

# Fine-Tuned Encoders versus Zero-Shot Prompting for Detecting Threat and Solution Framing in German Climate News

Raven Adam · David Maier · Marie Kogler University of Graz, Austria



## Zero-shot prompting now matches fine-tuned German encoders at classifying threat and solution framing — but only the newest prompting model does.

Threat · weighted F1

**.807** vs **.805**

MiMo-V2.6-Flash (zero-shot)

 vs 

GBERT-large (fine-tuned)

. Solution: .798 vs .802

Context gain

**+4.8** F1 pts

from adding four sentences on each side of the target (GBERT-large, solution)

Agreement

**>80** %

of sentences get the same label from both top models; the rest can go to manual coding

Inference cost

**≈ \$2.3** USD

for all MiMo-V2.6-Flash runs, vs \$12.4 (Llama 4 Maverick) and \$13.1 (Qwen3.8-27B)

### 1 Why threat vs. solution?

**Threat framing** without efficacy information can induce fatalism; **solution framing** promotes engagement and policy support. Tracking the balance across large news corpora matters for climate communication research.

Manual coding does not scale. Two automated routes dominate: 

fine-tune an encoder

 on labelled data, or 

prompt an LLM

 with the coding guidelines.

RQ1 · Can zero-shot prompting match fine-tuned German encoders at sentence level?

RQ2 · How much does surrounding context help?

### 2 Data

**440** Austrian news articles, Jan–Jun 2024: *Kronen Zeitung* (304), *Der Standard* (136)

**10,981** sentences, manually coded in MAXQDA; labels inherited from coded segments

Two **independent** binary labels: a sentence can be threat-oriented, solution-oriented, both, or neither.

**Threat** · risks & dangers 2,575 · 23.4%

**Solution** · actions & remedies 4,108 · 37.4%

Neutral sentences undersampled to balance each task: 5,150 (threat) and 8,216 (solution) sentences.

### 3 Two approaches, identical folds

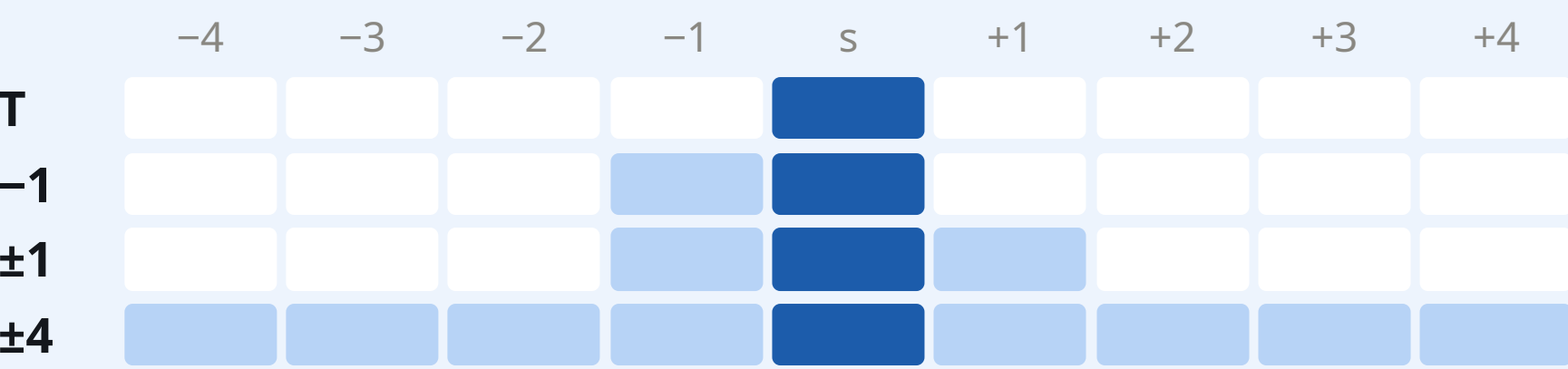
#### A · Fine-tuned encoders

GBERT-large · GELECTRA-large (both 335M) · GeistBERT-base

INPUT: SENTENCE PAIR

[CLS] target sentence [SEP] context [SEP]

FOUR CONTEXT WINDOWS



TRAINING

- One classifier per task × model × window
- AdamW, lr 2–3·10<sup>−5</sup>, 10–15 epochs
- No domain-adaptive pre-training

#### B · Zero-shot prompting

No training, no examples: the model gets only the written coding guidelines

PROMPT: SIX XML SECTIONS, IN GERMAN

- Role** · climate journalism analyst
- Speech-act taxonomy** · representative, directive, commissive, expressive, declarative
- Framing definition** · who, when, where, why, how
- Classification rules** · threat 9 incl./3 excl., solution 8 incl./7 excl.
- Context** · 4 sentences before and after
- Target sentence**

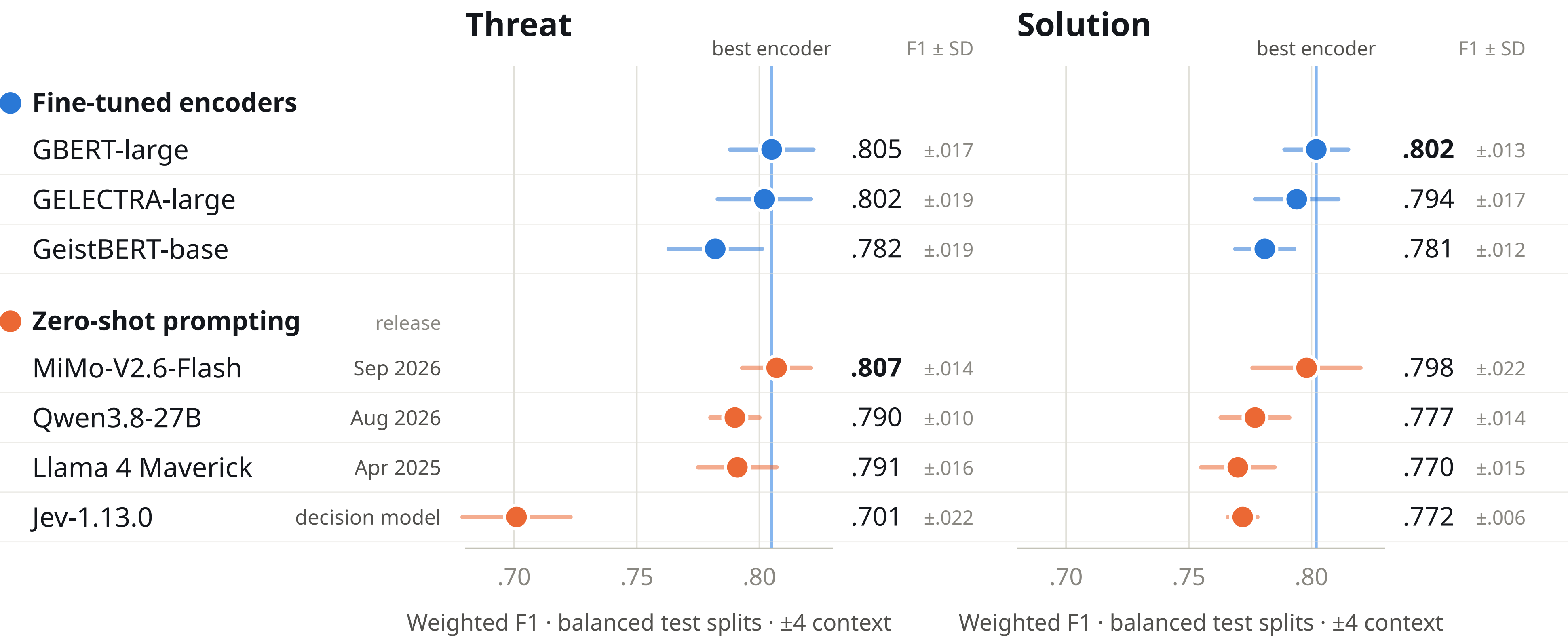
reason → classification → confidence  
JSON, reasoning first (chain of thought)

MODELS · TEMPERATURE 0

| Model            | Params   | Type     | Weights |
|------------------|----------|----------|---------|
| MiMo-V2.6-Flash  | 309B/15B | MoE      | open    |
| Qwen3.8-27B      | 27B      | dense    | open    |
| Llama 4 Maverick | 400B/17B | MoE      | open    |
| Jev-1.13.0       | n/a      | decision | closed  |

Same five article-grouped folds for every model (stratified, ≈80/20): no article appears in both training and test data. Metric: weighted F1, mean ± SD across folds.

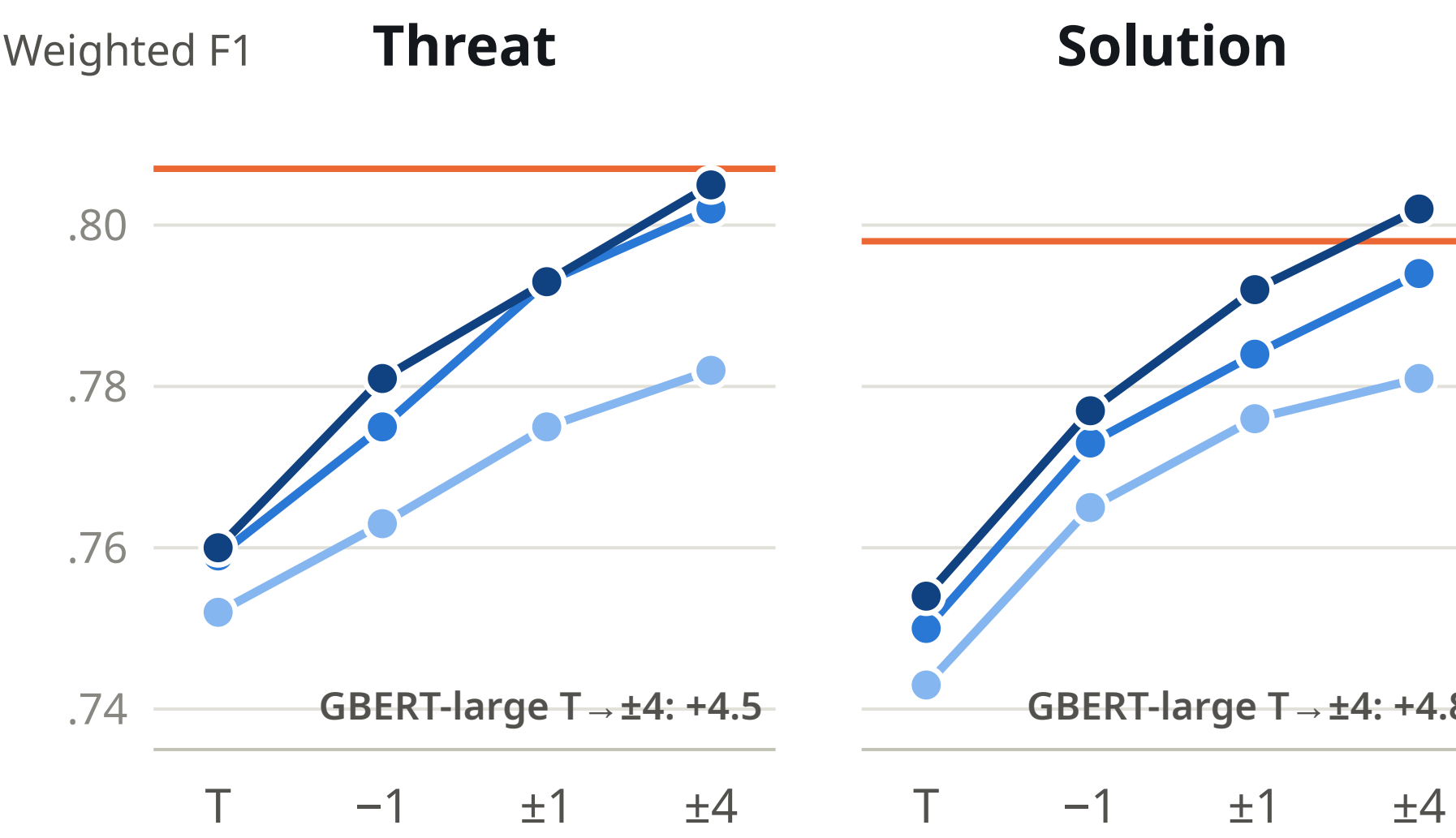
### 4 Parity at the top; earlier prompting models stay below



The top two differ by ≤ .004, well inside the fold SD (.013–.022): neither is superior. Llama 4 Maverick and Qwen3.8-27B only reach GBERT-large with ±1 context. Jev-1.13.0 (English-trained decision model) is stable on solution but weakest on threat.

### 5 Context helps every encoder

GBERT-large, GELECTRA-large, GeistBERT-base, MiMo-V2.6-Flash (zero-shot, ±4 context)



- Gains are monotonic for all encoders and both tasks.
- Biggest single step: adding the **preceding sentence** (+2.1 / +2.3 for GBERT-large).
- Model size matters more than pre-training objective: both large encoders are close; GeistBERT-base trails.

### 6 On the real class distribution

Best encoder vs best prompting model on all 10,981 sentences (pooled over folds).

| Task     | Model           | Pos. F1 | Weighted | Macro |
|----------|-----------------|---------|----------|-------|
| Threat   | GBERT-large     | .648    | .804     | .750  |
|          | MiMo-V2.6-Flash | .673    | .831     | .776  |
| Solution | GBERT-large     | .757    | .803     | .794  |
|          | MiMo-V2.6-Flash | .747    | .799     | .788  |

GBERT-large keeps high recall (.819 / .833) but loses **precision** (.536 / .694): trained on balanced data, it implicitly assumes a 50% base rate. Train on the natural distribution or adjust the threshold.

### 7 Where the two disagree

On contested sentences exactly one model is right. Share of contested sentences each gets right:



**Neither dominates.** GBERT-large says “yes” more readily (recall .612 / .583); MiMo is more conservative, especially for threat (recall .388, precision .561).

Hard cases for both: sustainability, legislation, climate litigation, activism, and framing that only a wider context reveals.

### 8 Take-aways

- Prompting is now a viable, low-cost alternative, even when labelled data exist.
- The gap closed only recently: **re-run such comparisons** as new models appear.
- Give encoders context: ±4 beats sentence-only for every model.
- Use agreement as a quality signal; **route disagreements** to manual coding.
- Encoders stay attractive where reproducibility and local execution matter.

#### LIMITATIONS

Sentence labels are inherited from coded segments and no formal intercoder reliability is reported. Hyperparameters were set without a separate validation split; no significance tests. One zero-shot prompt design, one pass per model; pre-training exposure to the articles cannot be ruled out. Two Austrian newspapers, six months.

#### SELECTED REFERENCES

- Chan, Schweter & Möller (2020). German’s next language model. *COLING*.  
Hart & Feldman (2014). Threat without efficacy? Climate change on U.S. network news. *Science Communication* 36(3).  
Riesinger (2024). Mixed-methods analysis of threat- and solution-oriented statements in Austrian newspaper articles on climate change. MA thesis, Univ. of Graz.  
Wei et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

[ QR code to paper ]

[ contact e-mail ]