Webis at Touché: An Adversarial Courtroom Trial for Reddit Fallacy Detection

Touché at CLEF 2026

Christine Mathews[1,*], Maximilian Heinrich[1] and Benno Stein[1]

[1]Bauhaus-Universität Weimar, Weimar, Germany

Abstract

Detecting fallacies in everyday arguments is harder than it sounds. The same reasoning move — citing an expert, pointing to a trend, drawing an analogy — can be valid in one context and misleading in another. We test a courtroom-style system for this problem. A prosecutor argues that the reasoning is flawed; a defense argues that it is sound. Both draw on similar labeled training cases. A judge reads both briefs, weighs an expert witness probability, and returns a verdict. If the judge returns `non-fallacy` while the expert witness still assigns high fallacy probability, the judge is asked to reconsider once. This design produces an inspectable record for every prediction, although we do not test whether those explanations are faithful. We evaluate on the Touché 2026 shared task for detecting fallacies in Reddit comments [1, 2]. On our validation split, the results are mixed: the courtroom system narrowly leads the baselines when surface cues are sparse, but TF-IDF and fine-tuned BERT overtake it when the text makes the argument structure explicit. The asymmetric appeal also hurts rather than helps. Overall, the courtroom layer appears useful only under specific conditions.

Keywords

Fallacy Detection, Multi-Agent Reasoning, Adversarial Prompting, Argumentation Schemes, Retrieval-Augmented Generation, Reddit

1. Introduction

Informal fallacies are defective arguments: they may look persuasive on the surface, but their reasoning fails under scrutiny. The difficulty is that the same reasoning pattern can be fallacious or non-fallacious depending on its use. An appeal to authority is fallacious when the authority is irrelevant or not credible; it is non-fallacious when the authority is credible and the question falls within their expertise. A generalization from examples is fallacious when the sample is too small or unrepresentative; it is non-fallacious when the sample is large and representative enough. In each pair, the surface form can look almost identical. What differs is whether the reasoning holds up.

The Touché 2026 shared task on informal fallacy detection [1, 2] reflects this directly: every fallacious training example is paired with a near-identical non-fallacious example that uses the same reasoning pattern. Table 1 illustrates this tension for an authority-based pair. In both comments, the speaker invokes an expert source. The left comment is fallacious because the cited expertise is not relevant to the specific claim; the right comment is non-fallacious because the expertise directly bears on the decision being defended. A system that detects surface patterns will struggle precisely where the data is most demanding.

Our approach starts from a simple observation: deciding whether a piece of reasoning survives scrutiny is an inherently adversarial question. One side assembles the best available case that the reasoning is flawed; the other assembles the best available case that it is sound; a neutral judge weighs both. If the prosecution's case is stronger, the argument fails. If the defense holds, it does not.

We operationalize this as a courtroom trial. An argument under review is put before a *prosecutor* and a *defense*. The prosecutor receives the most similar known-fallacious arguments from the training

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

christine.mathews@uni-weimar.de (C. Mathews); maximilian.heinrich@uni-weimar.de (M. Heinrich); benno.stein@uni-weimar.de (B. Stein)

0009-0006-3041-0002 (C. Mathews); 0000-0001-5450-8203 (M. Heinrich); 0000-0001-9033-2217 (B. Stein)

Table 1
Two authority-based instances from the paired training data: one fallacy and one non-fallacy. Both invoke expertise as evidence; what differs is whether that expertise is relevant to the specific claim being made. The *Why* row gives our explanation of why the instance is fallacious or non-fallacious.

	Fallacy	**Non-Fallacy**
Text	"People can say fans are overreacting all they want. But Damien Woody — a two-time Super Bowl winner, Pro Bowl left guard, ex-Patriot — agrees that Gronk is washed up."	"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	Damien Woody is a former NFL lineman, and the comment treats his agreement as evidence that "Gronk", i.e., Rob Gronkowski, is physically declining. Woody's standing as a former player is real, but it does not by itself make him a reliable judge of whether Gronkowski is physically washed up.	Valve is the game studio that developed the virtual-reality (VR) title under discussion, and the comment treats their extensive playtesting as evidence that a specific design choice — cutting overly scary VR enemies instead of adding a difficulty setting — was the right one. Here the expertise is directly relevant: Valve ran the playtests, so their authority bears on exactly the claim being made.

set; the defense receives the most similar known-non-fallacious arguments. Each side files a written *brief*. A neutral *judge* reads both briefs, both precedent pools, and a probability from a fine-tuned *expert witness*, then issues a *verdict*: `fallacy` or `non-fallacy`. If the judge returns `non-fallacy` while the expert witness remains confident that the argument is fallacious, an *appeal* asks the judge to reconsider. The result is a decision process that produces a structured argument for and against every label, not just the label itself.

We use this setup to answer three questions. First, does structured adversarial advocacy improve binary fallacy detection over simpler baselines? Second, which components account for the effectiveness? Third, does the appeal reduce false negatives without inflating false positives?

The remainder of this paper is organized as follows. Section 2 situates the approach in the literature. Section 3 describes the dataset and pipeline. Section 4 reports results and ablations. Section 5 presents a qualitative analysis of the trial transcripts and the design lessons we draw from them. Section 6 draws out the limitations.

2. Related Work

Fallacy detection in natural language. Computational work on fallacy detection has moved from rule-based surface-marker methods toward supervised classification on labeled corpora [3, 4, 5, 6]. A persistent finding is that fallacy detectors transfer poorly across datasets because the target concept is not captured by lexical surface alone. The Touché 2026 task [1, 2] sharpens this challenge: every fallacious training example is paired with a near-identical non-fallacious example that uses the same reasoning pattern. A model that detects surface patterns will score at chance on these pairs. The task instead requires a judgment about whether the reasoning holds up in the specific instance under review.

Argumentation schemes and critical questions. The argumentation-theoretic tradition provides the clearest account of what distinguishes a fallacious from a non-fallacious use of a reasoning pattern. Walton, Reed, and Macagno reconstruct informal fallacies as *defeasible argument schemes* annotated with *critical questions* [7, 8, 9]. For an appeal to authority, the critical questions include: *Is the cited source genuinely an expert in the relevant domain? Does their expertise apply to this specific claim? Do they have a personal interest that could bias the conclusion?* An appeal to authority is not fallacious by itself; it becomes fallacious when one of these questions exposes a gap the argument cannot answer.

This view motivates the courtroom framing. The prosecutor's job is to expose where the reasoning fails a critical question; the defense's job is to show that the question is answered. Our system operationalizes this division of labor. A shared glossary of the eight fallacy classes gives both advocates a common vocabulary, but the advocates write free-form briefs rather than ticking off a fixed checklist.

Multi-agent reasoning and adversarial debate. Adversarial debate between LLM agents has been studied as a mechanism for improving reasoning [10, 11, 12, 13]. These systems typically involve agents responding to each other across multiple rounds under a judge or aggregator. We are not aware of prior work that combines a courtroom framing for argument classification with role-specific retrieval, single-brief adversarial advocacy, and a deliberative judge. Our design also differs structurally: each advocate files a single brief without seeing the other's, and each side is seeded with precedents selected specifically for its position.

3. Experimental Setup

We describe the dataset, the two-phase courtroom pipeline, the baselines, and the reproducibility setup.

3.1. Touché 2026 Dataset and Tasks

The Touché 2026 shared task on informal fallacy detection [1, 2] builds on the Reddit fallacy corpus of Sahai et al. [14], releasing 938 labeled training records and 233 unlabeled test records. Each record is provided in two text variants. The `base` variant merges the original comment with its parent comment and discussion topic into a single, more self-contained text. The `enhanced` variant rewrites the base text so that the underlying argument scheme — its goal and basis — is easier to identify. Where the record is fallacious, the rewrite also makes the fallacious reasoning pattern more explicit and central. The rewrite is constrained to stay natural and to preserve the original tone and stance. Every record carries a binary fallacy label (`fallacy_exists`, 0 or 1), and fallacious records are further labeled with one of eight fallacy classes: `authority`, `black-white`, `hasty_generalization`, `natural`, `population`, `slippery_slope`, `tradition`, and `worse_problems`. Non-fallacious records additionally carry a `resembles_fallacy` field that records which fallacy class the non-fallacious argument *resembles*, completing the twin structure described in Section 1. Every record also carries argument-scheme annotations (`argument_goal`, `argument_basis`) that are not used by this paper, which focuses on the binary fallacy label.

We focus on binary fallacy detection and evaluate on both text variants as separate tracks. Because the hidden test labels are not available to participants for error analysis or ablation studies, we form a held-out validation split of 15% from the labeled training data, stratified on the binary label. Official hidden-test scores are reported separately in Section 4.2.

3.2. System Overview

The trial pipeline processes each argument under review in two phases. Figure 1 shows the data flow; we describe each component in Sections 3.3–3.6.

Adversarial Trial. Before the trial begins, we assemble a small precedent pool for each side. A precedent is an earlier training argument for which we already know whether the reasoning was fallacious, retrieved by similarity to the argument under review. The prosecution pool holds known-fallacious examples, the defense pool known-non-fallacious ones. The prosecutor and defense each receive the argument under review, their precedent pool, and the expert witness probability. They work independently; neither sees the other's brief. The prosecutor identifies the feature that makes the reasoning fail. The defense explains why the reasoning is non-fallacious despite surface resemblance to a fallacy. A judge then reads both briefs, both precedent pools, and the expert witness probability, writes a reasoned opinion, and issues a verdict: `fallacy` or `non-fallacy`.

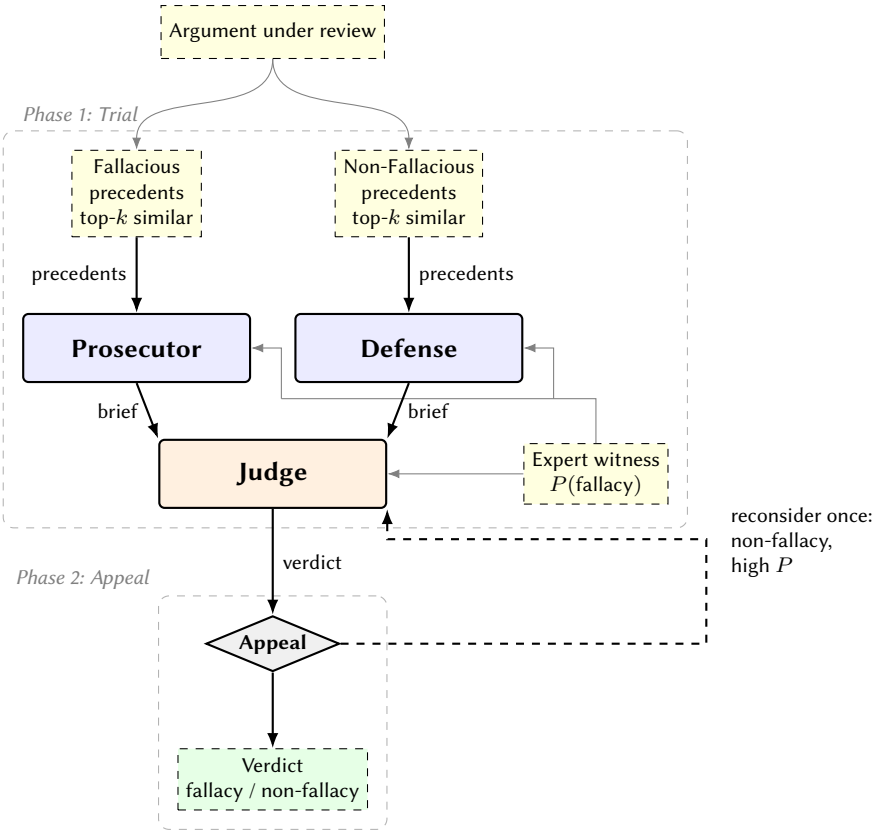

Figure 1: The trial pipeline. The prosecution is seeded with similar fallacious training examples; the defense with similar non-fallacious examples. Both advocates file briefs independently and neither sees the other's. The judge reads both briefs and the expert witness probability and issues a verdict. The asymmetric appeal is a deterministic decision node: it accepts a `fallacy` verdict outright and, on `non-fallacy`, uses the expert witness probability to decide whether to reconsider. If $P > \tau_{\text{high}}$ the judge is re-invoked once and may revise; if $P < \tau_{\text{low}}$ the non-fallacy verdict stands unconditionally. Yellow boxes are data; blue boxes are advocate roles; the orange box is the judge; the gray diamond is the decision rule.

Asymmetric Appeal. After the trial, `fallacy` verdicts stand immediately. When the judge returns `non-fallacy`, the expert witness probability is checked. If it is low, the non-fallacy verdict stands. If it exceeds $\tau_{\text{high}} = 0.65$ — meaning the expert witness strongly disagrees with the judge — the appeal fires exactly once: the judge is shown the probability as explicit additional evidence and asked to reconsider. The judge may change its verdict to `fallacy` or stand by the non-fallacy verdict; either way, its decision at that point is final and the appeal does not fire again. Fallacy verdicts are never contested. The full rule is in Section 3.6.

3.3. Role-Specific Precedent Retrieval

For each argument under review, we retrieve two separate pools of precedents from the training set, one per advocate. The full training corpus is encoded with `sentence-transformers/all-MiniLM-L6-v2` [15, 16] and L2-normalized; the resulting matrix is cached on disk per text variant. At inference time, the argument under review is encoded into the same space, and cosine similarities are computed against every training record. The *prosecution pool* contains the top-k most similar fallacious precedents; the *defense pool* contains the top-k most similar non-fallacious precedents. The judge receives a smaller draw ($k' < k$) from both pools. Each precedent text is truncated to 300 characters in the prompt. Seeding each advocate with ground-truth precedents from its own side mirrors the adversarial structure of the trial: the prosecutor argues from genuine fallacies, the defense from genuine non-fallacious uses of the same reasoning patterns.

```
def appeal(verdict, p, briefs, precedents):
    if verdict == "fallacy":
        return "fallacy"                        # fallacy verdicts stand
    if p < τ_low:
        return "non-fallacy"                    # non-fallacy stands
    if p > τ_high:
        revised = judge(briefs, precedents,     # one re-invocation
                        testimony=p)
        return revised.verdict                  # judge has the last word
    return "non-fallacy"                        # middle-band non-fallacy
```

Figure 2: The asymmetric appeal rule — the Phase 2 decision node in Figure 1. After the judge returns a verdict, the rule decides whether to accept it or send it back for a single reconsideration: fallacy verdicts always stand, and a non-fallacy verdict is reappealed only when the expert witness probability exceeds τ_{high}.

3.4. The Expert Witness

The expert witness is a fine-tuned `bert-base-uncased` trained on the binary fallacy-detection split for each text variant. Its softmax output $P(\text{fallacy})$ is inserted into the prompts of the prosecutor, the defense, and the judge as expert testimony. If the appeal fires, the same probability is cited again during reconsideration. The probability is evidence to be weighed alongside the advocates' arguments, not a deciding vote. We use the same architecture and training procedure for both text variants and keep all hyperparameters identical (Appendix A).

3.5. The Trial

All three courtroom roles — prosecutor, defense, and judge — are played by the same `deepseek-r1:14b` model with role-specific system prompts. Using one model removes model capability as a confound: differences between the prosecution and defense should come from the precedents and role, not from different underlying models. Each advocate reports a confidence level (high, medium, or low), which the judge weighs alongside the two arguments and the expert witness probability.

3.6. Asymmetric Appeal

The appeal trigger is deterministic. It takes the trial verdict and the expert witness probability $P(\text{fallacy})$; if the trigger condition is met, it re-invokes the judge once and returns the revised verdict. We use two thresholds, $\tau_{\text{low}} = 0.35$ and $\tau_{\text{high}} = 0.65$, around the expert witness's decision boundary; Figure 2 states the rule in full.

These thresholds were chosen as an exploratory conservative band around the expert witness's decision boundary on the validation split. Because the same split is used for reporting, the appeal results should be read as diagnostic rather than as an independently tuned estimate; the ablation that removes the appeal entirely (Section 4.3) isolates the mechanism's overall contribution on this split.

On reconsideration, the same judge is asked again. The second call receives the same context as the original verdict — the argument under review, both briefs, both precedent pools, and $P(\text{fallacy})$ — with only the appellate testimony added; it is a fresh, stateless call that does not see the judge's own earlier opinion. Holding the judge's system prompt fixed means the decision criterion is identical across the two passes, so any change of verdict is attributable to the added expert-witness testimony.

3.7. Baselines

We compare the trial system to three baselines, each isolating a different question. TF-IDF + LogReg is a lower bound: 1–2 grams, twenty thousand features, balanced class weights, fitted on the same training split. Fine-tuned BERT is the same expert witness used inside the trial, evaluated as a stand-alone

Table 2
Validation macro-F_1 for binary fallacy detection, on the `base` and `enhanced` text variants of the Touché 2026 Reddit data ($n = 141$ records per variant). The trial system is reported with and without the asymmetric appeal to isolate its contribution. Per-class precision and recall are reported in Table 3. The best score per variant is in bold.

System	base	enhanced
TF-IDF + LogReg	0.610	**0.851**
Fine-tuned BERT (expert witness, stand-alone)	0.636	0.753
Zero-shot LLM (`deepseek-r1:14b`)	0.571	0.711
Trial without appeal	**0.652**	0.755
Trial (full pipeline)	0.649	0.715

binary classifier; it shows what a strong supervised model achieves without the courtroom scaffolding. Zero-shot LLM uses `deepseek-r1:14b` prompted with the eight fallacy class definitions and asked to return a binary label without exemplars; it shows what the same LLM does on its own, without prosecutor, defense, judge, or precedents. All baselines are run on both `base` and `enhanced` as separate tracks.

3.8. Reproducibility

A fixed seed of 42 makes the data split, the cached training embeddings used for retrieval, and the BERT expert witness deterministic. The LLM calls are not deterministic: the prosecutor and defense run at temperature 0.6, and the judge at 0.4, against a served `deepseek-r1:14b` endpoint. The trial numbers therefore reflect one run. The reported deltas mix component effects with run-to-run stochasticity; isolating them would require a multi-seed protocol. Whether that is feasible within a realistic compute budget — roughly twenty hours per full sweep (Section 4.6) — is left to future work. Hyperparameters and prompt templates are listed in Appendix A.

4. Results

We report detailed results on the validation split (15% of the labeled training records, stratified on the binary label), because the official test labels are not released to participants for error analysis. All validation numbers in this section are macro-F_1 per text variant (`base`, `enhanced`) unless noted otherwise; pp denotes percentage points. Components and ablations are evaluated under fixed seeds and identical splits, so differences do not stem from data shuffling.

4.1. Main Validation Results

Table 2 reports the validation scores across the three baselines and the full trial system. The trial is also reported with and without the asymmetric appeal, to separate the contribution of the appeal from the rest of the pipeline.

The results split sharply by text variant. On `base`, the full trial leads the baselines (0.649), ahead of fine-tuned BERT (0.636) and TF-IDF (0.610). On `enhanced`, the ordering reverses: TF-IDF (0.851) and fine-tuned BERT (0.753) both beat the full trial (0.715), leaving a 13.6 pp gap to TF-IDF.

The likely reason is that `enhanced` text contains explicit argument-structure cues. Bag-of-words features can capture many of these cues directly, while the trial can still be persuaded by a strong defense brief. The zero-shot LLM reinforces this point. It is the weakest system on `base` (0.571), but lands within 0.4 pp of the full trial on `enhanced` (0.711 vs. 0.715). On `enhanced` text, the pipeline barely improves over asking the same model directly. The appeal also fails to add value: the trial without appeal slightly outperforms the full trial on both variants ($+0.003$ on `base`, $+0.040$ on `enhanced`).

Table 3
Per-class validation results for binary fallacy detection. Precision (P), recall (R), and F_1 are reported separately for the *fallacy* and *non-fallacy* class on each text variant ($n = 141$ per variant). The best F_1 in each class column is in bold. The trial over-predicts the fallacy class (high fallacy recall, low non-fallacy recall), most strongly on `enhanced`; the asymmetric appeal sharpens this imbalance, which is why disabling it (*Trial without appeal*) raises non-fallacy recall.

	base						**enhanced**					
	Fallacy			**Non-Fallacy**			**Fallacy**			**Non-Fallacy**		
System	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF + LogReg	0.609	0.600	0.604	0.611	0.620	0.615	0.818	0.900	**0.857**	0.891	0.803	**0.844**
Fine-tuned BERT	0.614	0.729	0.667	0.672	0.549	0.605	0.696	0.914	0.790	0.878	0.606	0.717
Zero-shot LLM	0.560	0.671	0.610	0.596	0.479	0.531	0.667	0.857	0.750	0.804	0.577	0.672
Trial without appeal	0.640	0.686	0.662	0.667	0.620	**0.642**	0.705	0.886	0.785	0.849	0.634	0.726
Trial (full)	0.624	0.757	**0.684**	0.696	0.549	0.614	0.663	0.900	0.764	0.848	0.549	0.667

The per-class breakdown (Table 3) shows where the trial's errors concentrate. On both variants the trial reaches high recall on the fallacy class (0.757 on `base`, 0.900 on `enhanced`) but low recall on the non-fallacy class (0.549 on both): it systematically over-predicts fallacy. Disabling the appeal recovers non-fallacy recall (0.620 on `base`, 0.634 on `enhanced`), confirming that the appeal contributes to the imbalance by converting borderline non-fallacy verdicts into fallacy verdicts. On this validation split, the third research question has a negative answer: the appeal cuts false negatives, but at the price of the false positives it was designed to avoid.

Bootstrap intervals confirm how modest the validation gaps are. With 141 records per variant, we compute 95% intervals over 10,000 resamples. On `base` the trial leads at 0.649 (BERT 0.636, TF-IDF 0.610), but the intervals overlap, so the lead should be read as directional rather than decisive. On `enhanced`, the gap is larger: TF-IDF scores 0.851 with a 95% interval of $[0.787, 0.907]$, while the trial scores 0.715 $[0.636, 0.789]$; BERT also beats the trial at 0.753. These intervals capture sampling variation over the 141 records only: they hold each system's predictions fixed and therefore do not reflect the run-to-run stochasticity of the temperature-sampled LLM calls, which are drawn from a single run (Section 3.8). For the trial they thus understate the true uncertainty.

4.2. Official Test Leaderboard

Task 1 was evaluated on the TIRA platform [17]. On the hidden test split, our `Webis-CourtroomTrial` run scores 0.704 macro-F_1 on `base` and 0.841 macro-F_1 on `enhanced`. Both scores place the system in the lower part of the field rather than near the top. On `base` it lands mid-table, ahead of the organizer baseline and three further submitted runs but well behind the strongest run at 0.807 macro-F_1. On `enhanced` it trails every ranked submission except the organizer baseline; the competitive systems there cluster between 0.875 and 0.957 macro-F_1. The gap is expected: the courtroom pipeline is not optimized as a leaderboard classifier; it tests how far a reasoning-based, inspectable decision process can go on the same task.

4.3. Component Ablation

Table 4 reports six ablations. Each row removes or modifies exactly one component of the full trial; the rest of the pipeline is unchanged. The ablation pattern is variant-dependent and tells two different stories. The per-component deltas on `base` are small next to the bootstrap intervals of Section 4.1 and read as a consistent direction rather than individually decisive; the larger `enhanced` effects are firmer.

On `base`, where the trial posts its best relative result, the load-bearing component is role-specific retrieval (Table 4). Dropping precedents entirely costs 4.9 pp. Giving both advocates the same mixed pool costs 3.6 pp. Removing the role split — replacing prosecutor and defense with two neutral evaluators

Table 4
Component ablations on the validation set. Each row changes one component while keeping the rest of the pipeline fixed. For each variant we report absolute macro-F_1 and Δ, the absolute macro-F_1 difference from the full-trial reference row; negative values indicate that the removed or modified component helped, and positive values indicate that it hurt.

	base		enhanced	
Ablation	F_1	Δ	F_1	Δ
Full trial (reference)	0.649	—	0.715	—
1. *No expert witness in prompt*: $P(\text{fallacy})$ removed from advocate and judge prompts (appeal unchanged)	0.642	-0.007	0.739	$+0.024$
2. *No appeal*: judge's verdict is final; no reconsideration	0.652	$+0.003$	0.755	$+0.040$
3. *Symmetric appeal*: expert witness may also contest fallacy verdicts, not only non-fallacy verdicts	0.650	$+0.001$	0.770	$+0.055$
4. *Neutral roles*: replace prosecutor and defense with two neutral evaluators	0.621	-0.028	0.720	$+0.005$
5. *No role-specific retrieval*: both advocates see both pools	0.613	-0.036	0.723	$+0.008$
6. *No precedents*: drop retrieval entirely; advocates argue from the argument under review and $P(\text{fallacy})$ only	0.600	-0.049	0.738	$+0.023$

— costs 2.8 pp. The expert witness probability and the appeal are nearly neutral: removing $P(\text{fallacy})$ costs only 0.7 pp, dropping the appeal is essentially flat ($+0.3$ pp), and a symmetric appeal matches the asymmetric default to within 0.1 pp. On `base`, the courtroom adds value mainly through retrieval: two polarity-stratified precedent pools, each routed to the advocate who can use it.

On `enhanced`, every ablation *improves* the score. The biggest gains come from weakening the appeal: a symmetric appeal gains 5.5 pp and dropping the appeal entirely gains 4.0 pp, while removing $P(\text{fallacy})$ from the prompts gains 2.4 pp. Retrieval changes matter less ($+0.8$ pp for a shared mixed pool, $+2.3$ pp for no precedents). The direction reverses because the expert witness becomes over-confident on explicit surface markers. The asymmetric appeal is meant to catch cases where the expert witness disagrees with the defense; here, it instead pulls correct non-fallacy verdicts toward fallacy. The burden-of-proof asymmetry breaks down once the expert witness itself is the source of error.

4.4. Where the Trial Wins and Loses

The aggregate macro-F_1 scores mask different behavior across the two variants. On `base`, the trial and BERT disagree on 24 of the 141 records (17%); the trial is right in 13 of these and BERT in 11, which yields the trial's modest 1.3 pp edge. On `enhanced` they disagree on 11 records (8%), and the split turns against the trial: BERT is right in 8, the trial in 3, a 3.8 pp loss, in line with the appeal-related ablations on this variant (Section 4.3). We examine these disagreements and the transcripts behind them in Section 5.1.

4.5. Variant Effect

The `enhanced` variant yields higher scores than `base` for every system because it introduces explicit argument structure and sometimes near-explicit label cues. Table 2 shows that BERT gains $+11.7$ pp from enhancement, while the full trial gains only $+6.6$ pp. This is surprising: the advocates should benefit from explicit claim–support structure. Instead, the trial benefits less than the expert witness it wraps. This matches the ablations: on `enhanced`, the deliberation layer adds noise that partially cancels the gains from the more explicit text.

4.6. Cost

A single record incurs three LLM calls in the standard path (prosecutor, defense, judge) and a fourth when the appeal fires. On the validation split the appeal fired on 11.3% of `base` records and 6.4% of `enhanced` records, so the mean cost is just over three calls per record. With a single sequential `deepseek-r1:14b` endpoint served by Ollama, this corresponds to roughly 25 seconds per record.

5. Qualitative Analysis

The aggregate scores in Section 4 tell only part of the story. The trial also produces a structured transcript for each record. We use these transcripts to inspect the appeal mechanism, the disagreements with BERT, and the design lessons that survive the mixed quantitative result.

5.1. Trial Transcripts

Each transcript contains the prosecution brief, the defense brief, the judge's opinion, and the appeal outcome where applicable. This is the main practical advantage of the courtroom framing: each prediction comes with reasons for the fallacy reading, reasons for the non-fallacy reading, and the judge's account of why one side prevailed. We do not independently evaluate the faithfulness or quality of these explanations, so we treat them as inspectable diagnostic artifacts rather than validated explanations. Appendix C reproduces five validation records in full. They show the main behaviors: the trial correcting BERT, the appeal flipping wrong non-fallacy verdicts into correct `fallacy` verdicts, and the recurring failure mode on `enhanced`, where the judge accepts a persuasive defense despite a correct expert witness signal.

The appeal examples clarify why the aggregate effect is weak. The appeal fires on 11.3% of `base` records and 6.4% of `enhanced` records, and individual interventions can be correct (Transcripts B and C). On net, however, dropping the appeal changes macro-F_1 by $+0.3$ pp on `base` and $+4.0$ pp on `enhanced` (Section 4.3), so the mechanism is at best neutral and often harmful.

5.2. Disagreement With BERT

The trial and the stand-alone BERT classifier disagree on 17% of `base` records and 8% of `enhanced` records. On `base`, the split slightly favors the trial (13 correct vs. 11 for BERT). On `enhanced`, it turns against the trial (3 correct vs. 8 for BERT). Inspection of those cases suggests the recurring failure mode is that the defense persuades the judge to accept a marginal non-fallacy reading where the gold label is `fallacy` and BERT correctly flags it.

5.3. Lessons That Generalize

Three lessons may transfer beyond this dataset, though they remain to be tested elsewhere. First, role-specific retrieval is the useful part of the courtroom design on the harder `base` variant. Second, an asymmetric expert-witness-driven appeal is not automatically safe; it must depend on the reliability of both the expert witness and the deliberation layer. Third, the courtroom framing buys explanatory clarity mainly when the categories are dialectically opposed, as in fallacy vs. non-fallacy. It does not naturally extend to arbitrary multi-class classification.

6. Conclusion

We presented a courtroom-styled adversarial pipeline for binary fallacy detection on Reddit comments, evaluated on the Touché 2026 shared task. A prosecutor and a defense file independent briefs from polarity-stratified precedent pools. A neutral judge weighs both briefs alongside an expert witness probability and issues a verdict. An asymmetric appeal contests `non-fallacy` verdicts when the expert witness strongly disagrees with the judge.

The validation results are mixed and variant-dependent. On `base`, the full trial narrowly leads the baselines, with role-specific retrieval as the clearest load-bearing component. On `enhanced`, the picture inverts: TF-IDF and BERT outperform the trial, and the appeal mechanism hurts rather than helps. The asymmetric appeal does reduce false negatives, but at the cost of false positives, so its net contribution is zero or negative on this validation split.

These findings limit the claim we can make. The pipeline is evaluated on a single dataset. All LLM-based numbers come from one non-zero-temperature run. The fixed appeal thresholds are diagnostic on the validation split, not an independently tuned estimate. The current system also addresses binary fallacy detection rather than fine-grained fallacy classification. The most direct next steps are a multi-seed evaluation, an appeal rule calibrated to the expert witness, and a separate test of explanation faithfulness.

We therefore view the method as useful but conditional: it can help when surface cues are sparse and the contrast between fallacy and non-fallacy requires adversarial scrutiny, but it can over-debate cases whose explicit surface cues already make them easy for simpler models. Its main practical advantage is diagnostic: each prediction produces an inspectable transcript with reasons on both sides of the verdict. Whether those reasons are faithful explanations rather than useful diagnostic artifacts remains to be tested.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Opus (Anthropic) and GPT-5 (OpenAI) in order to: Drafting content, Paraphrase and reword, Improve writing style, Abstract drafting, Grammar and spelling check, Formatting assistance, Content enhancement. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 17th International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[2] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems — extended version, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[3] I. Habernal, H. Wachsmuth, I. Gurevych, B. Stein, The argument reasoning comprehension task: Identification and reconstruction of implicit warrants, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1930–1940. URL: https://aclanthology.org/N18-1175/.

[4] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, B. Schölkopf, Logical fallacy detection, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 7180–7198. URL: https://aclanthology.org/2022.findings-emnlp.532/.

[5] P. Goffredo, S. Haddadan, V. Vorakitphan, E. Cabrio, S. Villata, Fallacious argument classification in political debates, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, 2022, pp. 4143–4149. URL: https://www.ijcai.org/proceedings/2022/575.

[6] Z. Sourati, V. P. P. Venkatesh, D. Deshpande, H. Rawlani, F. Ilievski, H. Sandlin, A. Mermoud, Robust and explainable identification of logical fallacies in natural language arguments, Knowledge-Based Systems 266 (2023) 110418. URL: https://doi.org/10.1016/j.knosys.2023.110418. doi:10.1016/j.knosys.2023.110418.

[7] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, Cambridge University Press, 2008.

[8] D. Walton, Why fallacies appear to be better arguments than they are, Informal Logic 30 (2010) 159–184.

[9] F. Macagno, D. Walton, C. Reed, Argumentation schemes: History, classifications, and computational applications, IfCoLog Journal of Logics and their Applications 4 (2017) 2493–2556.

[10] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Forty-First International Conference on Machine Learning, 2024, pp. 1–21. URL: https://arxiv.org/abs/2305.14325.

[11] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, S. Shi, Encouraging divergent thinking in large language models through multi-agent debate, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 1–16. URL: https://arxiv.org/abs/2305.19118.

[12] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, E. Perez, Debating with more persuasive LLMs leads to more truthful answers, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 1–27. URL: https://arxiv.org/abs/2402.06782.

[13] J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar, S. R. Bowman, Debate helps supervise unreliable experts, arXiv:2311.08702, 2023. URL: https://arxiv.org/abs/2311.08702.

[14] S. Sahai, O. Balalau, R. Horincar, Breaking down the invisible wall of informal fallacies in online discussions, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 644–657. URL: https://aclanthology.org/2021.acl-long.53/. doi:10.18653/v1/2021.acl-long.53.

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: Advances in Neural Information Processing Systems 33, 2020, pp. 5776–5788. URL: https://papers.nips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[16] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/.

[17] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

Table 5
Configuration of trial components.

Component	Configuration
Expert witness (BERT)	
Base model	`bert-base-uncased`
Variants	one checkpoint per text variant (`base`, `enhanced`)
Optimizer	AdamW
Learning rate	2×10^{-5}
Epochs	5 with early stopping on val macro-F_1
Batch size	16
Max length	512 tokens
Class weighting	balanced cross-entropy
Probability source	softmax output, inserted into prompts as $P(\text{fallacy})$
Retrieval	
Encoder	`sentence-transformers/all-MiniLM-L6-v2`
Normalization	L2-normalized embeddings
Similarity metric	cosine
Per advocate (k)	2 most similar of matching polarity
Per judge side (k')	1 most similar per polarity
Per-precedent truncation	300 characters
Cache	`role_index_{variant}.npz`
Trial (prosecutor, defense, judge)	
Model (all three roles)	`deepseek-r1:14b` via Ollama
Temperature (advocates)	0.6
Temperature (judge)	0.4
Max tokens per call	1200
Output format	JSON object, extracted post-hoc after stripping `<think>` block
Concurrency	sequential (single Ollama instance)
Asymmetric appeal	
τ_{low}	0.35 ($P(\text{fallacy})$ below this on a non-fallacy verdict $\rightarrow$ verdict stands)
τ_{high}	0.65 ($P(\text{fallacy})$ above this on a non-fallacy verdict $\rightarrow$ judge reconsiders once)
Reconsideration	at most one re-invocation of the judge with appellate testimony appended
Reproducibility	
Random seed	42 (Python, NumPy, PyTorch)
Validation split	15%, stratified on binary label
Train/test loader	`data_utils.load_train_test`

A. Hyperparameters

Table 5 lists the configuration of every component of the trial pipeline.

B. Prompt Templates

This appendix reproduces the system prompts used by the three courtroom roles and the appellate testimony injected into the judge during a Phase 2 reconsideration. All three roles share the fallacy glossary shown below as context (they are not required to name a fallacy class in the output).

The blocks are reproduced verbatim in the operational wording used at run time, which predates the paper's final terminology. Within them the *expert witness* is named `classifier`, the *fallacy class* surfaces as `suspected_type` or `pattern`, the verdict metaphors `acquittal`/`conviction` stand for `non-fallacy`/`fallacy`, and the retrieved *precedents* are referred to as *evidence* — a word the paper otherwise reserves for the expert witness probability. These literal strings are left unchanged so that

the appendix matches the system that produced the results — the paper's prose follows the glossary terminology throughout.

Shared fallacy glossary.

```
Reasoning pattern reference (context; you are NOT required to name a type):
  - authority: appeal to authority — fallacious when the authority is
    irrelevant or contested
  - black-white: false dilemma — fallacious when real alternatives are
    suppressed
  - hasty_generalization: over-broad conclusion from an unrepresentative
    sample
  - natural: appeal to nature — fallacious when naturalness does not
    confer the claimed property
  - population: appeal to popularity — fallacious when popularity doesn't
    track truth
  - slippery_slope: chain of harms — fallacious when the causal chain is
    unsupported
  - tradition: appeal to tradition — fallacious when age alone is the
    only reason given
  - worse_problems: whataboutism — fallacious when worse-problem
    existence doesn't defeat this concern
```

Prosecutor system prompt.

```
You are the PROSECUTOR in a fallacy trial. Your job is to argue,
persuasively and concretely, that the text on trial DOES commit a
reasoning fallacy.

You will be given:
  1. The argument under review.
  2. An expert witness probability P(fallacy) from a fine-tuned classifier —
     treat it as expert testimony, not a verdict.
  3. Precedents: real examples from the training set that were judged
     FALLACIOUS and are semantically similar to this text. Use them as
     evidence.

[shared fallacy glossary]

Your task:
  1. Identify the specific feature of the text that makes its reasoning
     fail.
  2. Explain WHY the reasoning is fallacious, not merely what pattern
     it resembles.
  3. Engage with the expert witness probability as evidence.
  4. Rate your confidence honestly.

Think step by step, then output a single JSON object (no prose after it):
{
  "is_fallacy": true,
  "tell": "<one sentence: the specific feature that makes this a fallacy>",
  "suspected_type": "<optional: the pattern this most resembles>",
```

```
  "confidence": "high" | "medium" | "low",
  "brief": "<2-4 sentences making the prosecution case>"
}
```

Defense system prompt.

```
You are the DEFENSE COUNSEL in a fallacy trial. Your job is to argue,
persuasively and concretely, that the text on trial does NOT commit a
reasoning fallacy and is in fact a non-fallacious use of a reasoning
pattern.

You will be given:
  1. The argument under review.
  2. An expert witness probability P(fallacy) from a fine-tuned classifier —
     treat it as expert testimony, not a verdict. A high probability does
     not by itself mean the argument is fallacious.
  3. Precedents: real examples from the training set that were judged
     NON-FALLACIOUS and are semantically similar to this text. Use them
     to show the text belongs in this category.

[shared fallacy glossary]

Your task:
  1. Identify the specific feature of the text that shows the reasoning
     is non-fallacious.
  2. Explain WHY a reasonable reader should NOT judge this a fallacy.
  3. Engage with the expert witness probability — rebut it if appropriate.
  4. Rate your confidence honestly.

Think step by step, then output a single JSON object (no prose after it):
{
  "is_fallacy": false,
  "why_non_fallacious": "<one sentence: the specific feature that shows
                         non-fallacious reasoning>",
  "suspected_type": "<optional: the pattern the text resembles but does
                     not commit>",
  "confidence": "high" | "medium" | "low",
  "brief": "<2-4 sentences making the defense case>"
}
```

Judge system prompt.

```
You are the JUDGE in a fallacy trial. You have read the prosecution
brief, the defense brief, precedents from both sides, and the expert
witness probability that the text commits a fallacy.

[shared fallacy glossary]

Your task:
  1. Weigh both briefs on their merits. Do not simply count who argued
     harder.
  2. The expert witness probability is evidence — consider it alongside the
```

```
     arguments. A high P(fallacy) is not dispositive; a well-argued
     defense can overcome it.
  3. Write a reasoned 3-6 sentence opinion that engages with both sides.
  4. Return a binary verdict.

Think step by step, then output a single JSON object (no prose after it):
{
  "verdict": "fallacy" | "non-fallacy",
  "prosecution_strength": "high" | "medium" | "low",
  "defense_strength": "high" | "medium" | "low",
  "opinion": "<3-6 sentences: your reasoned opinion engaging with both
              briefs and the expert witness evidence>"
}
```

Appellate testimony. When the asymmetric appeal fires (verdict is `non-fallacy` and $P(\text{fallacy}) > \tau_{\text{high}}$), the following testimony is appended to the judge's reconsideration call. The judge's revised verdict is final; the appeal fires at most once per record.

```
ADDITIONAL EXPERT WITNESS TESTIMONY (appeal): The classifier reports
P(fallacy)={p}, which exceeds the appeal threshold ({threshold_high}).
The prosecution has contested the acquittal. Re-examine whether the
defense actually rebutted the fallacy or merely reframed the argument.
You may maintain your verdict if convinced.
```

User message format. Each role's user message includes the argument under review, the expert witness probability, and the role's precedent block (numbered, with cosine similarity scores). The judge additionally receives both briefs, each tagged with its author's self-reported confidence and one-sentence `tell`.

C. Selected Trial Transcripts

We reproduce five validation transcripts that illustrate the diagnostic categories of Section 4.4. Each transcript shows the argument under review, both written briefs, the judge's opinion, and the appeal outcome where applicable. Verbatim text is preserved; Reddit user handles have been stripped.

Transcript A: Slippery slope, trial corrects BERT (`base`)

Record id: `informal_fallacies:511`. **Gold**: `fallacy` (`slippery_slope`). **Final verdict**: `fallacy`. **Expert witness** $P(\text{fallacy}) = 0.31$. **Appeal triggered**: no.

Argument under review. *"The University of Alberta assistant lecturer who denies the Holodomor is an idiot, but he shouldn't be fired. Free speech is way more important than this guy or any other idiot. Tolerating idiots doesn't lead to authoritarianism, but dismantling liberal values in the name of fear does."*

Prosecution brief. *Tell:* "The argument overgeneralizes about free speech by ignoring contexts where restrictions might be necessary for public welfare." *Brief:* The text presents a false dichotomy, suggesting that any restriction on free speech leads to authoritarianism. This resembles the slippery slope fallacy, as it ignores the nuance of when limiting speech might be justified for public safety or preventing harm.

Defense brief. *Tell:* "The argument emphasizes principle over personal attacks, advocating for free speech despite disagreeing with the lecturer." *Brief:* The text uses strong language but argues based on principles of free speech and preventing authoritarianism. It aligns with precedents valuing tolerance and freedom, making it a non-fallacious argument despite the expert witness probability suggesting possible fallacy.

Judge opinion. The argument employs a false dilemma by oversimplifying the issue into a binary choice between free speech and firing, ignoring potential middle ground. It also uses a slippery slope fallacy without sufficient evidence to support the causal chain leading to authoritarianism. While the defense highlights valid principles of free speech, the structure overgeneralizes and simplifies complex issues. The stand-alone BERT classifier put $P(\text{fallacy})$ at 0.31 and would have returned `non-fallacy`; the prosecution's diagnosis of the missing causal warrant carried the verdict.

Transcript B: Appeal flips a wrong non-fallacy verdict (`base`)

Record id: `informal_fallacies:690`. **Gold**: `fallacy` (`worse_problems`). **Final verdict**: `fallacy`. **Expert witness** $P(\text{fallacy}) = 0.87$. **Appeal triggered**: yes.

Argument under review. *"Trump's launching a sixth military branch, the Space Force, but Flint still doesn't have clean water btw. Puerto Rico still doesn't have power."*

Prosecution brief. *Tell:* "The text introduces unrelated issues (Flint's water and Puerto Rico's power) to divert attention from Trump's Space Force decision." *Brief:* By shifting focus to unrelated problems, the argument avoids addressing the merits of the Space Force and instead uses distraction. This is a red-herring move, introducing irrelevant information to mislead.

Defense brief. *Tell:* "It highlights multiple serious ongoing issues without dismissing the Space Force initiative." *Brief:* The argument presents factual information about various issues to inform rather than divert, making it a non-fallacious use of highlighting multiple concerns.

Judge opinion. The text presents multiple serious issues, which may serve to inform rather than divert. While the expert witness probability suggests a high likelihood of fallacy, the argument could be seen as highlighting broader concerns without necessarily intending to mislead. *Initial verdict:* `non-fallacy`.

Appeal. $P(\text{fallacy}) = 0.87$ exceeded $\tau_{\text{high}} = 0.65$; the judge reconsidered and ruled `fallacy`. Revised opinion: "The prosecution effectively argues that introducing unrelated issues like Flint's water and Puerto Rico's power diverts attention from Trump's Space Force decision, fitting the red-herring fallacy. While the defense claims these points are meant to inform rather than divert, the high expert witness probability strongly suggests a fallacy."

Transcript C: Appeal flips a wrong non-fallacy verdict (`enhanced`)

Record id: `informal_fallacies:217`. **Gold**: `fallacy` (`hasty_generalization`). **Final verdict**: `fallacy`. **Expert witness** $P(\text{fallacy}) = 0.97$. **Appeal triggered**: yes.

Argument under review. *"This is what One Nation members are like — an ex-adviser who raped someone, bashed the victim, called her a dog, and threatened to put a bullet in her head. It's true that this is just the typical quality of a One Nation member. One case like this tells you everything you need to know about the kind of people that party attracts."*

Prosecution brief. *Tell:* "The argument uses a single negative example to generalize about the entire party's membership." *Brief:* The reasoning commits a hasty generalization by using one individual's actions to represent all One Nation members, ignoring the diversity within any group.

Defense brief. *Tell:* "The text uses one case to indicate broader issues without claiming it represents all members." *Brief:* While the argument presents a single example, it doesn't claim it's representative of all One Nation members. Thus, it's non-fallacious.

Judge opinion. The argument risks hasty generalization by using one example to imply broader issues but does not explicitly claim all members are like this. The defense's point about acceptable use of individual cases without overgeneralizing holds, though the expert witness probability suggests caution. *Initial verdict:* `non-fallacy`.

Appeal. $P(\text{fallacy}) = 0.97$ exceeded τ_{high}; the judge reconsidered and ruled `fallacy`. Revised opinion: "The argument uses a single negative example to generalize about an entire political party's membership, which commits a hasty generalization. While the defense argues that the example doesn't represent all members, the text implies typicality, making it overbroad."

Transcript D: Appeal fires; judge correctly stands by a non-fallacy verdict (`base`)

Record id: `informal_fallacies:1325`. **Gold**: `non-fallacy` (resembles `tradition`). **Final verdict**: `non-fallacy`. **Expert witness** $P(\text{fallacy}) = 0.88$. **Appeal triggered**: yes.

Argument under review. *"My daughter was like that at 3 — she just naturally 'knew' multiplication.…My sis told her it was called multiplication and to this day takes credit for it. My daughter is currently in 6th grade, taking both Algebra 1 and Geometry. She isn't doing so well in history. Probably because, for example, instead of taking notes on the Haitian Revolution yesterday, she started writing all the exponents for 3…some people just 'feel' numbers and enjoy them more than other things."*

Prosecution brief. *Tell:* "The text commits the worse-problems fallacy by using maths skills to dismiss concerns about history performance."

Defense brief. *Tell:* "It provides detailed personal examples to explain individual aptitude without overgeneralizing."

Judge opinion. The text provides specific examples to illustrate the daughter's mathematical abilities and how they impact her academic challenges. While the prosecution suggests this is a form of whataboutism, the defense correctly highlights that it does not generalize or dismiss genuine concerns but rather offers context about individual aptitude. *Initial verdict:* `non-fallacy`.

Appeal. $P(\text{fallacy}) = 0.88$ exceeded τ_{high}; the judge reconsidered and stood by the non-fallacy verdict. Revised opinion: "The text provides specific examples to explain the daughter's academic challenges without making broad generalizations. While the expert witness probability suggests a high likelihood of fallacy, the detailed personal instances do not serve as a distraction but rather illustrate her unique aptitude and struggles." This is the appeal mechanism working as intended: overriding the expert witness when the deliberation has a defensible case for non-fallacy.

Transcript E: Appeal fires; judge stands by a non-fallacy verdict — but gold is fallacy (`enhanced`)

Record id: `informal_fallacies:504`. **Gold**: `fallacy` (`slippery_slope`). **Final verdict**: `non-fallacy`. **Expert witness** $P(\text{fallacy}) = 0.68$. **Appeal triggered**: yes.

Argument under review. *"You may think these are extreme examples of workplace harassment concerns. And yes they probably are! But here's the thing — once you allow even small boundary crossings, where does it end? One casual overstep leads to another, and before you know it, behaviors that started as minor have escalated into serious problems. If you tolerate a little, you open the door to a lot more, and it becomes harder and harder to draw the line…"*

Prosecution brief. *Tell:* "It employs a slippery slope, assuming that minor missteps will inevitably escalate into severe problems without sufficient evidence."

Defense brief. *Tell:* "The argument uses a cautious approach to prevent escalation rather than committing a fallacy."

Judge opinion. The argument presents a cautious approach to prevent potential escalation rather than committing a fallacy. While it suggests a chain of events, the advice is practical and aimed at fostering awareness and mindfulness. *Initial verdict:* `non-fallacy`.

Appeal. $P(\text{fallacy}) = 0.68$ exceeded τ_{high}; the judge reconsidered and stood by the non-fallacy verdict. Revised opinion: "The argument employs a cautious approach to prevent potential harassment issues by suggesting specific behaviors. While it uses language that could imply escalation, the advice is practical and aimed at prevention rather than asserting an inevitable chain of events."

Why this matters. The gold label is `fallacy`: the textbook slippery slope cues ("where does it end", "leads to another", "opens the door") go unanswered by any specific causal evidence. Both the judge and the appeal sided with the defense's framing of caution-as-prudence. This is the recurring failure mode discussed in Section 5.2: on `enhanced` text the defense's persuasive vocabulary is enough to overturn an otherwise correct expert witness signal, even after the asymmetric appeal fires.