JEVDB · SEMANTIC QUERY PROCESSING

Prune first. Decide fast.

Fewer candidate pairs.
Less generation per decision.

Scalable Semantic Query Processing with JEVDB

Zhengle Wang[1] · Hanxu Yan[1]
Fuheng Zhao[2] · Chunwei Liu[1]

[1] Purdue Data & AI System Lab
[2] University of Utah

Jihwan Ryu

Paper seminar for KAIST LDIS Lab

arXiv:2610.02046v1 (2026) · Independent academic project, not affiliated with or endorsed by TypeSafe AI (footnote 1)

Seminar overview

03–05 **Jev and the query engine**
Typed decisions and the three-tier engine.

06 **Operators and typed outputs**
Select, classify, rank, and join rows.

07–12 **Pruning, reuse, and reasoning**
Reduce candidates and repeated model work.

13–16 **Quality, time, and cost**
Read the gains together with their conditions.

17 **Planning and reusable state**
Choose execution paths; keep reused state valid.

18–19 **Verification and synthesis**
Verification costs and the paper’s main takeaway.

Jev: probabilities, categories, and scores

ILLUSTRATIVE REPORT R1

"40% battery. Opening the camera shuts down the phone."

JEV · SYSTEM 1 DECISION MODEL

Input text + task + output space produce a typed decision.

One forward pass per judgment

TASK FOR THE SAME REPORT	VALUE RETURNED	PRIMITIVE
Assess a battery-fault predicate	Probability that it is true	**Noul**
Identify the component	Category: battery / display / other	**Choice**
Rate defect severity	Expected level on a 0–4 scale	**Score**

These are output types, not measured predictions. Typed output does not guarantee a correct judgment.

Decisions without text generation

ILLUSTRATIVE BULLETIN B1

“Under peak load, battery voltage can drop below the shutdown threshold.”

PAIR TO JUDGE · REPORT R1 + BULLETIN B1

B1 covers the shutdown reported in R1.

Generative execution

Read pair → decode answer tokens → parse Boolean

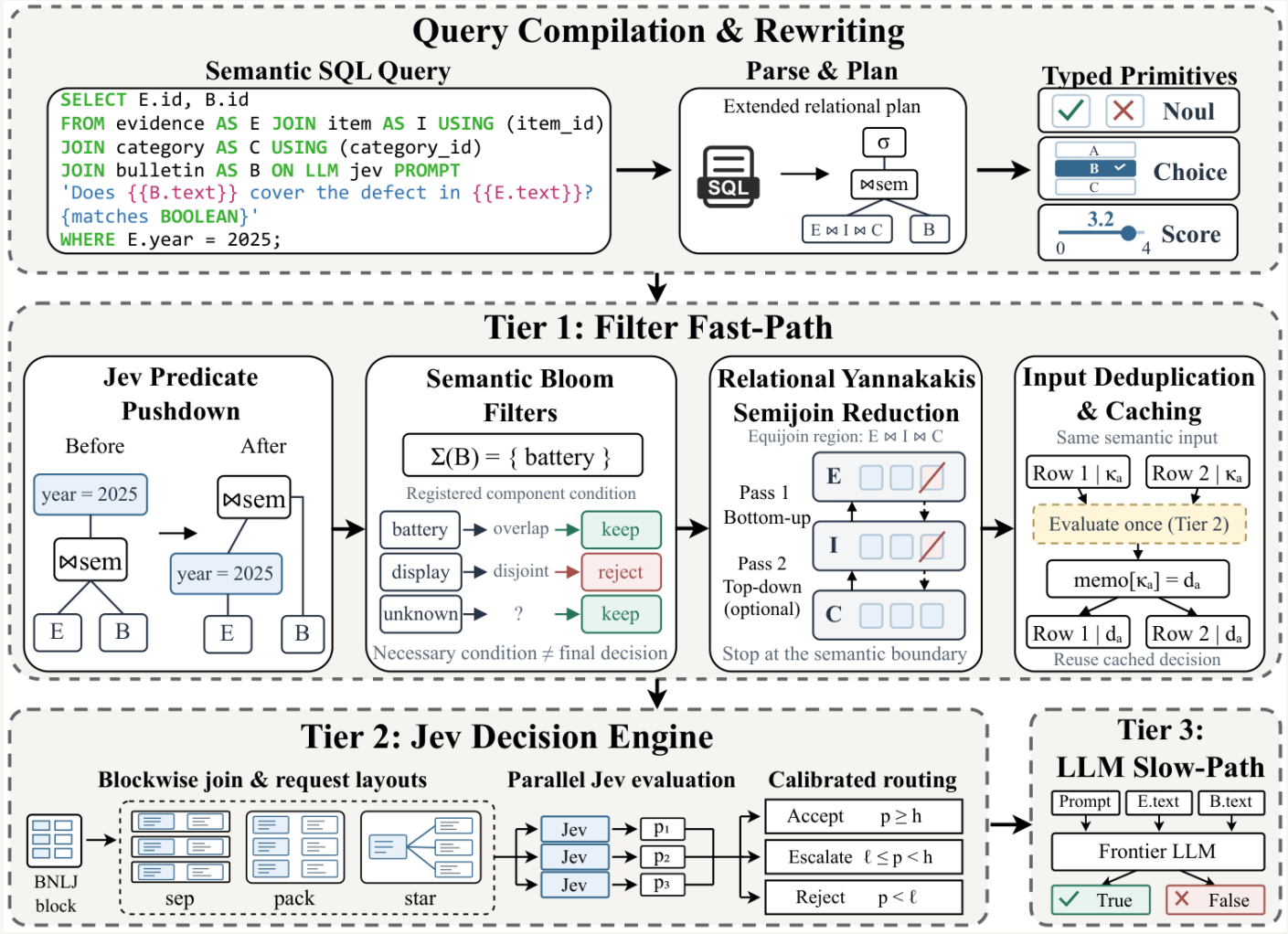
Jev / Noul

Read pair → return a truth probability

The DB uses the probability to make a decision.

Cheaper decisions still leave the candidate-count problem.

The adaptive three-tier engine



RELATIONAL–SEMANTIC MAPPING

Define each judgment and its output type

TIER 1 · CANDIDATE REDUCTION

Screen with relational keys and registered conditions

TIER 2 · JEV FAST PATH

Judge pairs or reuse document distributions

TIER 3 · LLM ESCALATION

Escalate uncertain pairs to a reasoning LLM

Map semantic operators to Jev

RELATIONAL–SEMANTIC ALGEBRA MAPPING · THE OUTPUT FOLLOWS THE OPERATOR

SQL operation	Input bound by the DB	Jev	Value used by the DB
Filter	One row + predicate	Noul	Boolean / probability
Classification	One row + finite labels	Choice	VARCHAR category
Ranking	One row + ordered levels	Score	DOUBLE expected score
Join	Candidate pair + predicate	Noul	Boolean match

Choice selects the most likely class.
Use the category to group reports.

Score averages ordered levels by probability.
Use the degree to rank reports.

JEVDB v1 §3.1, Table 1, Eqs. (1)–(5) · Illustrative report/bulletin predicate

Prune with relational structure

RELATIONAL REGION · EXPLICIT KEYS

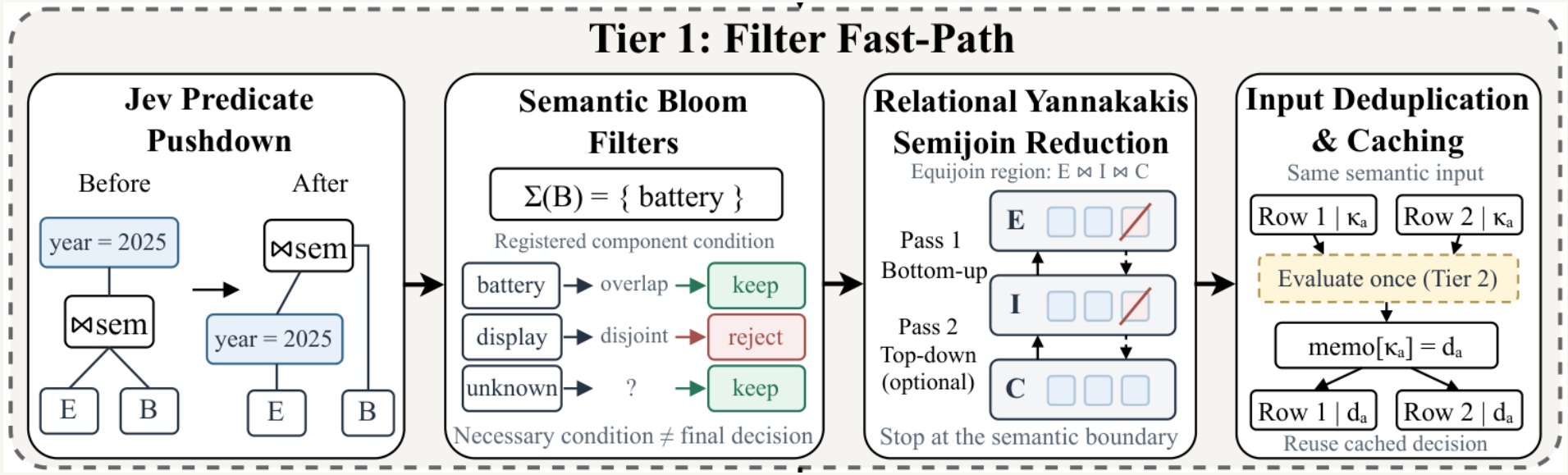
Report (E) · Item · Category

Year = 2025; matching item/category keys

SEMANTIC EDGE · LATENT MATCHES

Report · Bulletin

Surviving reports: R1 · R2 · R3
Their bulletin matches remain unknown.



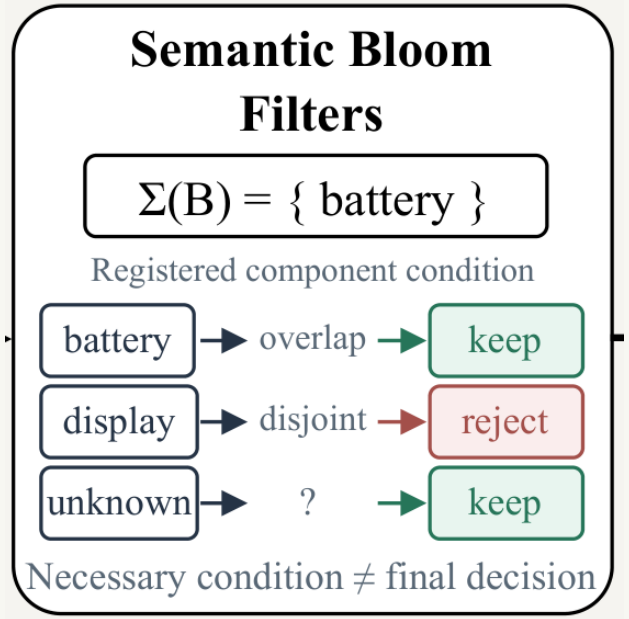
E denotes reports. R1–R3 are our illustrative survivors, not rows from the paper's dataset.

JEVDB v1 Figure 1 (Tier 1 crop), PDF p.5; §3.2 · Original figure

Screen pairs with necessary conditions

Registered condition: a valid defect match requires compatible components.

UNARY SBF · ONLY BATTERY BULLETINS



Original Figure 1: the display report cannot match any available bulletin.

ADD A DISPLAY BULLETIN · PAIRWISE SCREEN

Report	B1 battery	B2 display
R1 · battery	Evaluate	**Reject**
R2 · display	**Reject**	Evaluate
R3 · unknown	Evaluate	Evaluate

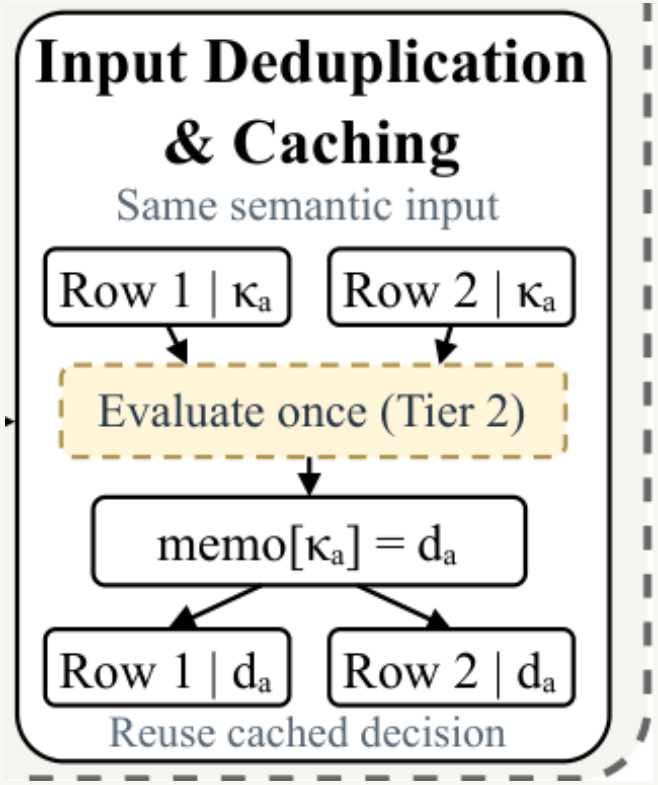
6 candidate pairs → 4 for semantic evaluation

Unknowns stay. Lossless pruning requires a sound condition and sound attribute extraction.

JEVDB v1 §3.2.1, Figure 1 (SBF detail), PDF p.5 · Original crop, CC BY 4.0 · Right: illustrative pairs

Reuse judgments, preserve result rows

ORIGINAL FIGURE 1



Exact inputs · one query · pointwise predicate

R1–B1 CAN OCCUR IN MULTIPLE RESULT ROWS

Predicate: the bulletin covers the reported defect.

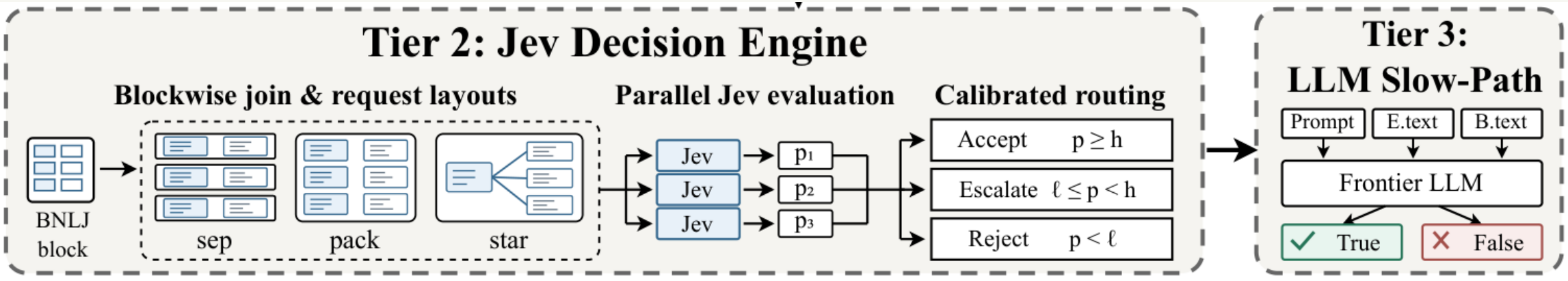
Row	Report R1	Bulletin B1
A	40% battery. Opening the camera shuts down the phone.	Under peak load, battery voltage can drop below the shutdown threshold.
B	40% battery. Opening the camera shuts down the phone.	Under peak load, battery voltage can drop below the shutdown threshold.

One distinct input → one cached decision

True: emit A and B **False: emit neither**

JEVDB v1 §3.2.3, Eq. (11), Figure 1 (deduplication detail), PDF p.5 · Original crop, CC BY 4.0 · Illustrative rows

Batch requests, share repeated text

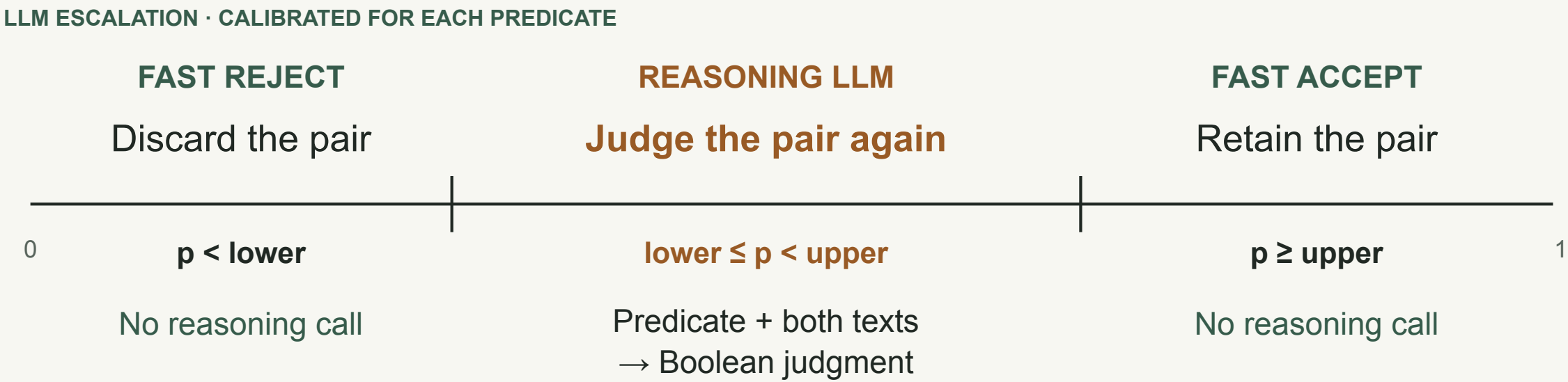


Back to the B1 column: R1–B1 and R3–B1 survive screening and remain distinct inputs.

SEP	PACK	STAR
B1 + R1 B1 + R3	Both pairs together; B1 appears twice	B1 once; R1 and R3 separately

Same two judgments. Pack reduces requests; star also shares the repeated bulletin.

LLM escalation for uncertain pairs



Full JEVDB calibrates a band against sampled LLM labels.

400 samples per predicate. Allowed recall / precision loss: 0.05 each, relative to the judge.
Confident errors can still bypass the LLM.

JEVDB-Flash uses a fixed 0.5 threshold and makes no reasoning calls.

JEVDB v1 §3.3.3, Eqs.12–13; §4.1 · Direct Jev path; threshold positions are schematic

Reuse each document’s interpretation

TWO ALTERNATIVES FOR THE FIRST SCORE · NOT TWO SUCCESSIVE STAGES

Direct Noul on surviving pairs

up to n × m

Pairwise model reads in a full candidate space.
Pruning reduces the pairs actually read.

OR

Reuse document interpretations

n + m

Interpret n reports and m bulletins once.
Combine stored distributions for each pair.

The index-derived score can replace direct Noul before accept, reject, or escalation.

20,558 documents · 15 s / $0.54 build · pair management and escalation remain

§3.3.4 leaves the combination function and score-specific calibration unspecified.

SemBench and Shelob: data and queries

SemBench

21 operator queries

Filter, join, classification, ranking
Text and structured attributes

Flash vs published baselines

Shelob

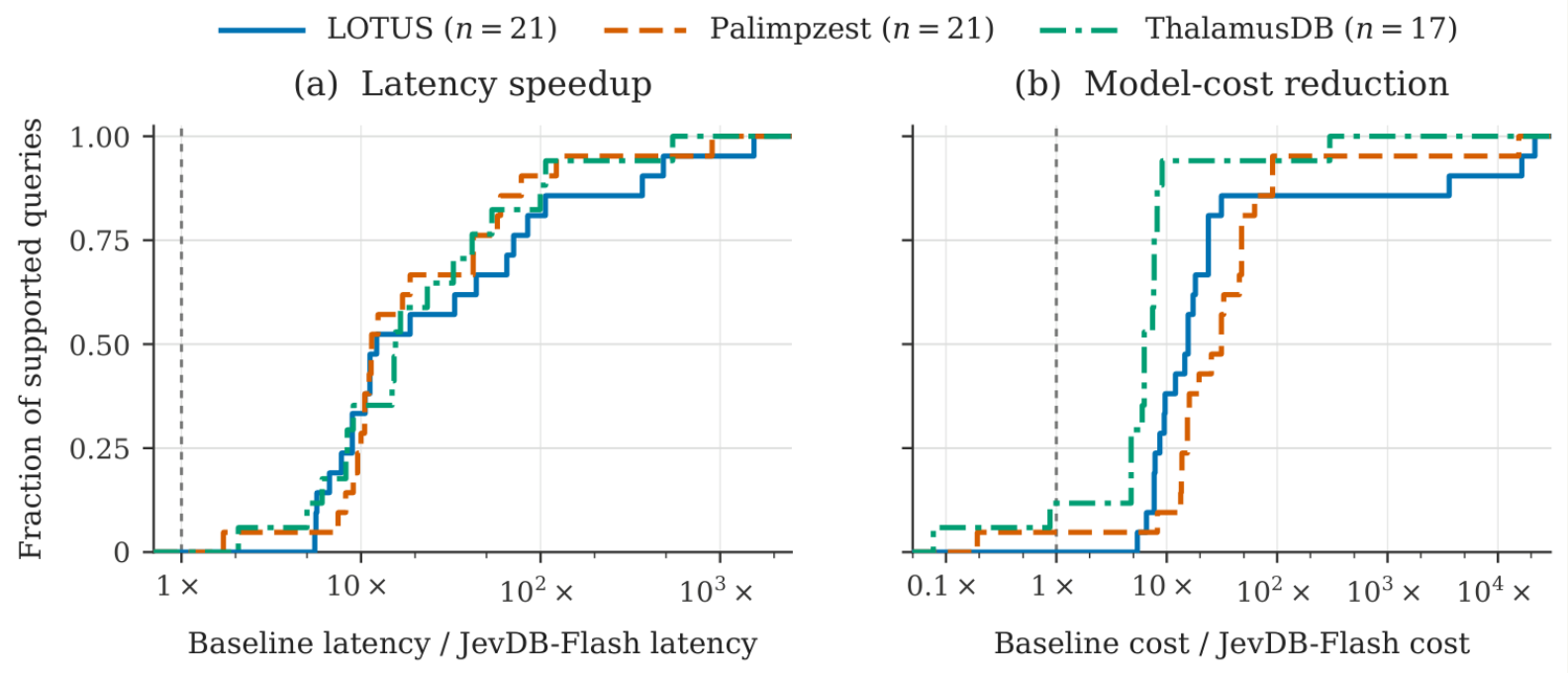
5 joins × 3 scales

TPC-DS with reports and bulletins
Linked release-note documents

Flash / full vs rerun baselines

For each query: answer quality, elapsed time and token-priced model cost

SemBench: a combined model and system gain



21-QUERY TOTAL TIME RATIO

89.4×

LOTUS / JEVDB-Flash
3,709 s / 41.50 s

14 / 21

best or tied in quality
Four more within 0.02

Original per-query ratio CDFs. Values above 1 favor Flash.

Jev vs Gemini 2.5 Flash: model and execution strategy both change.

Flash: 1 run, threshold 0.5, no reasoning. Baselines: 5-run means. Nominal concurrency: 20.

Shelob: 87.4% fewer candidate pairs

Q1–Q5 · XS/S/M AGGREGATE

SBF · CANDIDATE PAIRS

1.78M → 224K

87.4% removed before pair evaluation

INDEX SCORING · REASONING ESCALATIONS

194K → 87K

Among SBF survivors. Baseline details unspecified.

M TIER · ALL FIVE SYSTEMS · 1,200 s PER QUERY

	Completed · F1 (%)		Timed out · candidate pairs evaluated (%)		
Query	Full JEVDB	Flash	LOTUS	Palimpzest	ThalamusDB
Q1	96.3	85.3	0.0%	5.8%	1.1%
Q2	94.9	73.4	0.0%	23.6%	3.6%
Q3	91.5	73.8	0.0%	15.2%	3.0%
Q4	98.7	23.8	0.0%	5.5%	1.0%
Q5	96.9	59.8	83.7%	95.1%	29.6%

Timed out: evaluated-pair progress, not F1. LOTUS used CPU embeddings.
LOTUS 0%: embedding or threshold-learning timeout. GPU may reduce overhead (§4.1).

Shelob Q4: quality costs time

M Q4 · REPORT–BULLETIN JOIN · 540,180 CANDIDATE PAIRS

	JEVDB-Flash	Full JEVDB
F1 against reference answers	23.8%	**98.7%**
Elapsed query time	114 s	**883 s**
Model cost	$3.706	**$2.175**

How were the reference answers produced and checked?

Independence from escalation is unverified. S/M query cost excludes index setup: 15 s / $0.54.

Future work: planning and reuse

Choose a plan

Semantic-aware query optimization

Screen	Checks	Pairs left
Unary	3	6
Pairwise	6	4

Compare saved calls with screening cost.
Jev pushdown is the paper's default.

Keep reuse valid

Reusable semantic state

R1 changes: power stays on,
but the screen flickers.

battery → display

Refresh the affected interpretation
and decisions that depend on it.

Further directions: scale-out semantic joins and multimodal semantic queries

Research perspective: verification costs

6 TASKS × 3 DEFECT CONDITIONS

Astra executes and repairs. The completion verifier changes.

Astra verifier / Jev Choice verifier

T-284 · INJECTED CONFLICT

Lookup: out_for_delivery
Injected answer: processing

Choice flags it. Astra repairs. The verifier rechecks.

−18.17 s

Median paired runtime difference
Jev Choice − Astra verifier

18 / 18

Strict success in each setting
One run per case and setting

Agent loop only. No low-confidence Astra escalation configured. Shared-account load uncontrolled.

JEVDB reduces the work of inference

Prune candidates → Decide and reuse → Escalate

Relational keys
Necessary conditions

Typed Jev outputs
Stored interpretations

Reasoning for
uncertain cases

SemBench: fast end-to-end execution
Shelob: higher quality can require more time

Plan the cost of a result whose quality has been checked.