# helium at Touché

## The case for naturalistic evaluation of fallacy detection

**Diana Markova**

Sofia University

Touché at CLEF 2026 / online talk

**NATURALISTIC EVALUATION**

Does the model detect fallacious reasoning - or the rewrite format?

**binary fallacy detection**

# A high score can answer the wrong question

## Fallacy detection is not fact-checking

An argument can manipulate beliefs through implicit assumptions, weak inference, or emotive framing even when its factual claims are true.

**Capability we want**

**Recognize fallacious reasoning in ordinary online discourse**

**Shortcut a benchmark may reward**

**Recognize label-correlated cues in a rewritten input format**

**The evaluation must separate these explanations.**

# Three views of each Reddit argument

### RAW + CONTEXT

**Original comment**

Comment + thread title + parent comment

most natural

### BASE

**Normalized rewrite**

Cleaner statement with relevant context incorporated

moderate intervention

### ENHANCED

**Label-aware rewrite**

Generated with access to the gold label

diagnostic only

**938**
labeled Touché rows

**738 / 200**
train / held-out validation

**195**
Sahai-aligned test proxy

**4,898**
LOGIC-binary stress test

**Key concern: the enhanced view knows the answer while it is being written.**

METHOD

# Models, input views, and evaluation sets

## MODEL FAMILIES

| | |
|---|---|
| **TF-IDF** | SVC / logistic regression |
| **Embeddings** | OpenAI 3-large + logistic head |
| **Encoders** | ELECTRA-small / RoBERTa-large |
| **LLM** | Gemma 4 31B, zero- and four-shot |

## INPUT VIEW

- **raw + ctx**
- **base**
- **enhanced**

## EVALUATION

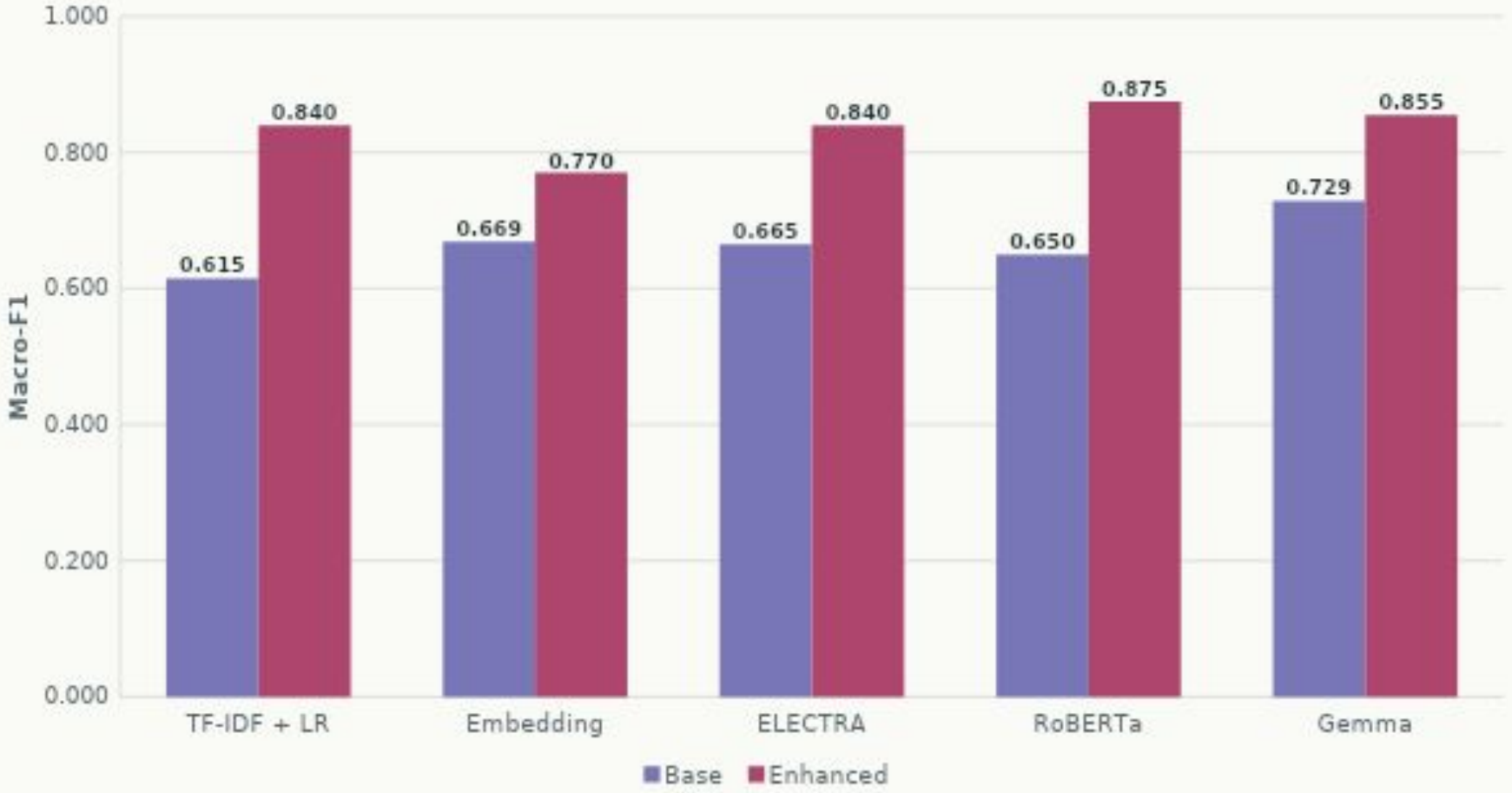
- **Touché validation**
  in-domain
- **Sahai proxy**
  Reddit-style auxiliary test
- **LOGIC-binary**
  independent-corpus stress test

**Primary metric: macro-F1**

**Control: hold the checkpoint fixed, change only the input view**

RESULT 1

# Enhanced inputs lift every in-domain model family




**+0.225**

TF-IDF + LR gain

**+0.225**

RoBERTa gain

**Even a bag-of-words model reaches 0.84.**

**High enhanced-view performance is therefore not, by itself, evidence of robust reasoning.**

FIXED-CHECKPOINT CONTROL

# One checkpoint loses about 0.33 macro-F1 when the format changes


SAME

**RoBERTa-large checkpoint**

ONLY CHANGE

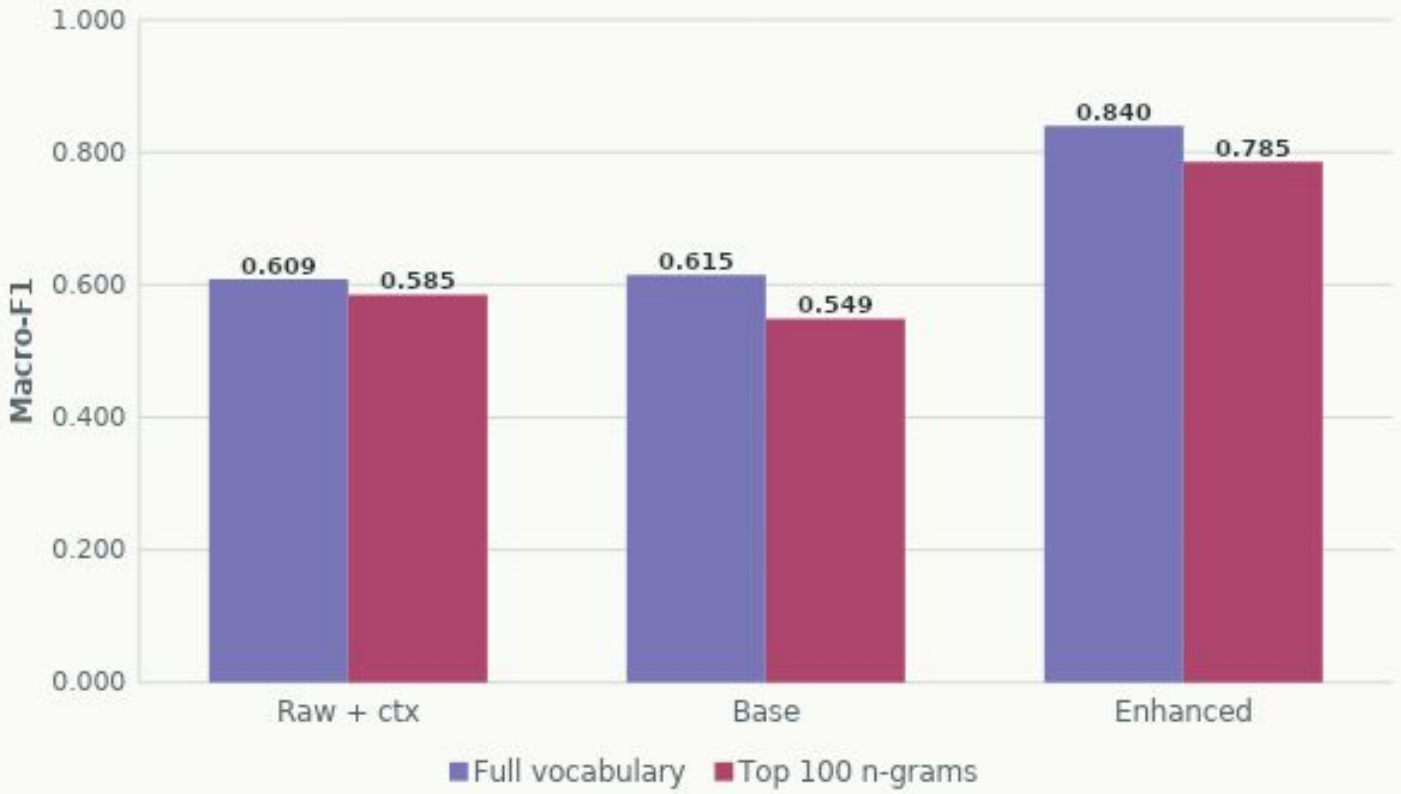
**enhanced text → base text**

**-0.329 / -0.382**

**The advantage is tied to the representation, not a representation-independent detector.**

LEAKAGE DIAGNOSTIC

# A small lexical feature set recovers most of the enhanced score




## Highest-weight enhanced cues

**push toward FALLACY**

problems · nature · there are · proves

**push toward NON-FALLACY**

reasoning · actually · actual · consequences

**RoBERTa enhanced attention probe**

**1.32×** fallacy rows

**1.68×** non-fallacy rows

Cue-attention lift, 12 selected rows per view
Routing diagnostic, not causal attribution

**Top 100 enhanced n-grams retain 93.5% of the full-vocabulary score.**

# Supervised systems do not transfer reliably


**Gemma 4 31B**

**0.93**

LOGIC macro-F1

**Only positive transfer-gap exception**

**CAVEAT**

LOGIC-binary is itself style-confounded. Read it as a corpus-transfer stress test, not a fair external benchmark.

CONCLUSION

# Naturalistic evaluation changes the model choice

**1**

## Treat enhanced rewrites as diagnostics

They reveal sensitivity, but their label-aware construction can introduce shortcuts.

**2**

## Run cross-view controls

Hold the checkpoint fixed and change the representation before claiming robust capability.

**3**

## Prefer raw or base evidence

Use inputs that resemble the discourse the detector will actually encounter.

SUBMITTED SYSTEM

**Gemma 4 31B**

**raw + context**

OFFICIAL BASE TRACK

**F1 0.736**

precision 0.745 / recall 0.737

**4th among ranked submissions**

**The benchmark design changed both the scientific conclusion and the submitted system.**

# Benchmarks for motivated and biased reasoning

## Does a model evaluate the same evidence differently when one interpretation helps it complete its task?

### A controlled benchmark

Hold the evidence fixed. Vary task pressure or incentives. Measure changes in interpretation, justification, and action.

### The lesson from fallacy detection

Test whether a monitor detects these shifts across unfamiliar tasks and writing styles, without relying on answer-revealing cues.

Motivation: Anthropic reports biased reasoning and recklessness in cyber incidents (2026).

**Interested in collaborating?**

I am looking for collaborators, mentors, and senior researchers.

dianamarkovakn@gmail.com

linkedin.com/in/diana-markova-178380306