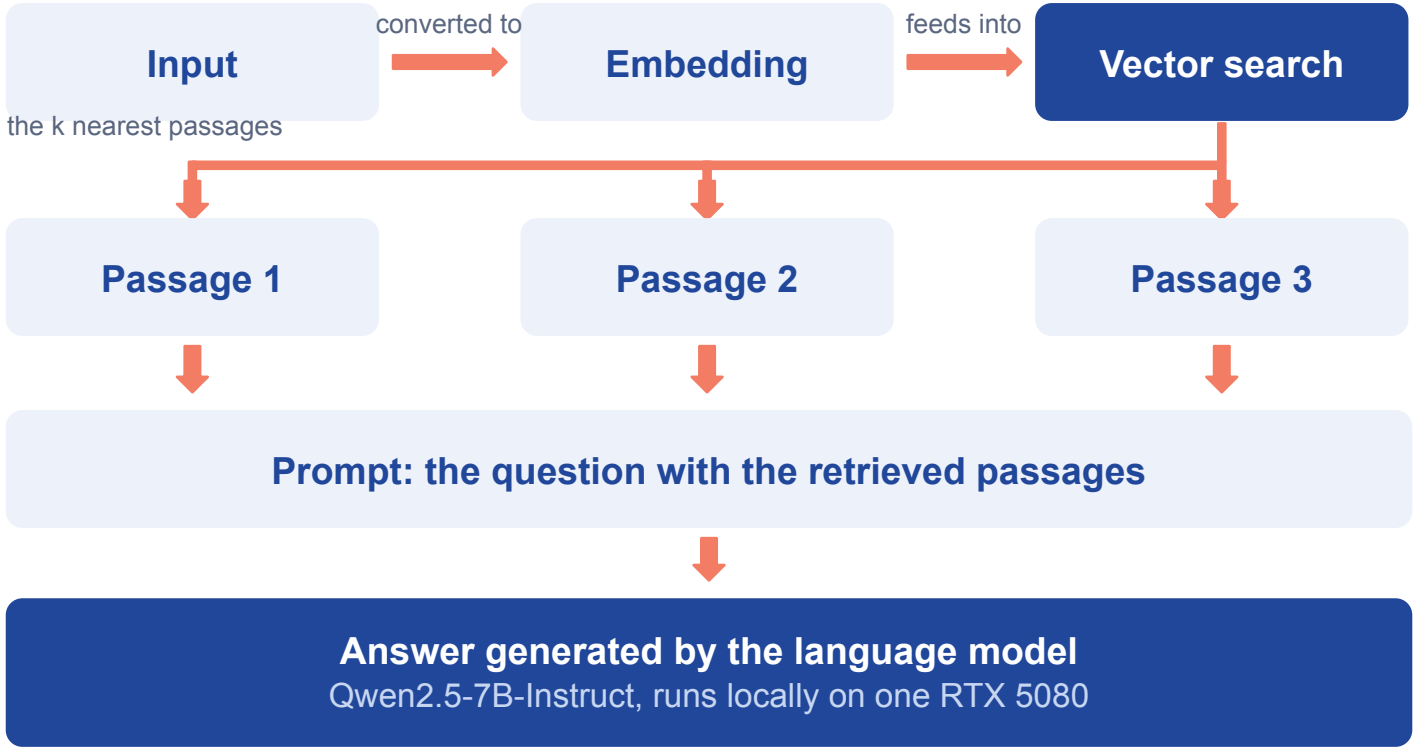
R. Taktashov, N. Biziukova, A. Dmitriev, O. Tarasova

Institute of Biomedical Chemistry, Moscow

MedHopper: Database, 2026, baag047

Russian Science Foundation, project No. 25-15-00300

Retrieval-augmented generation, and where it stops



Where this stops

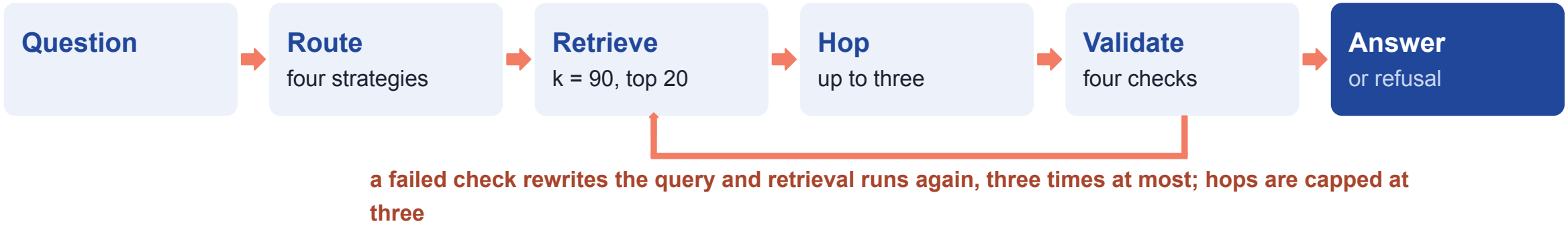
- One retrieval, one answer: a question whose answer needs a second search has no second step.
- Similarity returns passages on the topic, and the topic may not contain the answer.
- Nobody reads the draft: no grounding check, no stopping rule, no refusal.

Two answers: hybrid retrieval, where BM25 lexical scores and vector similarity are merged and reranked by a cross-encoder; and a controller around the loop that decides what to retrieve next, when to stop and when to refuse.

To which class does the drug belong that acts on the protein product of the gene mutated in cystinosis? multi-hop question

MedHopper: a bounded controller

Agency here is a fixed, enumerable set of decisions: the controller picks among them and the model fills the slots.



0.55 MedHopper

0.43 DecompRAG

0.59 GPT-4o, no RAG

Exact match, MedHopQA test set, N = 1000, official CodaBench evaluation.

SD 0.000
five seeds, κ = 1.0, agreement 100%

1 GPU
Qwen2.5-7B-Instruct on an RTX 5080, no cloud call

Component removed	EM	Δ
Full pipeline	**0.55**	
Single pass, no multi-hop	**0.50**	−0.05
Reranking top-N = 10	**0.51**	−0.04
Repair loop off	**0.52**	−0.03
Reranking off	**0.52**	−0.03
Any validator or Orphanet off	**0.54**	−0.01

The gain is spread across routing, reranking, bounded hopping and repair: there is no single component to point at.

Drug-drug interaction prediction interpretation ≡ multi-hop question

1. **Pair of structures**
   including compounds never synthesised
2. **Prediction models**
   mechanism, enzymes, adverse reactions, severity
3. **Evidence, round after round**
   one query per predicted claim
4. **Interpretation or hypotheses**
   with the status of every claim

Why the question is multi-hop

- A prediction arrives as a dozen tables, and each row is a separate claim.
- The evidence for one claim sits in several papers and handbooks, and often only part of it is there.
- Finding it takes several rounds of querying, with each round shaped by what the last one returned.

The state machine comes from MedHopper

- The same bounded controller: route, retrieve, hop, validate, answer.
- Hops are the rounds of evidence gathering, and the validators check that each claim is grounded.
- The route stays traceable, which matters when a report is later questioned.

ReAct inside the nodes, MCP to the tools

- Inside a node the model reasons, calls a tool and reads the result, until the local goal of that node is met [Yao et al., 2022].
- MCP connects PubMed, PubChem, DrugBank and the web search API through one interface [Hou et al., 2025].

Two tiers of evidence, four ways a claim can end

Curated local corpus — tier 1

161 handbooks, monographs and textbooks on interactions
and metabolism; membership is fixed and a citation resolves to a page

Web search — tier 2

current articles, labels and guidelines, reached through MCP when the corpus is silent; domains are validated

confirmed the source is quoted and cited

contradicted the source says otherwise, and that is shown

undocumented predicted, with nothing in the literature either way

insufficient a ranked list of biochemical hypotheses, each marked as unconfirmed

The curated corpus decides

- It is curated by hand, so it outranks the web.
- Where the two disagree, the corpus decides and the disagreement is displayed.
- The web keeps the system current without re-indexing the store.

Silence in the literature is a result

- A claim the models predict and the literature has never recorded is a candidate for study, and the report says so.
- An effect the literature describes and the models reject measures the limits of the predictor.
- Either way the asymmetry is printed instead of being smoothed away.

Drug-drug interaction prediction example (Nicotine + Melatonin)

Predicted by the models (way2drug.com/ddi/)

Severity	class 1, contraindicated: +0.820 five-class, +0.793 three-class
Risk	major, Pa − Pi +0.326
Mechanism	additive pharmacodynamics 0.793, altered metabolism 0.643
Cytochromes	CYP1A2 +0.114, weak in PASS terms
Reliability	class 1 in three-class carries the highest validation accuracy, IAP 0.876

What the sources say

Nicotine
no induction of CYP1A2

Tobacco smoke
its PAHs induce CYP1A2

↓

CYP1A2 activity rises
the smoke carries the induction, the modelled molecule does not

↓

Melatonin clearance rises
melatonin is cleared mainly by CYP1A2, so its plasma level falls

What three rounds of search add to the prediction

- One pair of structures gives one prediction, and the prediction holds: the combination is documented as major and contraindicated.
- The evidence rounds name the confounder. The induction belongs to tobacco smoke; nicotine on its own leaves CYP1A2 untouched. The report therefore ties the prediction to smoking and marks nicotine replacement therapy as a separate case, which the structures alone could never show.

Conclusions

The control structure carries the result

- 0.55 against 0.43 with the same 7B model and the same corpus.
- The ablation spreads the gain across routing, reranking, bounded hopping and repair.

Prediction interpretation has the same shape

- Scattered and partial literature, several rounds of search and validation.
- State machine from MedHopper, ReAct inside the nodes, MCP to the tools.

The output is graded

- Confirmed, contradicted, undocumented, or a ranked set of biochemical hypotheses.
- Most PASS-based services fit, including PRIS.

Repository | github.com/RustT883/MedHopper
Correspondence | rustam_taktashov@ibmc.msk.ru | rustam_taktashov@mail.ru
Poster | https://way2drug.com/bcadd2026/content/posters/ysc/Taktashov_R.pdf
Abstract | https://way2drug.com/bcadd2026/content/posters/ysc/Taktashov_R_abstract.pdf

1. Islamaj R. et al. MedHopQA. Database, 2026, baag057.
2. Taktashov R. R. et al. MedHopper. Database, 2026, baag047.
3. Islamaj R. et al. Overview of MedHopQA, 2026, baag046

Thank you!

Backup slides

Backup: the two corpora

MedHopper index

- About 225 000 health-related Wikipedia pages, selected with PETScan.
- Chunks of 512 characters with 32 characters of overlap.
- MedEmbed-small-v0.1 in a local Chroma store, dense retrieval only.
- Recall@20 up to 0.778 on the development set, N = 45.

Curated corpus for interactions

- 161 handbooks, monographs and textbooks on interactions and metabolism.
- PDF converted to Markdown with OCR, then split recursively and semantically.
- Chroma for the dense side and a BM25S index for the lexical side.
- Membership is fixed, so every citation resolves to a page.

Generator Qwen2.5-7B-Instruct, FlashRank cross-encoder, one RTX 5080 under 12 GB, 15 to 20 seconds per question.

Backup: hybrid search and reranking

Lexical side

- BM25 scores a passage by how often a term occurs in it and how rare that term is across the collection.
- It finds the exact token: a gene symbol, a locus, a trade name.
- Paraphrase escapes it, because it has no notion of meaning.

Dense side

- Query and passage are embedded and compared by cosine similarity.
- It finds the paraphrase and the synonym.
- It also returns passages that are on the topic and silent on the answer.

Reranking

- The two lists are merged and de-duplicated.
- A cross-encoder reads query and passage together and scores the pair, which is slower and considerably more accurate than either score alone.
- Only the top passages survive into the prompt.

MedHopper runs a dense index with cross-encoder reranking; in the curated corpus the BM25S index is merged with the dense one before reranking.

Backup: routing and answer types

Four strategies

- Direct: the answer sits in one passage.
- Definition: the question asks what an entity is.
- Intersection: two constraints hold of one entity.
- Multi-hop: a second retrieval keyed on the first result.

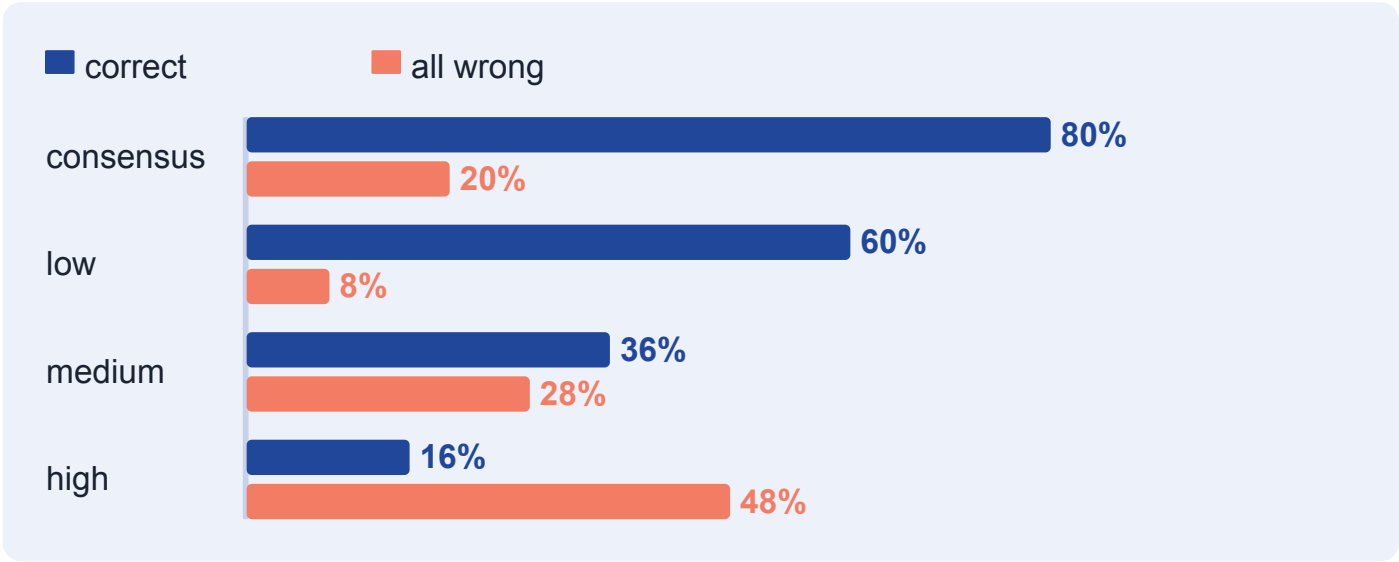
Observed split

- Direct and definition take 73.1% of the test questions.
- Multi-hop takes 25.6%.
- The bounded loop is the exception, which is why the system stays fast.

Answer types

- 18 types are predicted before generation.
- Gene, chromosome, drug class, process, symptom, yes or no, and so on.
- The type is checked again by a validator after the draft is written.

Backup: answer variation predicts error



100 random MedHopQA questions, 25 per bin.

Reading

- Accuracy falls from 80% to 16% as disagreement rises.
- In the hardest bin almost half the questions are missed by every configuration.
- Disagreement therefore marks questions that are objectively harder, and gives a confidence signal that needs no reference answer.

Backup: rare diseases and the Orphanet expansion

What the expansion does

- Orphanet supplies synonyms and candidate genes for a rare-disease mention.
- They go into the query only, and stay out of the evidence set.
- The generator still sees nothing but retrieved passages.

What it gives

- Rare-disease accuracy 55.3% against 42.6%, N = 47.
- About 0.01 exact match over the whole test set.
- The gain is concentrated where the vocabulary gap is widest.

Keeping the expansion out of the evidence set preserves the rule that every statement in the answer is grounded in a retrieved passage.

Backup: limitations

Against the leaderboard

- 0.55 against a leading track submission at 0.873.
- Those systems use commercial models, live web search and ensembles.
- Ours is one local 7B model over a fixed offline corpus.

What holds the score down

- The corpus is a Wikipedia subset, far from a curated biomedical library.
- Exact match penalises a correct answer in the wrong surface form.
- Term normalisation is absent, so a synonym scores as wrong.

What comes next

- Normalise answers against a controlled vocabulary.
- Retrieve over full text where licensing permits it.
- Calibrate the refusal threshold against the variation signal.

The informative comparison is against our own earlier pipeline under identical conditions: 0.55 against 0.43.

Backup: the prediction layer and the ORCA classes

PASS-based models

- Input: a pair of structures described by PoSMNA descriptors built from both molecules, so a combination absent from the training set can still be predicted.
- Output: Pa, the probability of activity, and Pi, of inactivity.
- They are independent estimates, so Pa + Pi differs from 1
- Ranking runs on the signed difference; a strong negative value asserts that the class is excluded, which the training set supports through documented non-interacting pairs.

ORCA severity classes

- 1, contraindicated: the combination must be avoided. IAP 0.876.
- 2, provisionally contraindicated: only in exceptional circumstances. 0.788.
- 3, conditional: significant, acceptable with monitoring. 0.683.
- 4, minimal risk: minor effects, intervention unlikely. 0.671.
- 5, no interaction expected. 0.752.

ORCA: J Am Pharm Assoc, 2001, 41(2), 161-5. | Dmitriev A. V. et al., Molecules, 2019, 24, 3955. | way2drug.com

Backup: how the prototype is instructed

The prompt carries the method

- Pa and Pi are defined in the system prompt as independent estimates, with ranking by their signed difference and the thresholds used here.
- Negative differences are passed through on purpose, because a rejected class carries information.
- Silence in the literature leaves the predicted activity standing.

Two stages

- Stage one gathers evidence with short literal queries.
- Stage two interprets the prediction tables together with that text, with no tools attached.
- The split keeps retrieval and reasoning separately inspectable.

Engineering

- A server-side proxy, so the key stays off the browser.
- Every result cached against the task identifier and reusable.
- Request and response are logged, so a report can be reproduced.
- The prototype currently calls a hosted model; moving it to the local 7B model is the next step.