Jeva: System One Models from First Principles

A native architecture for typed decisions

Paper by Qi Li and Xinchao Wang
National University of Singapore

arXiv:2610.18773v2 | Paper revised 18 Oct 2026
Reading presentation: 28 Oct 2026

Takeaway: separate reusable options from set-dependent judgment

The paper's central move

Encode each option independently; refine the current request against the *whole* option set; separately estimate whether any option is valid.

- This yields an explicit null outcome, permutation-equivariant probabilities, variable-size choice, and candidate-side caching.
- Reported evidence: 98.66% null AUROC, zero flips in tested permutations, and 27.98× lower ratio of median latencies for fully reused novel MLDR sets.

Publication: not found as of 28 Oct 2026; author lists a 2026 preprint.

Official code: LiQiiiii/Jeva.

Paper, Abstract and Secs. 3–4, PDF pp. 1, 4–9.

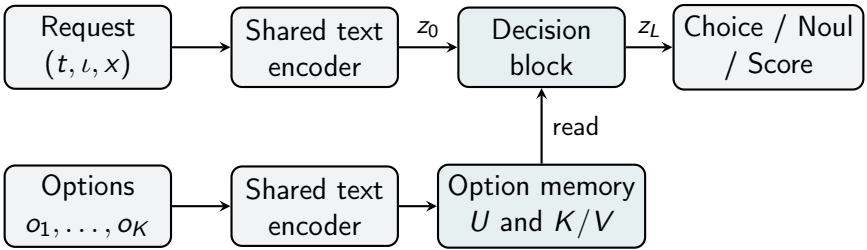
A decision problem has four structural requirements

Requirement	What a model should do
Null-aware	Distinguish no valid option from uncertainty among valid options.
Order-stable	Reordering alternatives should only reorder their probabilities.
Option-scalable	More candidates should not require a new label vocabulary.
Cacheable	Reuse computation for recurring options across new requests and sets.

Jointly serializing request and options ties reusable content to each request and gives unordered options incidental text positions.

Paper, Sec. 1, PDF pp. 2–3; Fig. 1, PDF p. 1.

Jeva updates the request; its option memory stays fixed



- Option memory is request-independent and may be cached.
- Only the request state is refined against the current candidate set.

Teaching schematic redrawn from Paper, Fig. 1, PDF p. 1; Sec. 3, PDF pp. 4–6.

Independent encoding makes each option reusable

$$z_0 = P_x E_\theta(T_x(t, \iota, x)), \qquad u_i = \mathrm{nrm}\big(P_o E_\theta(T_o(t, o_i))\big), \quad z_0, u_i \in \mathbb{R}^d.$$

- x: state; ι: question; t: decision type; o_i: candidate text.
- T_x includes type, question, state. T_o includes only type and one option.
- The shared backbone pools the last non-padding token; the option projection is ℓ_2 normalized.

Cache identity

The same typed option has the same encoder input under fixed weights, tokenizer, template, and length policy.

Paper, Eq. (1), Sec. 3.1, PDF p. 4; Appendix C, PDF pp. 16–18.

Independent candidate scores cannot reverse a retained pair

$$q_i(\mathcal{O}) = \frac{e^{g(x,\iota,o_i)/\tau}}{\sum_{j\in\mathcal{O}} e^{g(x,\iota,o_j)/\tau}}, \qquad \log \frac{q_A(\mathcal{O})}{q_B(\mathcal{O})} = \frac{g(x,\iota,A) - g(x,\iota,B)}{\tau}.$$

- Adding a third option changes q_A and q_B, but their odds ratio does not change.
- Jeva keeps option vectors reusable while letting the *request state* depend on the current set.

Paper, Eq. (2), Sec. 3.1, PDF p. 4.

A small mixer collects evidence from the whole set

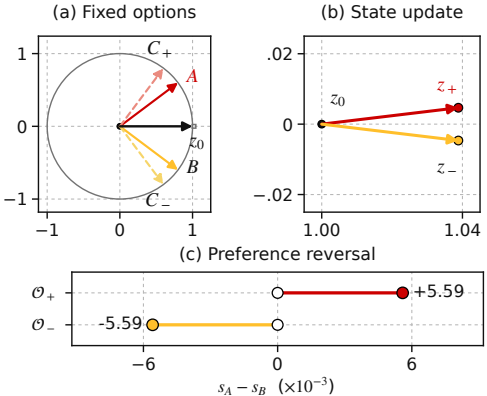
$$a_{\ell i}^{h} = \mathrm{softmax}_i\left(\frac{(q_\ell^h)^\top k_{\ell i}^h}{\tau_\ell}\right), \qquad m_\ell = W_\ell^O \,\mathrm{Concat}_h\left(\sum_i a_{\ell i}^h v_{\ell i}^h\right).$$

$$\widetilde{z}_\ell = z_\ell + \alpha_\ell m_\ell, \qquad z_{\ell+1} = \widetilde{z}_\ell + L^{-1/2} F_\ell(\mathrm{RMSNorm}(\widetilde{z}_\ell)), \quad \alpha_\ell = \frac{\tau_\ell \sigma(\gamma_\ell)}{\sqrt{L}}.$$

- Keys and values come from unchanged options; only z_ℓ evolves.
- $a_{\ell i}^h$ allocates *evidence*; it is not the final choice probability.

Paper, Eqs. (3)–(4), Sec. 3.2, PDF pp. 4–5; Appendix A.1, PDF p. 11.

Even one block can reverse a pair's preference



- Keep request, A, and B fixed; replace only C_+ by C_-.
- The refined state moves in opposite directions.
- The authors construct $s_A - s_B = \pm 24\alpha w/25 \approx \pm 0.00559$; A and B win in different sets.

Paper, Proposition 1 and Eq. (5), PDF p. 5; Fig. 3, PDF p. 3. Existence construction.

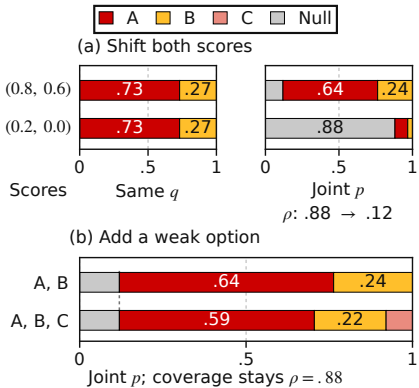
Competition and answer availability are different probabilities

$$s_i = z_L^\top u_i, \qquad q_i = \mathrm{softmax}_i(s/\tau_{\mathrm{dec}}),$$
$$\rho = \sigma\left(\frac{\max_i s_i - b_{\mathrm{cov}}}{\tau_{\mathrm{cov}}}\right), \qquad p_\perp = 1 - \rho, \qquad p_i = \rho q_i.$$

- q_i: preference *conditional on some valid option existing*.
- ρ: estimated probability that at least one candidate is acceptable.
- Open-set Choice uses $(p_\perp, p_1, \ldots, p_K)$; closed-set Choice uses q.

Paper, Eq. (6), Sec. 3.3, PDF p. 5.

The coverage gate avoids a forced candidate-count effect



- Shift all scores: q stays fixed, ρ changes.
- At *fixed final scores*, adding a below-max option keeps ρ fixed.
- Flat softmax with a null logit instead gives

$$p_{\perp}^{\text{flat}} = \sigma\left(\frac{b_0 - s}{\tau} - \log K\right).$$

Paper, Proposition 2 and Eq. (7), PDF pp. 5–6; Fig. 4, PDF p. 4. Set refinement can still alter scores.

One head supports Choice, Noul, and Score

Type	Readout
Choice	Closed set: select from q; open set: include $p_\perp$ and select from p.
Noul	Return q_{i^+}, probability of the positive Boolean candidate.
Score	Return $\sum_i q_i v_i$, expected value in the supplied rubric units.

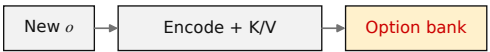
$$\ell_{\text{open}} = \text{BCEWithLogits}(\text{logit}\,\rho, c_y) - \sum_i y_i \log q_i, \qquad c_y = \sum_i y_i.$$

All-null examples train coverage without imposing a candidate ranking.

Paper, Eqs. (6), (9), PDF pp. 5–7; Appendix C examples, PDF pp. 16–17. Noul is not $p_\perp$.

Symmetry is structural; caching reuses candidate computation

(a) Prepare each option once



$$\mathcal{M}(o) = (u_o, \{K^o_\ell, V^o_\ell\}_\ell)$$

(b) Three requests, overlapping sets

Request / set	Option work	Hits
x_1, $\{A, B, C\}$	Prepare A, B, C	0/3
x_2, $\{B, C, D\}$	Reuse B,C; new D	2/3
x_3, $\{A, C, D\}$	Reuse A, C, D	3/3



- Reordering complete candidate entries reindexes the attention sum; state and coverage stay fixed, candidate probabilities permute.
- A cached entry contains u_i and each layer's key/value projections.
- New request: state encoding and decision readout still run.

Paper, Proposition 3, Eq. (8), PDF p. 6; Fig. 5, PDF p. 4. Exact arithmetic, fixed weights.

Prepared memory gives linear-in-K decision work

$$\mathcal{M}_t(o_i) = \left(u_i, \{k^h_{\ell i}, v^h_{\ell i}\}_{\ell,h}\right),$$

$$\text{decision cost} = O\big(L(d^2 + Kd + df)\big), \qquad \text{bank storage} = O\big(M(L+1)d\big).$$

- K: candidates in one request; M: options in the stored bank; f: FFN width.
- New options still need text encoding and key/value projection.
- Hybrid cache hits individual options even when the assembled set is new.

Jeva-4B configuration

Qwen3.5-4B backbone; $d = 768$, $L = 1$, $H = 8$, $f = 3072$; rank-64 LoRA plus 13.37M projection/head parameters.

Paper, Secs. 3.4–3.5, PDF pp. 6–7; Appendix B.6–B.7, PDF p. 15.

Evaluation separates quality, structure, and latency

Question	Protocol / metric
General accuracy	Typed Decisions (2,000); BANKING77/MASSIVE ($K = 8, 16$); Mind2Web (three $K = 16$ slices).
Null detection	3,080 BANKING77 requests at $K = 16$; 588 answer-absent; AUROC and TPR at near-5% FPR.
Order stability	600 Choice items; original, reverse, up to three seeded orders; align candidate identities.
Large sets	500 ToolRet requests in six disjoint $K = 32 \ldots 2048$ cohorts; support, accuracy, p50 latency.
Reuse	Off, exact-set, hybrid; MLDR novel sets and shared banks; priming separately timed.

Paper, Sec. 4, PDF pp. 7–9; Appendix C, PDF pp. 15–18. Open-set baseline adapters differ by interface.

Jeva leads seven of eight general-accuracy slices

Slice	Jeva-4B	Best other	Difference
Typed Decisions	78.30	78.70 (Intern-4B)	−0.40
BANKING77, $K = 16$	91.62	82.50 (Jev)	+9.12
MASSIVE, $K = 16$	90.22	81.61 (Jev)	+8.61
Mind2Web task	79.99	66.26 (Jev)	+13.73
Mind2Web website	62.25	59.97 (Jev)	+2.28
Mind2Web domain	66.23	62.47 (Jev)	+3.76

Accuracy is first-argmax %, including null when applicable; differences are percentage points. Comparators here have full slice support.

Paper, Table 1, PDF p. 5; Sec. 4.1, PDF p. 7.

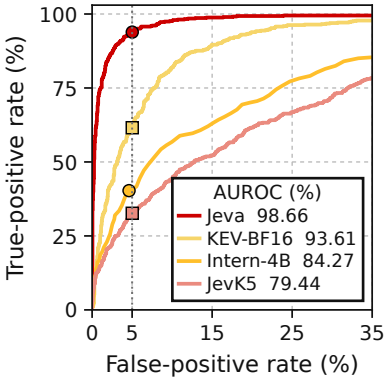
Typed fidelity is strong, but not every metric is best

Model	Noul Brier ↓	Score MAE ↓	Score RMSE ↓
Jeva-4B	**0.0184**	**0.2315**	0.3222
InternDecision-4B	0.0340	0.2488	**0.3208**
Jev-1.13.0	0.0342	0.3744	0.4795

- Noul: 600 items with soft probability targets; Score: 800 items.
- Score errors use the original rubric's numeric units.
- Brier and RMSE are different metrics; compare models within a column.

Paper, Table 2, PDF p. 6; Fig. 6(a), PDF p. 7.

A separate coverage gate detects missing answers

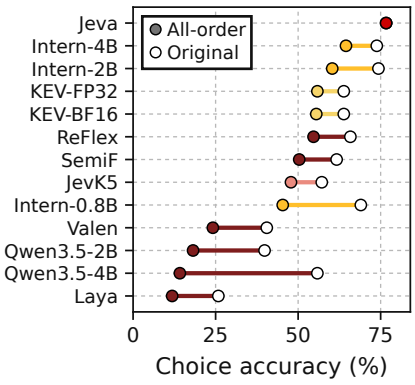


(b) Null detection

- BANKING77, $K = 16$: 3,080 matched requests; 588 have no valid option.
- Null AUROC: Jeva 98.66%; KEV-BF16 93.61%; Intern-4B 84.27%; JevK5 79.44%.
- Jeva detects 93.88% of absent answers at 5.02% FPR.

Paper, Sec. 4.2, PDF pp. 7–8; cropped from Fig. 6(b), PDF p. 7.

Observed Choice decisions do not flip under tested orders

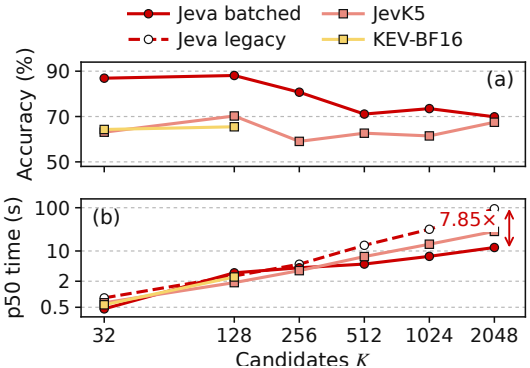


(c) Order stability

- 600 Choice items, original + reverse + up to three seeded permutations.
- Jeva: 0% flips, mean TV 3.17×10^{-8}, 76.67% all-order accuracy.
- Jev: 17.00% flips, 66.67% all-order accuracy; Intern-4B: 22.67%, 64.50%.

Paper, Table 2, PDF p. 6; Sec. 4.1 and cropped Fig. 6(c), PDF pp. 7–8.

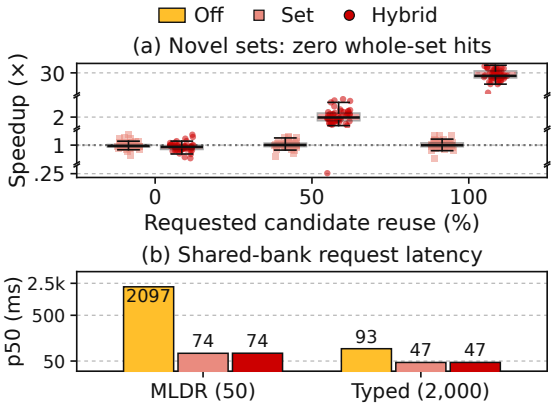
Large sets remain supported; batching cuts encoder latency



- ToolRet: 500 requests in six *disjoint-query* cohorts; added negatives include weak or unjudged items. Jeva covers all.
- Aggregate accuracy: 78.40% vs recursive JevK5's 64.00%. At $K = 2048$: 69.88% vs 67.47%.
- At $K = 2048$, p50 is 12.040 s vs about 28 s; length-aware batching within Jeva changes 94.459 s to 12.040 s (7.85×).

Paper, Sec. 4.3 and cropped Fig. 8(a,b), PDF p. 8. Batching: 256 rows / 131,072 padded tokens.

Hybrid reuse helps even when no whole set repeats



- Novel MLDR sets have *zero* set-cache hits; priming is timed separately. With full option reuse, p50 goes from 2,102.0 to 75.1 ms: 27.98× ratio of medians.
- At about 50% option reuse: 1.95×. Exact-set cache stays near 2.1 s on novel sets.
- Shared Typed bank: 93.1 to 46.7 ms with hybrid reuse; 99.2% candidate rows avoid encoding.

Paper, Sec. 4.4 and cropped Fig. 9(a,b), PDF pp. 8–9. Priming is timed separately; no answers are cached.

What the paper establishes, and what to carry forward

1. Independent option features can still support set-dependent choice when a small state reads the full set.
2. Coverage, competition, symmetry, and cacheability have explicit computational roles; the experiments test each one.
3. Reuse gains concern candidate-side work under fixed weights and measured cache conditions. Request encoding and decision still run.

Interpretation: cross-model gains also reflect differing training and evaluation adapters; the paper does not isolate every architecture component with a common-backbone removal study.

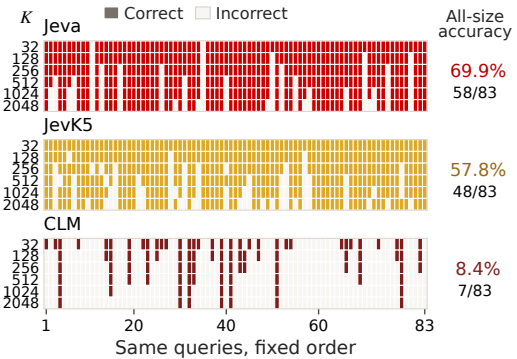
Paper, Secs. 3–5, PDF pp. 4–9; Appendix B.5–B.7, PDF pp. 14–15.

Paper and release links

- Qi Li and Xinchao Wang. *Jeva: System One Models from First Principles*, arXiv:2610.18773v2 (18 Oct 2026).
- Official implementation and documentation: LiQiiiii/Jeva.
- Released Jeva-4B model.
- Author publication page: lists Jeva among 2026 preprints (checked 28 Oct 2026).

Publication status at lookup: no formal venue found; the author lists this work as a preprint.

Backup: the same queries across growing candidate sets



Paper, Fig. 7, PDF p. 7; Sec. 4.3, PDF p. 8.

- 83 fixed queries; nested sets grow from 32 to 2,048 candidates.
- Correct at *every* size: Jeva 58/83 (69.88%); JevK5 48/83 (57.83%); CLM 7/83 (8.43%).
- This 69.88% uses a different denominator from the $K = 2048$ ToolRet cohort.

Backup: open-set evaluation and numerical limits

- Open-set baseline adaptation: JevK5/CLM receive a literal or descriptive None option; Jev/InternDecision/KEV use an auxiliary coverage Noul question.
- Exact symmetry and cache substitution assume fixed weights, deterministic exact arithmetic, and full candidate metadata permutation.
- Implementation: no decisions change across cache conditions, but maximum absolute probability difference is 0.00736.
- Novel-set sample count differs across manuscript locations: 50 matched queries in Sec. 4.4; 46 plotted per reuse level in Fig. 9(a) caption and Appendix C.4.

Paper, Sec. 4, PDF pp. 7–9; Appendix B.5 and C.4–C.6, PDF pp. 14, 16–18.