Webis at Touché 2026

An Adversarial Courtroom Trial for Reddit Fallacy Detection

Christine Mathews Maximilian Heinrich Benno Stein

Bauhaus-Universität Weimar

Bauhaus-Universität Weimar

webis.de

Informal Fallacy Detection

Example

*“Everyone on Twitter is calling **Spiderman: Brand new day** a masterpiece. But speaking as someone who has curated a Spotify playlist with over 5,000 followers, binged every prestige HBO drama since 2010, and owns a vintage film camera, I can objectively say the narrative structure is garbage. Trust me, I know what good art looks like.”*

Does this argument commit a fallacy?

Informal Fallacy Detection

The same reasoning pattern, two verdicts

Fallacious use

*“Everyone on Twitter is calling **Spiderman: Brand new day** a masterpiece. But speaking as someone who has curated a Spotify playlist with over 5,000 followers, binged every prestige HBO drama since 2010, and owns a vintage film camera, I can objectively say the narrative structure is garbage. Trust me, I know what good art looks like..”*

Why: The credentials are real, but curating playlists and watching TV do not qualify you to judge a film's structure.

Non-fallacious use

“It’s almost like Valve do massive amounts of playtesting and know that cutting enemies that were too scary in VR works better than just slapping a difficulty selection bar on it.”

Why: Valve’s playtesting track record is directly relevant to their own design decision.

Both invoke expertise as evidence. What differs is whether that expertise is relevant to the specific claim — not the surface form.

Why This Task Is Hard

The same move can be a fallacy — or not

- For every kind of fallacy, the data also contains look-alikes that are fine: the same kind of move, but not a fallacy.
- So what decides it is a question: does this expertise apply to this claim? Someone has to ask it.

So, do today's systems ask that question?

The Gap

What existing systems leave on the table

- Supervised fallacy detectors transfer poorly across datasets.
- A fine-tuned classifier returns a label and a probability.
- Argumentation theory (Walton, Reed, Macagno): a scheme becomes fallacious when a critical question exposes a gap the argument cannot answer.
  - *Is the cited source genuinely an expert in the relevant domain?*
  - *Does their expertise apply to this specific claim?*

Can the decision itself be made inspectable?

The Idea

Deciding whether reasoning holds up is an adversarial question

- One side assembles the best case that the reasoning is flawed.
- The other assembles the best case that it is sound.
- A neutral party weighs both and decides.

We build this as a courtroom trial: prosecutor, defense, an expert witness, a judge, and an appeal.

Every record leaves a written transcript of the case for and against the label.

An Adversarial Courtroom Trial

① **A two-phase courtroom pipeline**
Adversarial trial (prosecutor, defense, judge) followed by an asymmetric appeal.

Experiments:

② **Main validation results**
Trial versus TF-IDF, fine-tuned BERT, and a zero-shot LLM on both text variants.

③ **Component ablations**
Which parts actually help, and which are just decoration.

The Courtroom

The main idea: two opposing sides argue about the comment — and a judge decides

Prosecutor

Argues: this comment is a fallacy.

Sees examples of known fallacies.

Defense

Argues: this comment is not a fallacy.

Sees examples of known non-fallacies.

Expert witness

A trained classifier (fine-tuned BERT). Gives its estimate of how likely a fallacy is.

Never votes.

Judge

Reads both sides and the estimate.

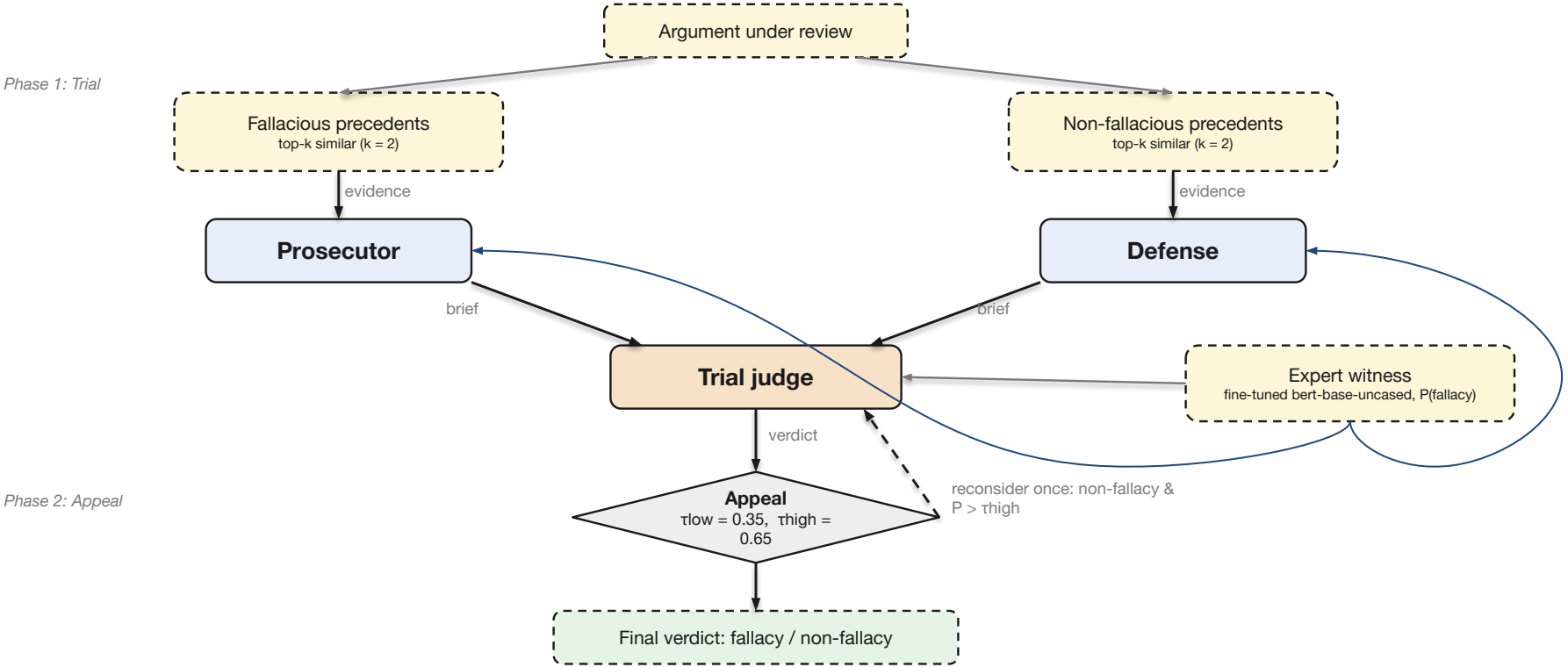
Issues the verdict.

Appeal. A “fallacy” verdict stands. A “not a fallacy” verdict the expert witness strongly disagrees with goes back to the judge — once.

Each advocate files one written brief and never sees the other’s. Prosecutor, defense and judge are all the same language model (deepseek-r1:14b, run locally); only the instructions differ.

System Overview

The trial pipeline



An Adversarial Courtroom Trial

① **A two-phase courtroom pipeline**
Adversarial trial (prosecutor, defense, judge) followed by an asymmetric appeal.

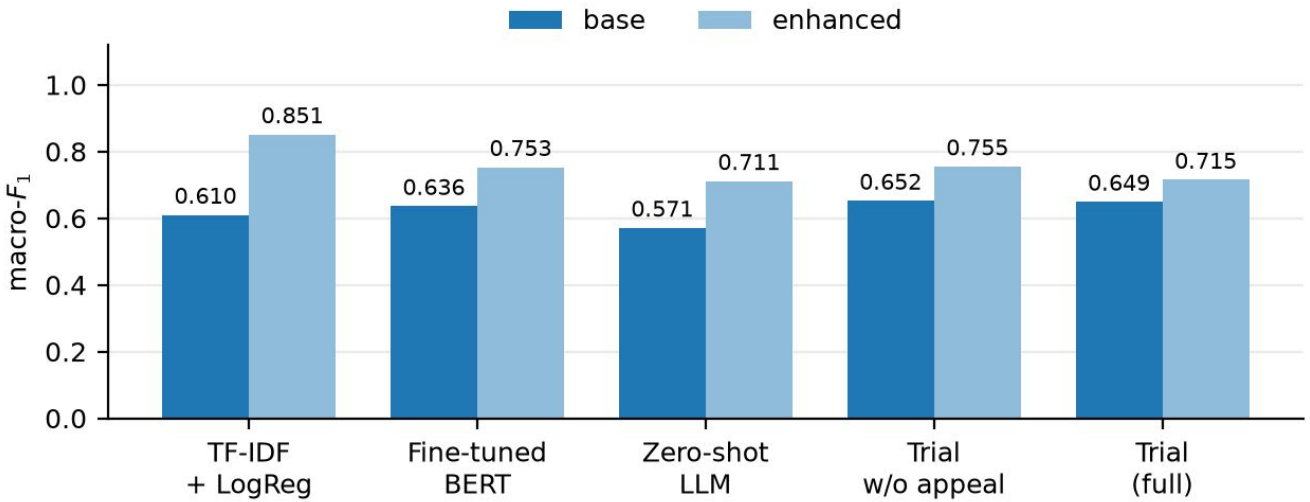
Experiments:

② **Main validation results**
Trial versus TF-IDF, fine-tuned BERT, and a zero-shot LLM on both text variants.

③ **Component ablations**
Which parts actually help, and which are just decoration.

Main Validation Results

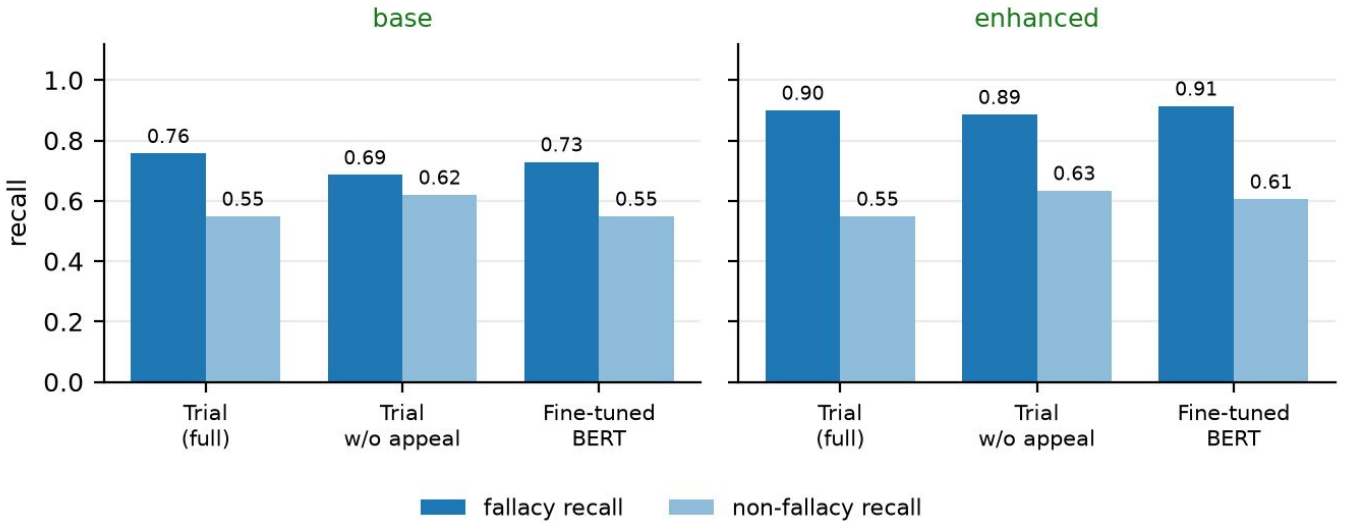
Macro-F1, n = 141 per variant



On `base` the trial narrowly leads the baselines (0.649); on `enhanced` a plain TF-IDF baseline wins by 13.6 pp. Note the trial scores higher without the appeal (0.652).

Per-Class Behavior

The trial over-predicts fallacy



- It catches most real fallacies (recall 0.757 / 0.900) but misses many real non-fallacies (0.549 on both).
- Switching the appeal off recovers non-fallacy recall (0.620 / 0.634): the appeal turns borderline “not a fallacy” decisions into “fallacy” ones.

An Adversarial Courtroom Trial

① **A two-phase courtroom pipeline**
Adversarial trial (prosecutor, defense, judge) followed by an asymmetric appeal.

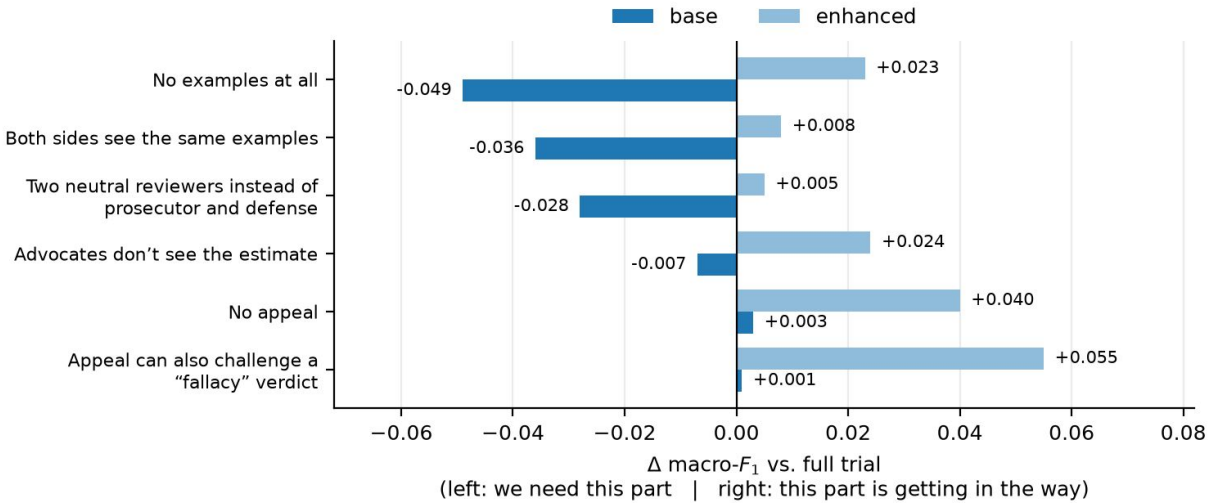
Experiments:

② **Main validation results**
Trial versus TF-IDF, fine-tuned BERT, and a zero-shot LLM on both text variants.

③ **Component ablations**
Which parts actually help, and which are just decoration.

Component Ablations

Each row removes or modifies exactly one component



Conclusion

① **We put each comment on trial**

A prosecutor argues it is a fallacy, a defense argues it is not. Each side gets examples that support its own position. A judge decides — and every comment leaves a written record of both sides.

② **The result depends on the text**

On the plain comments the trial leads the baselines: 0.649, against 0.636 for BERT and 0.610 for TF-IDF. On the rewritten comments TF-IDF wins, 0.851 against our 0.715.

③ **The appeal did not pay off**

Switching it off changes almost nothing on the plain comments (+0.3 pp) and improves the score by 4.0 pp on the rewritten ones. It cuts one kind of error by adding another.

Lessons That Generalize

1. Give each side only the examples that support it.
   This is the part that actually carries the courtroom on the harder text. It should work for any task where every positive example has a near-identical negative one.

2. Only let a classifier overrule a debate if you know which of the two is weaker.
   The appeal helps when the debate is the weak part and the classifier is reliable. When the classifier is already strong, the same rule makes things worse.

3. A courtroom only fits a yes/no question.
   Fallacy against not-a-fallacy works, because the two sides genuinely oppose each other. Eight classes that do not contradict one another would not.