

JevAlt: Technical report

Mert Kaya

Eschatia Labs, 2026, eschatialabs.com

Technical report (not peer-reviewed). Every figure comes from the published sources listed at the end.

What it is

JevAlt comprises Deem-4B for English, Karar-4B for Turkish and Wähler-4B for German. They speak the Jev API from TypeSafe: a request sends a state and typed questions (Choice, Score, Noul), and the answer gives a calibrated probability for every option. The Q4_K_M builds run on a laptop CPU in about 3 GB of RAM.

Code and weights are published under Apache-2.0. JevOss provides probes and calibration for any model that speaks the same API, and Emberwick uses the three models to decide what villagers do in a browser game.

Method

The published recipe starts from Intern-Decision-4B by InternLM. LoRA training uses one shared run over all three languages, followed by one run per language. The loss compares option probabilities with soft labels, and a temperature fitted on held-out decisions calibrates the outputs.

Training combines public datasets with known answers, situations written directly in each language, requests from Emberwick and targeted rows for planted instructions, missing facts, policies, negations, dates and numbers. A written row counts only when the writer and both teacher models agree on its answer. Test rows were split off by group before the final training runs.

Comparisons use the same items and the same client for every model, each as shipped. The named baseline is Kev-4B, an open decision model built on Qwen3.5-4B and released under Apache-2.0. Accuracy measures how often the answer is right; the Brier score measures how good the reported probabilities are.

Held-out results against Kev-4B

Each model is scored in its own language. Higher accuracy and a lower Brier score are better.

Model	Accuracy	Kev-4B accuracy	Brier	Kev-4B Brier
Deem-4B (English)	94.7%	84.7%	0.091	0.256
Karar-4B (Turkish)	96.8%	87.1%	0.058	0.214
Wähler-4B (German)	92.0%	81.1%	0.119	0.298

Targeted probes

The project also reports comparisons on planted instructions, policies, negations and missing facts. For planted instructions the figure is the share of answers that a hidden line changes, so lower is better.

Kev-4B offers no unknown option, and the start checkpoint, Intern-Decision-4B, already answers unknown in 9 of the 11 cases.

Probe	Kev-4B	JevAlt
Planted instructions: answers changed	36.0%	Deem-4B 14.0%, Karar-4B 19.0%, Wähler-4B 17.5%
Policies: correct answers	59.3%	80.0%
Negated questions: correct answers	76.7%	96.7%
Missing facts: answered unknown	0 of 11	9 of 11

Tests on the live models

On 4 October 2026 the three live models received 390 requests with known answers, 130 per language. Deem-4B answered 122 of 130 correctly, Karar-4B 113 and Wähler-4B 122. The benchmark dataset holds every request and answer.

Limits

- The held-out tests come from the project’s own data pipeline, so they favor JevAlt.
- Kev-4B leads elsewhere. On 10kGNAD, a German suite outside the training data, Kev-4B answers 65.3% correctly and Wähler-4B 62.5%. With 600 words of unrelated records in front of the facts, Kev-4B loses 5.4 accuracy points, Deem-4B and Karar-4B lose 17.4 and Wähler-4B 12.2.
- These are 4B models: general knowledge is limited, and long, noisy text is still a weak spot. Dates are shaky, also with reasoning on.
- Probabilities are calibrated on the project’s held-out data. Refit them on your own data with JevOss before setting thresholds.

Sources

- [JevAlt README](#)
- [JevAlt benchmark and live-test results](#)
- [JevAlt source code](#)

Citation

```
@techreport{kaya2026jevaltreport,  
  author = {Kaya, Mert},  
  title = {JevAlt: Technical report},  
  institution = {Eschatia Labs},  
  year = {2026},
```

```
url = {https://eschatialabs.com/research/jevalt-report/},  
note = {Technical report, not peer-reviewed}  
}
```