

When AI-Text Detectors Become Evidence: Rethinking Detection for Human Decisions

Jiatao Li^{1,2} Mao Ye¹ Xinyu Hu¹
Cheng Peng¹ Xunjian Yin¹ Xiaojun Wan¹

¹Wangxuan Institute of Computer Technology, Peking University

²Information Management Department, Peking University
leejames@stu.pku.edu.cn, wanxiaojun@pku.edu.cn

Abstract

AI-generated-text detectors are evaluated as classifiers, yet consequential deployments use detector outputs as evidence that people inspect, trust, contest, or use to escalate decisions. We call this mismatch the classification–evidence gap. The gap makes scalar detector scores a weak deployment interface: they hide the basis of a judgment, invite overtrust, and can turn false positives into accusations. We argue that AI-text detection should be evaluated as evidence-grounded decision support rather than as authorship proof. We introduce AGENT-XTrust, a guideline-expert detector that returns direct MetaAgent labels without target-domain scalar-cutoff calibration and converts each document-level judgment into inspectable evidence records containing local labels, rationales, confidence estimates, and linked spans. Across three domains and four generators, AGENT-XTrust preserves strong binary accuracy while producing inspectable, contestable evidence. In a controlled fixed-stimulus user study, participants assigned to the full AGENT-XTrust system-and-interface package achieved 19.8 and 11.3 points higher decision accuracy than those assigned to the score-only and score-plus-explanation packages, respectively; overtrust on each condition’s detector-wrong cases was 17.8 and 18.3 points lower. These results suggest a broader evaluation target for AI-text detection: not only whether a detector is correct, but whether complete detector-and-evidence packages support reliable human decisions under uncertainty.

1 Introduction

A wrong AI-text detector judgment can become an accusation. Yet the field still evaluates detectors mostly as classifiers: given a document, return a score or label, and measure whether it matches a ground-truth author label. This creates a classification–evidence gap: detectors are optimized for label agreement, but deployed as evi-

dence in human decisions. The classifier framing has supported progress in supervised detectors and zero-shot statistical detectors such as DetectGPT and Fast-DetectGPT (Mitchell et al., 2023; Bao et al., 2024), but it misses the object that matters in deployment. In schools, reviewing workflows, hiring pipelines, and moderation systems, detector outputs are not final truths. They are evidence objects that people inspect, trust, challenge, cite, or use to defer action. Closing this gap requires changing both the detector output and the evaluation target.

The classification–evidence gap creates a design problem that higher classifier accuracy alone cannot solve. A scalar score hides the basis of a judgment. A confident false positive can invite punitive action before a human understands what the model actually saw. A generic explanation can make the same score sound more authoritative without giving users concrete grounds to contest it. A fourth risk is threshold dependence: deployment requires converting scores into actions through scalar cutoffs that may not transfer across domains, generators, or institutional risk tolerances. The deployment problem is therefore not only to predict whether text was generated, but to help people decide which evidence is reliable enough to trust, question, or defer.

We frame AI-generated text detection as *evidence-grounded decision support*: a task of exposing usable evidence under uncertainty, not proving authorship from surface cues. Under this view, the unit of detector output is not merely a score; it is an inspectable, contestable evidence record: a decision with uncertainty, local or global evidence, cue-level rationales, and enough structure for a human to inspect or challenge the model’s basis. Detector outputs should not be treated as dispositive proof of authorship. Precisely because final decisions remain uncertain, the evidential basis must be exposed. The evidence vocabulary should there-

fore occupy a middle ground: more structured than a free-form rationale, but less brittle than a hand-written rule system. We build this vocabulary from externally sourced writing cues and treat them as human-readable priors over possible evidence, motivated by a study in which frequent LLM users identified English AI-generated nonfiction through recurring lexical, stylistic, formality, originality, and clarity cues (Russell et al., 2025).

AGENT-XTrust operationalizes this score-to-evidence view by routing each document to semantic, stylistic, and structural guideline experts grounded in stylistics and authorship analysis. Each expert produces a local AI/human judgment, a rationale, a confidence estimate, and candidate evidence spans; a MetaAgent then aggregates these reports into a document-level decision. Figure 1 summarizes this router-expert-MetaAgent backbone and its two deployment-facing components: a source-grounded detection guideline inventory and a trust layer for span-level inspection. The architecture is not merely an ensemble for improving classification. It changes the detector output object from a scalar score to a contestable decision package: a direct MetaAgent judgment plus inspectable evidence records that can be challenged and compared.

This framing makes the detector more than a score-producing classifier, and it makes the evaluation target larger. We separate three questions that are often collapsed: Is the detector correct? Is its evidence inspectable and grounded? Does the complete assistance package support reliable human reliance? The detector must be accurate enough to be useful, the displayed evidence must be concrete enough to inspect, and the interface should help users avoid blind deferral to an uncertain system. The goal is not to certify authorship, but to make fallible detector outputs more contestable and less prone to blind trust. We therefore evaluate false accusation, overtrust, deferral, calibration, and evidence localization alongside benchmark accuracy.

Our contributions are:

- We identify a deployment mismatch in AI-text detection: detectors are optimized as classifiers, but used as evidence. We argue that detector evaluation should include whether outputs support reliable human judgment under uncertainty.
- We introduce AGENT-XTrust, which converts document-level detection into structured ev-

idence records linking guideline cues, local labels, confidence estimates, rationales, and spans.¹

- We evaluate classification quality, evidence quality, and human reliance quality, showing that AGENT-XTrust preserves strong detector accuracy, produces inspectable evidence, and that participants assigned to its complete assistance package achieve higher decision accuracy and lower overtrust than participants assigned to the two baseline packages.

2 Background

AGENT-XTrust draws on authorship attribution without treating detector output as authorship proof. Classical authorship attribution studies infer textual source from recurring regularities in lexical choice, syntactic structure, rhetorical organization, and semantic preference, rather than from deterministic markers (Mosteller and Wallace, 1964; Burrows, 2002; Stamatatos, 2009). For AI-generated text, the useful premise is narrower: generated texts may be fluent, but they can exhibit recurring stylistic regularities that serve as fallible evidence rather than deterministic markers (Gehrmann et al., 2019; Ippolito et al., 2020; Uchendu et al., 2020).

We organize detection evidence into three linguistic dimensions. The *semantic dimension* concerns meaning-making: specificity of claims, originality of commentary, conceptual depth, attribution of knowledge, and whether the text contributes situated information rather than generic explanation. The *stylistic dimension* concerns expressive realization: lexical precision, hedging, formality, rhetorical smoothness, idiosyncratic phrasing, and the presence or absence of human-like irregularity. The *structural dimension* concerns discourse organization: paragraph rhythm, transition patterns, narrative distribution, argumentative progression, and whether coherence is adaptive or mechanically uniform. This division follows a long tradition in stylistics and functional linguistics that treats textual meaning as jointly produced by content, expression, and discourse organization (Halliday, 1978; Halliday and Matthiessen, 2004).

This taxonomy gives the detector a structured alternative to prompt ensembling. Each guideline expert is asked to inspect a specific kind of evidence. A semantic expert may examine whether a

¹Our code and release materials are available at <https://github.com/leejames/AGENT-XTrust>.

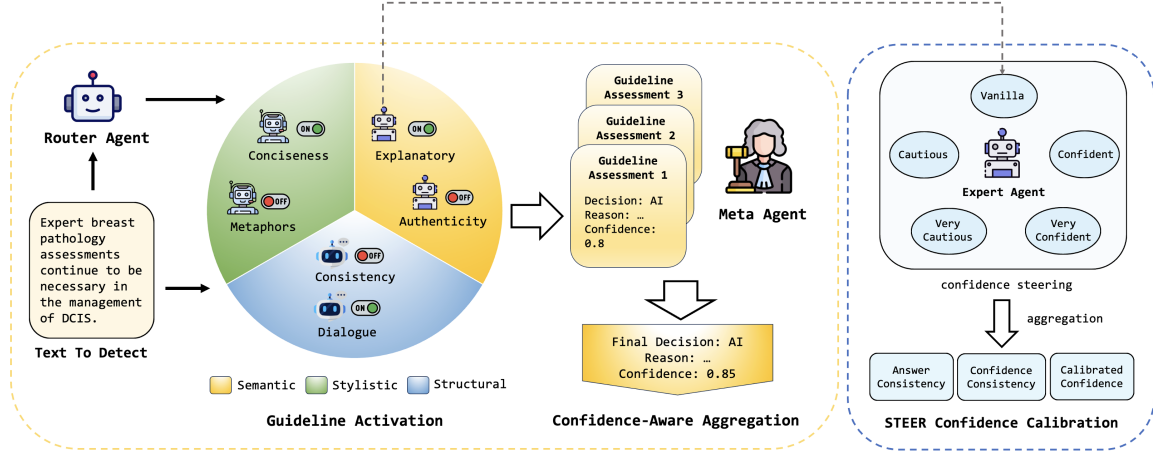


Figure 1: Overview of AGENT-XTrust as a score-to-evidence detector. Rather than exposing only a document-level score, the router selects relevant linguistic guideline experts, each expert produces a local decision, rationale, confidence estimate, and evidence spans, and the MetaAgent aggregates these assessments into a direct document decision without target-domain scalar-cutoff calibration while preserving an inspectable evidence record.

document makes unsupported notability claims; a stylistic expert may inspect generic balanced phrasing; a structural expert may inspect repetitive paragraph scaffolding. The taxonomy also makes disagreement meaningful. If stylistic experts identify AI-like polish but structural experts find realistic source attribution, the MetaAgent can expose that tension alongside the document-level score.

The Wiki-derived guideline inventory instantiates this taxonomy with public, source-grounded diagnostic cues. This makes the cue set easier to inspect while preserving the central design principle: detection should be based on structured, inspectable linguistic evidence. The cue set is not intended as a list of universal authorship laws. It is a human-prior vocabulary: a way to name lexical, stylistic, semantic, and discourse evidence resembling the cues reported by frequent LLM users when explaining AI-text judgments (Russell et al., 2025).

3 Task

Standard AI-text detection maps a document x to a binary label $y \in \{\text{AI}, \text{human}\}$ and often a score $p(y = \text{AI} \mid x)$. For human decision-making, this is underspecified: a reviewer, instructor, or moderator needs to know why a score was produced, which passages support it, whether the evidence is strong enough to act on, and whether a tuned threshold is appropriate for the current domain. In this setting, the detector output should not be treated as dispositive proof of authorship; it should support a decision process in which evidence can be in-

spected and challenged. We therefore define an evidence-grounded detector as returning

$$f(x) = (\hat{y}, s, E, H),$$

where $\hat{y}$ is the document prediction, s is an AI-evidence score, E is a set of guideline-level evidence records, and H is a set of highlighted spans linked to those records. Each evidence record contains a cue identifier, local judgment, confidence, rationale, and quoted or sentence-level support. The detector is intended as decision support: its role is to make the support for a judgment inspectable rather than to certify misconduct. This definition changes the evaluation target. A detector is no longer evaluated only by label agreement; it must also be evaluated by whether its evidence is inspectable and whether the resulting output package supports reliable human decisions under uncertainty. This separates three goals that score-only detectors conflate: classification quality, evidence quality, and human reliance quality.

4 Method

4.1 Overview

Given a document x , AGENT-XTrust first invokes a router agent, then runs a selected set of guideline-specific base agents, and finally uses a MetaAgent to synthesize their reports. The detector also preserves expert outputs for an evidence layer that supports span-level inspection in the deployed interface. Thus the architecture has two coupled outputs: a direct document decision without target-domain scalar-cutoff calibration and an inspectable

evidence package. This design choice is central to our framing: if the detector is wrong, the evidence package should still allow a human to see why the system erred and to challenge weak support.

The pipeline returns structured guideline reports $e_i = (g_i, \hat{y}_i, c_i, r_i)$, where g_i is the activated guideline, $\hat{y}_i$ is the local judgment, c_i is confidence, and r_i is the rationale. Optional quoted spans from expert outputs are preserved when available and otherwise augmented with sentence-level span candidates, so the evidence layer preserves intermediate reports alongside the final label.

4.2 Guideline Inventory

Guidelines are the central detection primitive. We use a fixed guideline inventory derived from Wikipedia’s community-maintained “Signs of AI writing” page (Wikipedia contributors, 2026). The inventory contains observable cues such as undue emphasis on significance or legacy, over-attribution and notability claims, superficial analysis, promotional language, generic balanced framing, formatting artifacts, repetitive discourse structure, and low-specificity claims.

The guideline inventory is fixed before evaluation and grounded in public provenance, linguistic interpretability, and coverage of content, expression, and discourse evidence. Wikipedia’s guidance collects descriptive field observations from a high-volume editorial setting. Prior work further shows that frequent ChatGPT users can be accurate AI-text detectors and explain judgments using lexical, stylistic, formality, originality, and clarity cues (Russell et al., 2025). We therefore treat the inventory as a codified human-prior vocabulary over evidence types, not as deterministic authorship rules; its value comes from routed, instance-level judgments and MetaAgent synthesis rather than any single cue.

Fixed here means frozen before evaluation, not architecturally immutable. Cue modules can be versioned, retired, or specialized for a domain, and each revised inventory should be revalidated for its intended population and writing setting.

4.3 Router Agent

The router agent determines which guidelines and corresponding expert agents should be activated for the input text. The router performs instance-level selection rather than invoking the full guideline set. It first analyzes the input’s domain and stylistic attributes, such as whether the document resem-

bles biomedical explanation, news summarization, creative narrative, formal argumentation, or promotional prose. It then selects the subset of guidelines whose cues are likely to be relevant for that input.

This adaptive routing matters because AI-writing cues are domain-conditional rather than universally diagnostic. A cue that is informative for one genre can be uninformative for another, so the router reduces noise by preventing irrelevant guidelines from voting on every document. Activated expert outputs are retained as evidence for downstream aggregation and interface display; implementation details are provided in Appendix F.2.

4.4 Expert Agents

Each activated guideline-specific expert independently evaluates the document according to its assigned cue. Each expert produces three core outputs: a binary classification decision, a textual rationale grounding that decision in the input, and a confidence estimate. Experts are additionally encouraged to quote or identify candidate spans that support the rationale, so that their reports can be transformed into user-facing evidence.

The decomposition is deliberate: a final decision can be traced back to the activated guidelines and the local evidence each expert found, rather than to a monolithic score.

4.5 Confidence Elicitation

Prompt-based LLMs are known to produce unstable confidence estimates. To reduce sensitivity to a single verbal confidence judgment, AGENT-XTrust adopts SteerConf’s mechanism for eliciting prediction–confidence pairs (Zhou et al., 2025) and augments each pair with a rationale. Each activated expert evaluates the same input under a small set of *confidence-steered prompts* ranging from *very cautious* to *very confident*. Under each steering level $P_k \in \mathcal{P}$, the expert returns a prediction $f_k(x)$, rationale, and confidence value $c_k(x)$.

We compute three quantities from the set $\{c_k(x)\}$:

(1) the mean confidence,

$$\mu_c = \frac{1}{|\mathcal{P}|} \sum_k c_k(x),$$

(2) an answer-consistency score capturing how often the predictions agree,

$$\kappa_{\text{ans}} = \frac{1}{|\mathcal{P}|} \max_{y \in \{\text{AI}, \text{Human}\}} \sum_k \mathbb{I}[f_k(x) = y],$$

(3) a confidence-stability term that down-weights large fluctuations,

$$\sigma_c = \sqrt{\frac{1}{|\mathcal{P}|} \sum_k (c_k(x) - \mu_c)^2}, \quad \kappa_{\text{conf}} = \begin{cases} \frac{1}{1+\sigma_c/\mu_c}, & \mu_c > 0, \\ 0, & \mu_c = 0. \end{cases}$$

These behavior-derived quantities are multiplicatively combined:

$$c_{\text{cal}}(x) = \mu_c \cdot \kappa_{\text{ans}} \cdot \kappa_{\text{conf}}.$$

Finally, the guideline-specific decision is selected from the steering run whose original confidence is closest to c_{cal} :

$$k^* = \arg \min_k |c_k(x) - c_{\text{cal}}(x)|, \quad f_{\text{final}}(x) = f_{k^*}(x).$$

This mechanism is used as a support component for confidence stability, not as a complete calibration solution. Its role is to reduce sensitivity to a single confidence phrasing and to expose confidence stability as part of the evidence record.

4.6 Meta Agent

The MetaAgent synthesizes all activated guideline reports into a document-level decision. It receives each expert’s local label, rationale, confidence value, guideline identity, and linguistic dimension, then produces a final AI/human judgment and a global rationale. This meta decision is the primary detector output. We use *threshold-free* only in the operational sense that the final label is generated directly by the MetaAgent rather than by comparing a scalar detector score against a tuned cutoff. The system still has a binary decision rule, but it does not require selecting a scalar threshold on target-domain validation data before returning that decision.

The MetaAgent performs evidence-weighted synthesis. It compares the quality and relevance of rationales, weighs confidence values, and reconciles disagreement among linguistic dimensions. For example, stylistic experts may identify formulaic balance while structural experts may identify realistic source attribution. In such cases the MetaAgent summarizes both sides and makes a consensus decision based on the strongest activated evidence. This preserves the core design goal: interpretable aggregation over specialized linguistic agents alongside document-level scoring.

In AGENT-XTrust, the MetaAgent produces the detector decision, while local guideline records, confidence values, and rationales are preserved as first-class evidence for user inspection.

4.7 Evidence Interface

The interface layer converts guideline evidence into span-level presentation. For guidelines supporting an AI decision, quoted spans are used when available; otherwise, sentence-level evidence candidates are linked to the supporting guideline records. The interface presents highlights, guideline titles, confidence values, short rationales, and a document-level evidence summary. Appendix Figure 4 shows the decision-support interface: users can inspect the aggregate judgment, highlighted-evidence coverage, and the guideline citations attached to each suspicious span. This turns expert assessments into inspectable evidence for human decision support.

5 Experiments

5.1 Datasets

We evaluate document-level detection using common AI-text-detection settings based on XSum for news summarization, WritingPrompts for creative writing, and PubMedQA for biomedical question answering (Narayan et al., 2018; Fan et al., 2018; Jin et al., 2019). This benchmark evaluates generations from GPT-3.5, GPT-4, Claude-3 Opus, and Claude-3 Sonnet.

We additionally evaluate the full AGENT-XTrust system across four generators and three writing domains: GPT-3.5, GPT-4, Claude-3 Opus, and Claude-3 Sonnet over XSum, WritingPrompts, and PubMedQA. All reported AGENT-XTrust evaluations use a fixed setup across datasets and generators: a single source-grounded guideline inventory, consistent prompting, and direct MetaAgent decisions without target-domain scalar-cutoff calibration. Appendix F.2 gives model, decoding, and parsing details.

5.2 Baselines

For document-level detection, AGENT-XTrust is compared against supervised RoBERTa detectors (Liu et al., 2019) and zero-shot detectors including likelihood, entropy, rank, LogRank, LRR, DNA-GPT, DetectGPT, and Fast-DetectGPT (Gehrmann et al., 2019; Ippolito et al., 2020; Su et al., 2023; Yang et al., 2024; Mitchell et al., 2023; Bao et al., 2024). Baseline thresholds are selected on a held-out validation set when required by the experimental setup; Appendix A.3.1 gives the exact selection rule.

5.3 Metrics

Accuracy is the primary detector metric because the method is intended to return an actionable binary decision without target-domain scalar-cutoff tuning. Many zero-shot detectors report AUROC, but practical use still requires translating continuous scores into binary actions, usually by selecting thresholds on an external calibration set. Such calibration data may be unavailable in new domains, and different institutions may tolerate false accusations and missed AI text differently. We therefore treat binary accuracy under a fixed decision procedure as central. AUROC remains important, but it answers a different question about ranking quality rather than deployment-time action.

We also report AUROC as a secondary metric because it measures ranking quality independent of a fixed threshold. Reporting both metrics separates two questions: whether the system can make strong binary decisions without target-domain scalar-cutoff calibration, and whether its confidence scores produce a useful ranking of suspiciousness.

For the evidence and decision-support evaluation, we report six corresponding quantities: manual evidence quality, detector-confidence calibration, human decision accuracy, false accusation, overtrust, and deferral. Manual evidence quality measures whether displayed support is concrete, grounded, and guideline-relevant and is reported through the evidence audit in Table 3. Detector-confidence calibration is reported as ECE before and after confidence steering in Figure 2. Human decision accuracy, false accusation, overtrust, and deferral are reported through the controlled user study in Table 4.

6 Results

We evaluate AGENT-XTrust along three axes. First, we ask whether an evidence-grounded detector can retain strong classification quality without target-domain threshold tuning. Second, we test whether the emitted evidence is concrete, grounded, and inspectable rather than merely fluent. Third, we ask whether this evidence changes human reliance behavior when detector outputs are uncertain or wrong.

6.1 Detection Accuracy

RQ1: Can evidence-grounded detection preserve classification quality? We first test whether

Generator	Baseline	AGENT-XTrust	Δ
GPT-3.5	82.56	86.01	+3.45 pts
GPT-4	68.00	85.92	+17.92 pts
Claude-3 Opus	82.00	83.78	+1.78 pts
Claude-3 Sonnet	78.56	80.00	+1.44 pts
Average	77.78	83.93	+6.15 pts

Table 1: Main detector benchmark summary. Values are average binary accuracies in percent across XSum, WritingPrompts, and PubMedQA for one generator family. “Best baseline” is the strongest result among all supervised and zero-shot baselines introduced in Section 5 and detailed in Appendix A.3; Δ reports AGENT-XTrust’s absolute percentage-point improvement over that strongest baseline. AGENT-XTrust reports direct MetaAgent decisions without target-domain scalar-cutoff calibration, while baseline binary accuracies use validation-threshold tuning when required.

converting detector outputs from scores into structured evidence records sacrifices benchmark accuracy. It does not: across representative GPT- and Claude-family generators, AGENT-XTrust remains competitive with supervised and zero-shot baselines while requiring no target-domain deployment threshold. Evidence grounding therefore complements, rather than replaces, detection quality.

Table 1 reports the main detector benchmark; Appendix A.3 gives the full per-domain tables. Each “Best baseline” entry is the strongest competitor after evaluating the supervised and zero-shot detectors described above, so the comparison is against the best baseline available for that generator family rather than a single hand-picked method. Across GPT-3.5, GPT-4, Claude-3 Opus, and Claude-3 Sonnet, AGENT-XTrust has the highest average point estimate in each generator-family setting. Some score-based baselines remain stronger on specific news and creative-writing subsets; the main pattern is averaged cross-domain robustness together with evidence support. The comparison also highlights a deployment distinction: score-based baselines require a selected cutoff for binary accuracy, while AGENT-XTrust reports the direct MetaAgent decision under the same fixed guideline inventory and prompting setup. Thus the evidence layer is attached to a competitive detector rather than compensating for weak labels.

Ablations. Table 2 reports detector-backbone ablations. Removing the guideline structure causes the largest accuracy drop, and weakening adaptive routing also hurts average performance. Rout-

Method	Acc. (%)	AUROC (%)
Full AGENT-XTrust	85.92	90.07
w/o Multi-Agent	78.68	<u>87.85</u>
w/o Guidelines	61.39	78.25
w/o LLM Router	71.97	77.38
w/o Adaptive Routing	75.88	86.65
w/o Confidence Steering	<u>79.91</u>	84.80

Table 2: Component ablations on GPT-4 generations, averaged across XSum, WritingPrompts, and PubMedQA. Accuracy and AUROC are percentages.

Criterion	Strict yes	Agree	κ
Understandable	100.0	100.0	–
Grounded	100.0	100.0	–
Span relevant	87.8	97.8	0.89
No hallucination	98.9	100.0	1.00
All four	87.8	99.4 [†]	0.91 [†]

Table 3: Main evidence-audit summary over 90 sampled records. Values are percentages except κ . “Strict yes” uses conservative consensus labels. For the “All four” row, 87.8% is the per-record four-criterion composite; [†]agreement and κ are pooled across the 360 individual criterion labels.

ing prevents domain-irrelevant cue families from voting on every document, while the semantic, stylistic, and structural experts preserve distinct evidence for MetaAgent reconciliation. The guideline, router, and multi-agent drops therefore support the decomposition motivated in Section 2. Appendix C.1 additionally shows a qualitative case of how guideline outputs become inspectable evidence.

6.2 Evidence Audit

RQ2: Are the evidence records inspectable and grounded? We separate document-label performance from evidence presentation. We conduct a two-annotator manual evidence-quality evaluation over sampled AGENT-XTrust records, measuring whether each displayed rationale is understandable, grounded in concrete text, relevant to the cited guideline, and free of unsupported claims. The criterion is stricter than perceived helpfulness: the displayed span should support the cited guideline-level claim.

Table 3 reports the compact main-paper audit. The conservative consensus evaluation finds that all 90 sampled records are understandable and concretely grounded, while 79/90 (87.8%) satisfy all four strict criteria. The main error category is overly narrow or misaligned span selection. Ap-

pendix D.2 describes the rubric and provides domain and detector-correctness breakdowns. These results indicate that AGENT-XTrust usually emits evidence records that annotators can inspect against the source text, not only labels and explanations.

Routing statistics in Appendix B.4 show that routing does not collapse onto a single evidence type: PubMedQA receives slightly more semantic activation, WritingPrompts receives more structural activation, and XSum leans stylistic, but all three dimensions contribute across datasets. This supports the design choice of using a multi-dimensional expert network rather than a single flat prompt or a single scalar detector score.

6.3 Human Decision Support

RQ3: How do complete assistance packages differ in human reliance? Automatic accuracy does not establish whether a detector helps people use uncertain evidence well. Beyond automatic detector metrics, we therefore evaluate three complete system-and-interface conditions in a controlled between-subject decision-support study: Fast-DetectGPT score-only, a direct-LLM score plus global explanation, and the full AGENT-XTrust package with spans, guideline citations, local rationales, and confidence. The stimulus set contains 18 fixed items, with six from each domain and a 9/9 human–AI label balance; items are stratified to include detector-correct, detector-wrong, high-confidence, low-confidence, and system-disagreement cases. Because detector identity and presentation co-vary across conditions, the study estimates differences among complete assistance packages rather than an isolated evidence-format effect.

The primary outcomes are human accuracy over all main decisions, false-accusation rate on human-written texts, overtrust when the displayed detector is wrong, and deferral rate. The main balanced analysis contains 30 retained participants per condition ($N = 90$), after the pre-specified quality filter and condition-balancing rule described in Appendix E.1. We report participant-bootstrap confidence intervals and participant-level permutation tests for the primary study contrasts.

Table 4 shows the decision-support results. Participants assigned to the full AGENT-XTrust condition achieved higher accuracy and lower overtrust than participants assigned to the Fast-DetectGPT score-only and direct-LLM explanation conditions. Participants in the direct-LLM explanation condi-

Condition	Acc. ($\uparrow$)	False acc. ($\downarrow$)	Overtrust ($\downarrow$)	Defer ($\downarrow$)
Score only	54.3 [49.4, 59.3]	23.3 [16.7, 30.4]	64.4 [53.3, 75.6]	19.1 [14.1, 24.1]
Score + explanation	62.8 [57.6, 67.8]	33.3 [28.9, 37.8]	65.0 [55.8, 74.2]	15.2 [11.7, 18.7]
AGENT-XTrust	74.1 [67.8, 80.0]	17.8 [13.3, 23.0]	46.7 [36.7, 55.6]	8.7 [5.2, 12.8]

Table 4: Human-study results after pre-specified quality filtering. Each interface condition has 30 retained participants. Arrows indicate the preferred direction for decision support. Accuracy and deferral are computed over all main decisions; false accusation is measured on human-written items; overtrust is measured on detector-wrong items. Small gray brackets show participant-bootstrap 95% confidence intervals. Bold marks the best point estimate for each metric.

tion also achieved higher accuracy than those in the score-only condition, without a corresponding reduction in false accusation or overtrust. These are condition-level comparisons of complete system-and-interface packages: detector identity, scores, explanations, and presentation differ together, and detector-wrong subsets are defined by each condition’s own displayed detector errors. The results therefore do not identify the causal contribution of structured evidence, any individual display field, or evidence format alone. Appendix E.1, Tables 14–17 provide domain, participant-level, and mixed-effects robustness breakdowns for the package contrasts.

7 Related Work

AI-generated text detection. Most work frames AI-text detection as document-level classification. SemEval-2024 Task 8 evaluates multidomain, multigenerator, and multilingual detection, source attribution, and single human-to-machine change-point detection (Wang et al., 2024); participant systems such as TrustAI address its document-level detection and attribution subtasks (Urlana et al., 2024). Supervised detectors commonly use pre-trained discriminative encoders such as RoBERTa (Liu et al., 2019), while zero-shot methods use probability, rank, perturbation, or intrinsic-feature signals (Gehrmann et al., 2019; Ippolito et al., 2020; Mitchell et al., 2023; Su et al., 2023; Yang et al., 2024; Bao et al., 2024; Yu et al., 2024). Recent work also studies cross-model and proprietary-model settings (Mireshghallah et al., 2024; Bao et al., 2025). These approaches remain primarily score-based; AGENT-XTrust instead preserves evidence for human inspection.

Evidence signals. Authorship-attribution and stylistics research emphasizes recurring lexical and stylistic patterns (Koppel et al., 2009; Simpson,

2004), while discourse theory formalizes organization and rhetorical structure (Mann and Thompson, 1988). Frequent LLM users in a study of English nonfiction explained AI-text judgments using lexical, stylistic, formality, originality, and clarity cues (Russell et al., 2025). AGENT-XTrust operationalizes these cues as inspectable priors. Its confidence elicitation builds on calibration and confidence-steering work (Jiang et al., 2023; Zhou et al., 2025), while its role-specialized collaboration draws on multi-agent systems and debate (Guo et al., 2024; Wu et al., 2024; Du et al., 2024; Hong et al., 2024).

ExaGPT is the closest example-based, human-facing design, exposing retrieved human- and machine-generated spans as analogical evidence (Koike et al., 2026). In contrast, AGENT-XTrust emits criterion-linked local judgments, rationales, confidence estimates, and source spans, providing a complementary diagnostic evidence format.

Decision support. Our work is closest to evaluations that treat model outputs as inputs to human decisions. XAI and human-centered AI research argues that explanations should be evaluated against user needs and downstream decision quality (Doshi-Velez and Kim, 2017; Mohseni et al., 2021; Poursabzi-Sangdeh et al., 2021). AGENT-XTrust applies this principle to AI-text detection by jointly evaluating classification, evidence quality, false accusation, overtrust, and deferral. Because fluent rationales can affect trust independently of correctness, we separately audit whether displayed spans and rationales support their cited guideline-level claims.

8 Conclusion

AI-text detectors are already used as evidence, but the field still often evaluates them as if a scalar score were the end of the task. This paper presents AGENT-XTrust and a broader argu-

ment: the classification–evidence gap should become a first-class evaluation problem for AI-text detection. Detectors should be evaluated by the human decisions their evidence supports, not only by the labels their classifiers emit. AGENT-XTrust combines a guideline expert network with direct MetaAgent decisions that require no target-domain scalar-cutoff calibration, span-level evidence presentation, and decision-support evaluation. The benchmark results show strong binary accuracy, the evidence audit shows that displayed records are usually concrete and guideline-relevant, and the controlled study shows differences among three complete assistance packages. Future AI-text detectors should be judged by three linked criteria: whether they classify documents correctly, whether their evidence can be inspected and contested, and whether complete detector-and-evidence packages support reliable human decisions.

Limitations

Our results characterize complete detector-and-interface packages under a controlled protocol covering unedited English documents from three domains, four generators, and one pinned backbone and prompt package. Claims are therefore scoped to package-level comparisons in this setting. Evaluating individual interface components, edited or multilingual text, mixed authorship, and alternative backbones is future work.

Ethical Considerations

AI-text detection errors can have consequences in high-stakes settings. AGENT-XTrust is intended as a review aid: deployments should communicate uncertainty and preserve human review and appeal rather than automate sanctions. Displayed spans and rationales should be treated as supporting evidence. The compensated, minimal-risk human study collected no direct identifiers, and released study materials will exclude participant identifiers.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (L253001), Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology) and National Engineering Research Center of New Electronic Publishing Technologies. We appreciate the anonymous reviewers for their

helpful comments. Xiaojun Wan is the contact author.

References

- Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. 2025. [Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection](#). In *International Conference on Learning Representations*, pages 51844–51865.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. [Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature](#). In *International Conference on Learning Representations*, pages 24814–24836.
- John Burrows. 2002. ‘delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.
- Finale Doshi-Velez and Been Kim. 2017. [Towards a rigorous science of interpretable machine learning](#).
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.
- Patrizio Giovannotti and Alexander Gammernan. 2024. [Calibrated large language models for binary question answering](#). In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 218–235. PMLR.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.
- M.A.K. Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Open University set book. Edward Arnold.

- M.A.K. Halliday and C.M.I.M. Matthiessen. 2004. *An Introduction to Functional Grammar*. Arnold.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for a multi-agent collaborative framework](#). In *International Conference on Learning Representations*, pages 23247–23275.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.
- Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. [Calibrating language models via augmented prompt ensembles](#). In *Workshop on Challenges in Deployable Generative AI at ICML 2023*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Ryuto Koike, Masahiro Kaneko, Ayana Niwa, Preslav Nakov, and Naoaki Okazaki. 2026. [ExaGPT: Example-based machine-generated text detection for human interpretability](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7702–7715, San Diego, California, United States. Association for Computational Linguistics.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. [Computational methods in authorship attribution](#). *Journal of the American Society for Information Science and Technology*, 60(1):9–26.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- William C. Mann and Sandra A. Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Niloofer Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2024. [Smaller language models are better zero-shot machine-generated text detectors](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–293, St. Julian’s, Malta. Association for Computational Linguistics.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. [Detectgpt: Zero-shot machine-generated text detection using probability curvature](#). In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.
- Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. [A multidisciplinary survey and framework for design and evaluation of explainable ai systems](#). *ACM Trans. Interact. Intell. Syst.*, 11(3–4):1–45.
- Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. [Manipulating and measuring model interpretability](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI ’21*, pages 1–52, New York, NY, USA. Association for Computing Machinery.
- Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. [People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5342–5373, Vienna, Austria. Association for Computational Linguistics.
- P. Simpson. 2004. [Stylistics: A Resource Book for Students](#). Language workbooks. Routledge.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. [Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023*, pages 12395–12412. Association for Computational Linguistics.
- Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395.

- Ashok Urlana, Aditya Saibewar, Bala Mallikarjunarao Garlapati, Charaka Vinayak Kumar, Ajeet Kumar Singh, and Srinivasa Rao Chalamala. 2024. [TrustAI at SemEval-2024 task 8: A comprehensive analysis of multi-domain machine generated text detection techniques](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 927–934, Mexico City, Mexico. Association for Computational Linguistics.
- Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. [SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.
- Wikipedia contributors. 2026. Wikipedia:signs of AI writing. https://en.wikipedia.org/w/index.php?title=Wikipedia:Signs_of_AI_writing&oldid=1353263698. Accessed 2026-05-09.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. [AutoGen: Enabling next-gen llm applications via multi-agent conversations](#). In *Conference on Language Modeling*.
- Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. 2024. [DNA-GPT: divergent n-gram analysis for training-free detection of gpt-generated text](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- WonJin Yoon, Ian Bulovic, and Timothy A. Miller. 2025. [Using tournaments to calculate auroc for zero-shot classification with llms](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23583–23591, Suzhou, China. Association for Computational Linguistics.
- Xiao Yu, Kejiang Chen, Qi Yang, Weiming Zhang, and Nenghai Yu. 2024. [Text fluoroscopy: Detecting LLM-generated text through intrinsic features](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15838–15846, Miami, Florida, USA. Association for Computational Linguistics.
- Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. [Don’t go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.
- Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. 2025. [Steerconf: Steering llms for confidence elicitation](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 19460–19486. Curran Associates, Inc.

A Evaluation Details

This appendix gathers supplementary metrics, baseline settings, threshold selection, full accuracy tables, and additional AUROC results.

A.1 Evaluation Metrics

In this study, we use accuracy as the primary metric because the system returns direct binary decisions without target-domain cutoff tuning, and report Area Under the Receiver Operating Characteristic (AUROC) as a secondary ranking metric.

Accuracy measures the proportion of correctly classified samples to the total number of samples, aligning with practical requirements for clear binary decisions. Existing score-based detection methods report threshold-independent AUROC (Mitchell et al., 2023; Bao et al., 2024), but converting their scores into deployment actions still requires selecting and validating a decision threshold. In new domains, labeled calibration data may be unavailable. In contrast, AGENT-XTrust returns direct MetaAgent labels without target-domain scalar-cutoff calibration, making accuracy a central evaluation metric while AUROC remains a secondary ranking measure.

AUROC measures ranking performance across all possible thresholds, quantifying the probability that a randomly selected AI-generated text receives a higher machine-generation score than a randomly chosen human-written text. AUROC ranges from 0.0 to 1.0, with 0.5 indicating random ranking. For prompt-based LLM classifiers, label-token or verbalized confidence can be poorly calibrated and sensitive to prompting (Giovannotti and Gamerman, 2024; Jiang et al., 2023; Zhang et al., 2024). Standard zero-shot LLM classifiers may also return discrete decisions without an adjustable boundary; tournament-based pairwise rankings provide continuous scores for AUROC and AUPRC in that setting (Yoon et al., 2025).

To empirically illustrate this issue, we evaluated the calibration of confidence scores generated by

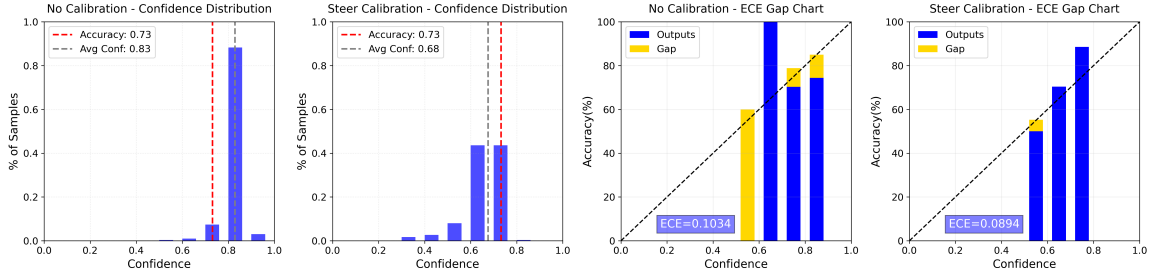


Figure 2: Confidence distributions and reliability diagrams for GPT-4-generated PubMed texts before and after confidence steering in the AGENT-XTrust detector. Gray dashed lines denote mean confidence; red dashed lines denote observed accuracy; yellow bars indicate miscalibration gaps.

our LLM-based expert agent, both before and after confidence steering (see Figure 2). Uncalibrated confidence scores showed measurable miscalibration (ECE = 0.1034). Confidence steering reduced ECE to 0.0894, while leaving room for further calibration improvements. This illustrates the challenge of relying directly on generative LLM confidence values without explicit calibration.

Consequently, we emphasize accuracy as the primary metric and report AUROC as an auxiliary measure of ranking performance across thresholds.

A.2 Baseline Details

Log-Likelihood (Gehrmann et al., 2019). This zero-shot method uses a language model to calculate token log probabilities. A higher average log-likelihood indicates a greater likelihood that the text was generated by an LLM.

Entropy (Su et al., 2023). This baseline averages predictive token-distribution entropy. Following the DetectLLM evaluation, larger entropy is treated as stronger machine-generation evidence, with score orientation validated on the held-out split.

Rank (Gehrmann et al., 2019). Rank-based detection assigns each token its numerical rank under a language model conditioned on the preceding context. Generated text tends to concentrate among higher-probability, lower-rank tokens; the decision orientation is frozen on the validation split.

Log-Rank (Su et al., 2023). Log-rank averages the logarithm of each token’s model rank across the passage. Generated text typically has lower average log-rank, with the score orientation frozen on the validation split.

LRR (Su et al., 2023). The Log-Likelihood Log-Rank Ratio combines likelihood and rank information into a zero-shot detector. Text with a higher

LRR score is treated as more likely to be generated by an LLM.

NPR (Su et al., 2023). Normalized Perturbed Log-Rank compares log-rank scores between the original text and perturbed variants. A higher NPR score indicates stronger machine-generation evidence.

DetectGPT (Mitchell et al., 2023). DetectGPT detects machine-generated text through probabilistic curvature: it perturbs a passage and checks whether the original text has unusually high likelihood relative to its perturbations.

DNA-GPT (Yang et al., 2024). DNA-GPT compares the original continuation of a truncated passage with model-generated continuations, using N-gram divergence in the black-box setting or probability divergence in the white-box setting. A higher discrepancy score suggests stronger evidence of LLM generation.

Fast-DetectGPT (Bao et al., 2024). Fast-DetectGPT replaces DetectGPT’s perturbation step with a more efficient sampling procedure. We follow the reported setting using GPT-Neo-2.7B as the scoring model and GPT-J-6B (Wang and Komatsuzaki, 2021) as the sampling model.

RoBERTa Classifier (Liu et al., 2019). The RoBERTa baseline represents supervised detector training on labeled text. It is included as a competitive discriminative baseline against zero-shot statistical methods and the guideline-based AGENT-XTrust detector.

A.3 Detailed Detector Accuracy

The following tables expand the main benchmark summary in Table 1 by domain and generator.

A.3.1 Baseline Threshold Selection

For score-based baselines that return a continuous statistic rather than a binary label, thresholds are selected only on a separate validation split and then frozen before test evaluation. For each baseline and score orientation, we scan the observed validation-score range and choose the threshold that maximizes validation-set binary accuracy. The scan uses uniformly spaced candidate thresholds with step size equal to 1% of the observed validation-score span. The selected threshold is then applied unchanged to the independent test split for the corresponding reported accuracy. The validation set is the SQuAD-GPT-Neo-2.7B split, and the key fairness constraint is that no threshold is fit on the test examples summarized in Tables 5–6. AGENT-XTrust does not use this validation-threshold procedure for its main binary decision; it reports a direct MetaAgent output without target-domain scalar-cutoff calibration under the fixed guideline inventory and prompting setup.

A.4 Additional AUROC Results

B Method Details

This appendix collects the guideline inventory, construction rationale, prompts, and routing statistics for AGENT-XTrust.

B.1 Detection Guidelines

The reported detector uses a fixed inventory of detection guidelines organized into content, expression, and structure cues. The inventory is adapted from publicly documented Wikipedia editorial guidance, and each cue is treated as contextual evidence generalized from Wikipedia editorial guidance: it supports an AI-writing judgment when the corresponding pattern is expressed in the input and is aggregated with other activated cues. The cue titles and compact descriptions are listed below.

Detection Guidelines

Content / Semantic Evidence

WC1: Broad significance, legacy, and trend framing
Cue: LLM writing often adds broad statements about why a subject matters, how it symbolizes a larger trend, or how it contributes to a wider context with limited concrete support.
Watch words / patterns: stands/serves as; is a testament/reminder;
vital/significant/crucial/pivotal/key role/moment;
underscores/highlights its
importance/significance; reflects broader; symbolizing its...

WC2: Notability, attribution, and media-coverage framing
Cue: LLMs act as if the best way to prove that a subject is notable is to emphasize claims of notability, often by listing sources that a subject has been covered in.
Watch words / patterns: independent coverage; local/regional/national/[country name] media outlets; music/business/tech outlets; profiled in; written by a leading expert; active social media presence

WC3: Superficial analyses
Cue: AI chatbots tend to insert superficial analysis of information, often in relation to its significance, recognition, or impact.
Watch words / patterns: highlighting/underscoring/emphasizing ...; ensuring ...; reflecting/symbolizing ...; contributing to ...; cultivating/fostering ...; encompassing ...; valuable insights;...

WC4: Promotional and advertisement-like language
Cue: LLM outputs can exhibit non-neutral or promotional tone.
Watch words / patterns: boasts a; vibrant; rich; profound; enhancing; showcasing; exemplifies; commitment to; natural beauty; nestled; in the heart of; groundbreaking; renowned; featuring; diverse array;...

WC5: Vague attributions and overgeneralization of opinions
Cue: AI chatbot outputs may attribute opinions or claims to vague authorities, a pattern often described as weasel wording.
Watch words / patterns: Industry reports; Observers have cited; Experts argue; Some critics argue; several sources/publications; such as; ...

WC6: Outline-like conclusions about challenges and future prospects
Cue: Many LLM-generated Wikipedia articles include a "Challenges" section, which typically begins with a sentence like "Despite its [positive/promotional words], [article subject] faces challenges..." and ends with either a vaguely positive assessment of the article subject, or speculation about how ongoing or potential initiatives could benefit the subject.
Watch words / patterns: Despite its... faces several challenges...; Despite these challenges; Challenges and Legacy; Future Outlook; ...

WC7: Leads treating Wikipedia lists or broad article titles as proper nouns; vague see also sections
Cue: In AI-generated articles about topics with a title that is not a proper name, such as a list, the first sentence of the lead may introduce and/or define the article's title as if it were a standalone real-world entity.
Watch words / patterns: refers to; is the; is a curated compilation of; is the chronological list of; ...

Expression / Stylistic Evidence

WE1: High density of "AI vocabulary" words
Cue: Many studies have demonstrated that LLMs overuse specific words.
Watch words / patterns: Additionally; align with; boasts; bolstered; crucial; delve;

Method	GPT-3.5				GPT-4			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
RoBERTa-base	83.67	60.00	52.67	65.45	62.42	49.67	43.96	52.02
RoBERTa-large	65.67	50.33	48.00	54.67	53.36	49.33	47.32	50.00
Likelihood (Neo-2.7)	80.00	91.33	76.33	<u>82.56</u>	71.00	60.67	<u>72.33</u>	68.00
Entropy (Neo-2.7)	38.00	37.67	35.33	37.00	50.00	47.67	39.67	45.78
Rank (Neo-2.7)	67.67	60.67	58.00	62.11	60.00	51.00	57.67	56.22
LogRank (Neo-2.7)	83.00	89.00	75.33	82.44	<u>72.67</u>	59.00	72.00	67.89
LRR (Neo-2.7)	84.33	73.00	57.67	71.67	68.46	54.67	55.70	59.61
DNA-GPT (Neo-2.7)	71.00	84.67	50.00	68.56	66.00	71.00	50.00	62.33
NPR (T5-11B/Neo-2.7)	70.33	88.67	60.00	73.00	49.66	61.00	55.37	55.34
DetectGPT (T5-11B/Neo-2.7)	64.67	76.67	49.67	63.67	50.67	50.33	50.00	50.33
Fast-Detect (GPT-J/Neo-2.7)	94.67	93.33	54.33	80.78	71.00	<u>76.00</u>	52.67	66.56
AGENT-XTrust	<u>89.67</u>	<u>92.33</u>	<u>76.04</u>	86.01	86.24	87.05	84.46	85.92

Table 5: Detector benchmark accuracy (percent) on GPT-3.5 and GPT-4 generations. Baseline thresholds are optimized on a separate validation set. Highest values are **bold**; second-highest values are underlined.

Method	Claude-3 Opus				Claude-3 Sonnet			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
RoBERTa-base	79.67	63.00	43.33	62.00	70.00	57.00	44.33	57.11
RoBERTa-large	66.00	51.33	43.67	53.67	58.67	49.00	43.33	50.33
Likelihood (Neo-2.7)	80.33	90.00	75.00	81.78	78.33	83.67	<u>73.67</u>	<u>78.56</u>
Entropy (Neo-2.7)	42.33	33.67	38.00	38.00	46.67	39.00	37.33	41.00
Rank (Neo-2.7)	66.00	64.67	57.33	62.67	62.00	61.33	58.33	60.56
LogRank (Neo-2.7)	81.67	88.67	<u>75.67</u>	<u>82.00</u>	<u>79.33</u>	82.00	72.33	77.89
LRR (Neo-2.7)	81.67	81.00	63.33	75.33	73.67	73.33	65.67	70.89
DNA-GPT (Neo-2.7)	<u>82.00</u>	86.67	50.00	72.89	77.33	84.67	50.00	70.67
NPR (T5-11B/Neo-2.7)	74.00	<u>92.33</u>	69.00	78.44	73.33	90.67	68.00	77.33
DetectGPT (T5-11B/Neo-2.7)	50.00	50.00	65.67	55.22	50.00	50.00	65.67	55.22
Fast-Detect (GPT-J/Neo-2.7)	87.67	94.00	57.00	79.56	80.67	<u>87.00</u>	53.00	73.56
AGENT-XTrust	81.33	91.00	79.00	83.78	75.67	86.33	78.00	80.00

Table 6: Detector benchmark accuracy (percent) on Claude-3 Opus and Claude-3 Sonnet generations. Baseline thresholds are optimized on a separate validation set. Highest values are **bold**; second-highest values are underlined.

emphasizing; enduring; enhance; fostering; garner; highlight; interplay; intricate/intricacies; key; landscape;...

WE2: Avoidance of basic copulatives ("is"/"are" phrases)
Cue: LLM-generated text often substitutes constructions like "serves as a" or "marks the" for their simpler counterparts that use copulas such as "is" or "are".
Watch words / patterns: serves as/stands as/marks/represents [a]; boasts/features/offers [a]

WE3: Negative parallelisms
Cue: When LLMs describe a subject, their output may seem as though it is clearing up a common misconception, or as though the audience may be reaching an incomplete or incorrect conclusion about that subject.
Watch words / patterns: not only ... but ...; it is not just ..., it's ...; not ..., but ...; however; no ..., no ..., just ...

WE4: Rule of three
Cue: LLMs overuse the "rule of three".

Watch words / patterns: adjective, adjective, adjective; phrase, phrase, and phrase

WE5: Elegant variation
Cue: Generative models often avoid obvious repetition, which can lead them to vary references to the same entity in a conspicuously polished way.
Watch words / patterns: protagonist; key player; eponymous character

WE6: Title case
Cue: In section headings, AI chatbots strongly tend to capitalize all main words.
Watch words / patterns: Capitalize All Main Words

WE7: Overuse of boldface
Cue: AI chatbots may display various phrases in boldface for emphasis in an excessive, mechanical manner.
Watch words / patterns: boldface; emphasis; key takeaways

WE8: Inline-header vertical lists

Method	GPT-3.5 AUROC (% , $\uparrow$)				GPT-4 AUROC (% , $\uparrow$)			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
RoBERTa-base	91.50	70.84	61.88	74.74	67.78	50.68	53.09	57.18
RoBERTa-large	85.07	54.80	67.31	69.06	68.79	38.21	60.67	55.89
Likelihood	95.78	97.40	<u>87.75</u>	<u>93.64</u>	79.80	85.53	81.04	82.12
Entropy	33.05	19.02	27.67	26.58	43.60	37.02	32.95	37.86
Rank	74.94	80.64	59.79	71.79	66.44	71.46	59.65	65.85
LogRank	95.82	96.56	86.87	93.08	79.75	82.86	80.03	80.88
LRR	91.62	89.58	74.33	85.18	74.47	70.28	68.14	70.96
DNA-GPT	91.24	94.25	79.59	88.36	73.47	80.32	75.65	76.48
NPR	78.99	89.24	67.84	78.69	52.80	61.22	63.28	59.10
DetectGPT	84.16	88.11	74.44	82.23	56.60	62.17	68.05	62.28
Fast-Detect	99.07	99.16	90.21	96.15	<u>90.67</u>	96.12	85.03	90.61
AGENT-XTrust	<u>96.28</u>	<u>97.94</u>	81.95	92.06	93.67	<u>92.06</u>	<u>84.47</u>	<u>90.07</u>

Method	Claude-3 Opus AUROC (% , $\uparrow$)				Claude-3 Sonnet AUROC (% , $\uparrow$)			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
RoBERTa-base	89.75	71.15	40.09	67.00	75.11	57.88	37.99	56.99
RoBERTa-large	81.46	55.48	37.98	58.31	67.25	47.72	38.56	51.18
Likelihood	93.22	97.34	<u>86.03</u>	<u>92.20</u>	88.62	94.84	<u>83.60</u>	<u>89.02</u>
Entropy	38.71	17.92	29.10	28.58	41.46	21.56	29.89	30.97
Rank	73.33	79.50	60.80	71.21	70.19	78.12	60.17	69.49
LogRank	93.57	96.79	85.08	91.81	88.67	94.01	82.96	88.55
LRR	89.56	91.78	74.48	85.27	83.59	87.46	74.36	81.80
DNA-GPT	94.24	96.53	78.06	89.61	85.58	94.15	76.47	85.40
NPR	84.26	<u>97.64</u>	80.94	87.61	83.49	<u>96.31</u>	79.42	86.41
DetectGPT	77.18	83.35	77.52	79.35	81.50	86.75	73.47	80.57
Fast-Detect	97.79	98.32	89.47	95.19	95.14	97.63	86.34	93.04
AGENT-XTrust	90.42	96.75	85.26	90.81	84.43	92.31	83.24	86.66

Cue: AI chatbots output often includes vertical lists formatted in a specific way: an ordered or unordered list where the list marker (number, bullet, dash, etc.) is followed by an inline boldfaced header, separated with a colon from the remaining descriptive text.

Watch words / patterns: 1.; *; -; -; #; emoji; :

WE9: Emoji

Cue: AI chatbots often use emoji.

Watch words / patterns: ; ; ; ; ; ; ; ; ; ;

WE10: Overuse of em dashes

Cue: While human editors and writers often use em dashes (–), LLM output uses them more often than nonprofessional human-written text of the same genre, and uses them in places where humans are more likely to use commas, parentheses, colons, or (misused) hyphens (–).

Watch words / patterns: -

WE11: Unusual use of tables

Cue: AI chatbot outputs sometimes include small tables where prose would be more natural.

Watch words / patterns: Metric; Figure

WE12: Curly quotation marks and apostrophes

Cue: ChatGPT and DeepSeek typically use curly quotation marks ("..." or '...') instead of straight quotation marks ("..." or '...'). Watch words / patterns: " "; ' '; '

Watch words / patterns: " "; ' ';

WE13: Subject lines

Cue: User messages and unblock requests generated by AI chatbots sometimes begin with text that is intended to be pasted into the Subject field on an email form.

Watch words / patterns: Subject:

WE14: Skipping heading levels

Cue: AI chatbots tend to skip level 2 headings (==) and start sections from the third level (===).

Watch words / patterns: ==

WE15: Thematic breaks before headings

Cue: AI chatbots sometimes include a thematic break (----) before each heading in a text (this is common in Markdown output).

Watch words / patterns: ---

WE16: Collaborative communication

Cue: Editors sometimes paste text from an AI chatbot that was meant as correspondence.

prewriting or advice, rather than article content.
Watch words / patterns: I hope this helps; Of course!; Certainly!; You're absolutely right!;
Would you like; is there anything else; let me know;
more detailed breakdown; here is a

WE17: Sudden shift in writing style

Cue: A sudden shift in an editor's writing style, such as unexpectedly flawless grammar compared to their other communication, may indicate the use of AI tools.

Watch words / patterns: unexpectedly flawless grammar; formal prose; American English; UK English

WE18: Overwhelmingly verbose edit summaries

Cue: AI-generated edit summaries are often unusually long, written as formal, first-person paragraphs without abbreviations, and/or conspicuously itemize Wikipedia's conventions.

Watch words / patterns: I revised; I formalized; ensured neutrality; improved clarity; encyclopedic standards; tone was adjusted

Structure / Discourse-Structural Evidence

WS1: Collaborative communication

Cue: Editors sometimes paste text from an AI chatbot that was meant as correspondence, prewriting or advice, rather than article content.

WS2: Knowledge-cutoff disclaimers and speculation about gaps in sources

Cue: A knowledge-cutoff disclaimer is a statement used by the AI chatbot to indicate that the information provided may be incomplete, inaccurate, or outdated.

WS3: Phrasal templates and placeholder text; use of Markdown

Cue: AI chatbots may generate responses with fill-in-the-blank phrasal templates for the user to replace later, but some users forget to fill in the blanks.

WS4: Broken wikitext and markup artifacts

Cue: Since AI chatbots are typically not proficient in wikitext and templates, they often produce faulty syntax.

WS5: Non-existent or out-of-place categories and non-existent templates

Cue: LLMs may hallucinate non-existent categories, sometimes for generic concepts that seem like plausible category titles, and sometimes because their training set includes obsolete or renamed categories.

WS6: Broken external links, invalid DOI and ISBNs, outdated access-dates, and unconventional references

Cue: If a new article or draft has multiple citations with external links and several of them are broken, this is a strong sign of an AI-generated page.

WS7: Miscellaneous workflow signals

Cue: Outside the prose itself, AI use may show up as a sudden shift in writing style, overwhelmingly verbose edit summaries, 'submission statements' in AFC drafts, or pre-placed maintenance templates.

WS8: Signs of human writing and ineffective indicators

Cue: Individual signs are treated as contextual rather than conclusive evidence of AI authorship.

B.2 Guideline Basis

B.2.1 Guideline Construction

The reported system uses the Wikipedia-aligned inventory in Appendix B.1. We use this inventory because it has explicit provenance, is publicly documented and inspectable, and is written around observable textual phenomena rather than opaque detector scores. All main evaluations use this fixed inventory; alternative inventory variants are reserved for ablation analysis.

The guideline inventory is constructed in three stages.

(1) Source cue extraction. We begin from Wikipedia's community-maintained "Signs of AI writing" page (Wikipedia contributors, 2026). Its value for this work is its public, revision-linked collection of editorial observations and concrete examples from contributors who inspect generated and human-written prose. The resulting cues are concrete: undue emphasis on significance, vague attribution, generic promotional phrasing, over-balanced exposition, formatting artifacts, and repetitive paragraph structure.

(2) Guideline normalization. We convert the source cues into a stable inventory with identifiers, titles, short descriptions, watch patterns, and examples. The normalization step preserves the source meaning but makes each cue suitable for an expert agent: the expert is asked whether the cue is expressed in the current input, to quote or identify supporting spans when possible, and to avoid treating the cue as decisive in isolation.

(3) Routing alignment. The normalized guidelines are grouped into content, expression, and structure dimensions, corresponding to the semantic, stylistic, and discourse-structural evidence types used in the main paper. This architectural grouping supports routing and MetaAgent comparison across heterogeneous evidence types; domain sensitivity is handled by allowing the router to leave irrelevant cues inactive.

B.2.2 Guideline Foundations

The guidelines are *source-grounded diagnostic cues*: concrete textual phenomena with public provenance, editorial motivation, and direct inspectability. A single phrase such as "plays a vital

Source cue family	Norm. ID	Dimension	Watch pattern	Example evidence
Inflated significance / legacy framing	WC1	Content	Broad claims that a topic “serves as” a symbol or vital contribution	“stands as a vital cultural landmark”
Notability and vague attribution	WC2/WC5	Content	Claims backed by vague media coverage, experts, observers, or unspecified sources	“Observers have cited...”
Superficial analysis and promotional tone	WC3/WC4	Content	Generic impact statements, celebratory adjectives, or advertisement-like praise	“showcasing a vibrant commitment”
High-density AI vocabulary	WE1	Expression	Overused lexical cues such as “crucial”, “delve”, “fostering”, or “landscape”	“plays a crucial role in the broader landscape”
Formulaic contrast and parallelism	WE3/WE4	Expression	“not only...but also” structures, rule-of-three phrasing, and balanced list rhythm	“not only preserves, but also inspires”
Formatting artifacts	WE6–WE15	Expression	Title-case headings, bolded inline headers, tables, Markdown breaks, or unusual quotation marks	“Key Takeaways:” followed by bolded bullets
Template / assistant residue	WE16, WS1, WS3	Structure	Pasted chat, placeholders, collaborative phrasing, or fill-in-the-blank templates	“Would you like me to...”
Reference and workflow artifacts	WS4–WS7	Structure	Broken markup, impossible categories, invalid links, outdated access dates, or workflow residue	malformed wikitext or non-existent categories

Table 9: Representative mapping from source cue families to the normalized guideline inventory. The full inventory is listed in Appendix B.1; this table shows how source observations are converted into expert-readable cue identifiers, dimensions, watch patterns, and evidence examples.

role” can appear in human writing, and polish is treated as contextual evidence only when supported by other activated cues. The scientific claim is distributional and ensemble-based: modern LLMs often overproduce recurring rhetorical, attributional, and structural patterns, and AGENT-XTrust makes these patterns inspectable by aggregating multiple routed, contextual cues.

The foundation has four parts. First, the inventory has external provenance: the cue families come from a public, community-maintained editorial source with documented provenance (Wikipedia contributors, 2026). Second, the cue families align with authorship-attribution assumptions that writers reveal themselves through recurring lexical, syntactic, rhetorical, and semantic choices (Koppel et al., 2009). Third, prior work suggests that frequent ChatGPT users detect AI-generated text robustly and explain judgments with vocabulary, formality, originality, clarity, and stylistic cues (Russell et al., 2025); this supports operationalizing such cues as human priors, not proof. Fourth, the detector aggregates multiple contextual cues rather than treating any single indicator as decisive.

The guidelines serve as a transparent, inspectable evidence vocabulary rather than a private prompt list. Their role is to make the detector’s reasoning inspectable: which cue fired, what span supported it, how confident the local expert was, and how the MetaAgent reconciled agreement or disagreement across cues. Routing then handles context sensitivity by activating the cues that match the input domain and style.

Content / semantic evidence. Content cues capture what the passage chooses to say: exaggerated significance, notability claims, vague sourcing, superficial analysis, generic balanced framing, and low-specificity claims. These cues are grounded in the observation that generated text often fills informational gaps with broadly plausible but under-specified explanation. They connect naturally to authorship-attribution work on semantic idiosyncrasy: human text often contains uneven, locally grounded detail, whereas generated text often drifts toward generic, consensus-like meaning.

Expression / stylistic evidence. Expression cues capture how the passage says it: AI-vocabulary density, promotional adjectives, hedging formulas,

formulaic transitions, over-polished neutrality, and characteristic phrase templates. These are not mere word blacklists. The expert must judge whether the wording pattern functions as evidence in context. For example, promotional language is more suspicious in an encyclopedic entry than in an advertisement; routing and meta aggregation are meant to account for that genre dependence.

Structure / discourse evidence. Structure cues capture organization-level regularities: collaborative chat phrasing, list-like formatting, invented sections, over-regular paragraph development, and conclusion-style summaries. These cues reflect discourse planning. LLM outputs often have highly legible structure because they are optimized to satisfy instruction-following expectations, while human writing can be more uneven, elliptical, or genre-specific.

Overall, the guideline foundation is source-grounded and publicly curated. AGENT-XTrust’s contribution is to integrate this evidence vocabulary into a routed expert ensemble with confidence, rationales, span evidence, and meta-level reconciliation.

B.3 Prompts

We provide the prompts used in our experiments, including the router agent (Prompts B.1, B.2), MetaAgent (Prompt B.3), and expert agent prompt (Prompt B.4).

Prompt B.1: Router Agent Domain and Style Analysis

You are an expert stylistic analyzer. Analyze the given text and clearly state:

1. The text’s domain (e.g., medical abstract, literary fiction, news article, etc.).
2. Key stylistic features that strongly characterize the text.

Input Text:
“<input text>”

Respond strictly in the following JSON format:

```
{
  "Domain": "<identified domain>",
  "Stylistic Features": ["<feature 1>", "<feature 2>", "..."]
}
```

Prompt B.2: Router Agent Guideline Selection

Given a text’s domain and stylistic features, your task is to strictly select **ONLY** the most relevant stylistic guidelines from the provided set for the “<aspect>” aspect.

IMPORTANT INSTRUCTIONS:

- Be extremely conservative in your selection.
- Activate a guideline **ONLY** if it clearly and strongly matches **BOTH** the domain and the listed stylistic features.
- If unsure or only moderately relevant, do **NOT** activate.

- Typically, each aspect should have very few (often 1–5) activated guidelines.

Text Domain: “<identified domain>”

Stylistic Features:

- <feature 1>
- <feature 2>

...

“<aspect>” Guidelines:

G1 <guideline description>

G2 <guideline description>

G3 <guideline description>

... [...]

Output **STRICTLY** as JSON:

```
{
  "Activated Guidelines": [
    "<Guideline ID 1>",
    "<Guideline ID 2>", "..."]
}
```

Confidence-steering substitutions. For both expert and MetaAgent calls, the steering placeholders below are instantiated with the same fixed five-level set: *very cautious* / “make your confidence low,” *cautious* / “somewhat low,” an unqualified condition / “report your true confidence,” *confident* / “somewhat high,” and *very confident* / “high.” All five variants retain the same task text and JSON schema.

Prompt B.3: MetaAgent

System: You are <confidence steering condition>. Provide a concise rationale, then reply only in JSON format:

```
{
  "reasoning": "<brief rationale>",
  "final_decision": "AI" or "Human",
  "confidence": <number between 0 and 1>
}
```

Make your confidence <confidence steering instruction>.

User: You are a Meta Agent responsible for summarizing and integrating responses from multiple aspect-specific Base Agents. No prior guidelines are provided; rely solely on the provided Base Agent outputs.

The text to evaluate is:

Text: “<input text>”

Base Agent Responses by Aspect:

Aspect <aspect 1>:

- Decision: [AI/Human], Confidence: <confidence score>

- Reasoning: “<reasoning>”

Aspect <aspect 2>:

- Decision: [AI/Human], Confidence: <confidence score>

- Reasoning: “<reasoning>”

...

Aspect <aspect N>:

- Decision: [AI/Human], Confidence: <confidence score>

- Reasoning: “<reasoning>”

Provide a final integrated reasoning and decision, explicitly referencing all provided Base Agent decisions, confidence scores, and reasoning. Pay particular attention to Base Agents with higher confidence; they should have greater influence on your final decision. If most agents agree, trust that direction more strongly. If disagreements arise, carefully consider both confidence levels and the number of supporting agents.

Respond strictly in JSON without extra explanations:

```
{
  "reasoning": "(Integrated reasoning referencing provided text, all aspects, and base agent outputs)",
}
```

Dataset	Depth	Sem.	Styl.	Struct.
XSum	4.2	35.68%	43.52%	20.80%
PubMedQA	5.1	38.89%	39.62%	21.49%
WritingPrompts	4.6	19.99%	40.07%	39.95%

Table 10: Routing statistics. Depth is the mean number of activated guideline experts per instance; the remaining columns report the proportion of activated experts from each linguistic dimension.

```
"final_decision": "AI" or "human",
"confidence": "( between 0 and 1 )"
}
```

Prompt B.4: Expert Agent

System: You are <confidence steering condition>. Provide a concise rationale, then reply only in JSON format:

```
{
  "reasoning": "<brieﬂ rationale>",
  "final_decision": "AI" or "Human",
  "confidence": <number between 0 and 1>
}
```

Make your confidence <confidence steering instruction>.

User: Guidelines for distinguishing AI-generated from human-written texts:
 <guideline text>
 Evaluate the following text. Determine if it is human-written or AI-generated:
 Text: "<input text>"

Please strictly respond in JSON without any additional explanations:

```
{
  "reasoning": "(Provide detailed reasoning explicitly addressing the
Content aspect, clearly justifying your decision and referencing specific
examples from the provided text.)",
  "final_decision": "AI" or "human",
  "confidence": "( between 0 and 1 )"
}
```

B.4 Routing Statistics

C Ablation Details

This appendix provides supplementary ablations and examples illustrating how the detector and evidence interface behave on concrete inputs.

C.1 Ablations

C.1.1 Component Ablations

Table 2 in the main text reports the averaged component ablations because they directly support the detector architecture rather than serving only as supplementary detail. Table 11 provides the corresponding per-domain breakdown.

C.1.2 Case Study

Figure 3 illustrates how the detector exposes intermediate evidence rather than only returning a score. The highlighted spans and quoted rationales are presented as illustrative evidence cues for human inspection. The router identifies the input domain and style, activates relevant semantic, stylistic, and structural guidelines, and each guideline expert produces a local decision with a confidence value and rationale. The MetaAgent then integrates

both agreement and disagreement across experts. This example is important for the paper’s decision-support framing: a user can inspect which cues support the decision, where experts disagree, and whether the final label is driven by broad consensus or by a small number of high-confidence cues.

Method	Accuracy (% , $\uparrow$)				AUROC (% , $\uparrow$)			
	XSum	Writing	PubMed	Avg.	XSum	Writing	PubMed	Avg.
Full AGENT-XTrust	<u>86.24</u>	87.05	84.46	85.92	93.67	92.06	84.47	90.07
w/o Multi-Agent	86.91	79.67	69.46	78.68	<u>93.57</u>	86.74	<u>83.23</u>	<u>87.85</u>
w/o Guidelines	77.85	56.33	50.00	61.39	93.19	70.78	70.78	78.25
w/o LLM Router	80.54	80.67	54.70	71.97	86.52	83.41	62.21	77.38
w/o Adaptive Routing	86.91	<u>82.67</u>	58.05	75.88	92.36	88.54	79.06	86.65
w/o Confidence Steering	85.90	81.00	<u>72.82</u>	<u>79.91</u>	91.44	<u>89.14</u>	73.83	84.80

Table 11: Per-domain ablation results on GPT-4 generations. Accuracy and AUROC are percentages. Highest values are **bold**; second-highest values are underlined.



Figure 3: Example AGENT-XTrust detection case. The router activates guideline-specific experts based on inferred textual characteristics; each expert provides a local decision, confidence value, and linguistic rationale; the MetaAgent aggregates the evidence into an interpretable direct decision without target-domain scalar-cutoff calibration.

C.2 Detection Case

Figure 3 and the case below illustrate the evidence interface for the detection guideline set. The example shows activated cue identities, local rationales, confidence values, and meta-level reconciliation.

Detection Case
Input Text The museum stands as a vital cultural landmark, showcasing the city's rich heritage while serving as a testament to the community's enduring commitment to education and preservation. Its diverse exhibits offer visitors a unique glimpse into the region's vibrant past, highlighting the important role the institution plays in shaping local identity.
Router Assessment The router identifies the passage as an encyclopedic description of a cultural institution, with promotional adjectives, legacy framing, generic significance claims, and a smooth summary style. It activates content cues WC1, WC3, and WC4; expression cues WE1 and WE2; and the structural counter-signal WS8.
WC1: Broad significance framing Reasoning: The phrases "stands as a vital cultural landmark", "serving as a testament", and "enduring commitment" inflate the subject's significance through broad legacy language rather than concrete evidence. Decision: AI, Confidence: 0.80
WC3: Superficial analyses Reasoning: The text adds generic interpretive claims such as "shaping local identity" without attributing a source or giving specific factual support. Decision: AI, Confidence: 0.76
WC4: Promotional and advertisement-like language Reasoning: Phrases such as "rich heritage", "diverse exhibits", "unique glimpse", and "vibrant past" resemble promotional description more than neutral encyclopedic prose. Decision: AI, Confidence: 0.82
WE1: High density of AI-vocabulary words Reasoning: The passage concentrates common generated-text vocabulary: "vital", "showcasing", "rich", "enduring", "diverse", "unique", "vibrant", and "highlighting". Decision: AI, Confidence: 0.78
WE2: Avoidance of basic copulatives

Reasoning: The sentence uses "stands as" and "serving as" where simpler copular phrasing would be more direct, matching the guideline's expression-level cue.
Decision: AI, Confidence: 0.70

WS8: Human-writing counter-signals

Reasoning: The expert finds no strong human-specific counter-signal. The text is fluent, but fluency alone is not evidence of human authorship.
Decision: AI, Confidence: 0.58

MetaAgent

Reasoning: Multiple content and expression cues agree: inflated significance, promotional phrasing, superficial interpretation, and high-density AI vocabulary all support an AI-writing judgment. The MetaAgent treats these as defeasible but mutually reinforcing evidence rather than as a single decisive phrase match.
Decision: AI, Confidence: 0.79

D Evidence Audit

This appendix reports supplementary rationale-quality audits and the manual evidence evaluation protocol used to assess whether produced evidence records support their intended decision-support role.

D.1 Rationale Audit

This appendix reports an auxiliary audit of rationale quality for the detector backbone. This auxiliary audit characterizes rationale quality for the detector backbone; the primary claims rely on the benchmark evaluation, evidence interface, and controlled user study.

D.1.1 Auxiliary Rationale Audit

This subsection summarizes an auxiliary rationale-quality audit for the detector backbone as complementary evidence to the AGENT-XTrust span-localization audit in Appendix D.2. The audit is descriptive and is not used to claim mechanistic explanation faithfulness.

The audit samples 100 detection cases from XSum, WritingPrompts, and PubMed. Each case includes the input text, classification, and explanation. Rationale quality is evaluated along two reader-facing dimensions: *correctness*, which measures whether the rationale describes linguistic features present in the text, and *usefulness*, which measures whether it helps a reader understand the model's stated decision. These dimensions follow prior XAI work emphasizing explanation evaluation against user-facing criteria (Poursabzi-Sangdeh et al., 2021; Doshi-Velez and Kim, 2017; Mohseni et al., 2021).

The audit judges 86% of explanations correct, 79% useful, and 74% as satisfying both criteria.

Qualitative analysis shows that stronger rationales tend to highlight observable cues such as hedging, structural repetition, or sentence rhythm, whereas weaker rationales rely on vague or generic statements. Overall, this auxiliary audit suggests that the detector can produce useful local rationales, while the stricter span-level evidence layer is evaluated separately below.

D.2 Manual Evidence Evaluation

To assess whether system-produced evidence records align with their intended support function, we conduct a manual evaluation over sampled AGENT-XTrust outputs. The evaluation is separate from the decision-support study in Appendix E.1: it measures evidence understandability, grounding, guideline relevance, and unsupported-claim avoidance, rather than participant decisions or trust. This isolates the reliability of the rationale layer from the behavioral question of how end users respond to that layer. Two annotators independently label all sampled records; we report inter-annotator agreement before consensus and use conservative consensus labels for the final rates.

D.2.1 Sampling

The audit samples 90 evidence records, with 30 from each evaluation domain: XSum, Writing-Prompts, and PubMedQA. Records are stratified by ground-truth label, detector correctness, confidence, and evidence agreement so that the audit includes easy true positives, easy true negatives, detector errors, and cases where experts disagree. The stratified sample includes uncertain and detector-error cases to estimate evidence quality beyond high-confidence correct predictions.

For each sampled record, the annotation sheet includes (i) the original passage, (ii) the system’s document-level label and confidence, (iii) the selected guideline name and dimension, (iv) the local rationale, and (v) the highlighted span or quoted evidence shown to the user. Detector correctness is used only for aggregate error analysis, not as the target of the evidence-quality labels.

D.2.2 Labels

Each record is labeled along four dimensions.

Understandability. This label records whether the rationale is clear enough for a reader to understand what textual property the system is pointing to.

Grounding. This label records whether the rationale refers to concrete textual evidence present in the passage rather than unsupported generalities.

Span relevance. This label records whether the highlighted span or quoted evidence supports the rationale and the cited guideline.

No hallucination. This label records whether the rationale avoids factual claims, stylistic descriptions, or source attributions not supported by the passage.

Labels are recorded as YES, PARTIAL, or NO. YES defines the strict success rate, while YES+PARTIAL defines a permissive support rate. The audit is intended as a descriptive quality check of displayed evidence rather than a mechanistic faithfulness benchmark. For final reporting, a strict YES requires both annotators to mark YES; any NO is treated as NO; remaining disagreements are treated as PARTIAL. The resulting success rates and error categories are presented in Table 13.

D.2.3 Reporting

We summarize annotation rates over understandability, concrete grounding, span relevance, and unsupported-claim avoidance, using both strict and permissive rates. Inter-annotator agreement before consensus is reported in Table 12. Table 13 reports the strict consensus rates overall, by source domain, and by detector correctness. The important failure mode is localization: most records are understandable and grounded, but the strict all-four success rate is limited by spans that are too narrow or misaligned with the rationale. This pattern indicates that the evidence layer usually provides usable textual support, while span selection remains the main target for improvement. Qualitative examples cover a clearly grounded rationale, a partially grounded rationale where the span is relevant but incomplete, and a failure case where the rationale sounds plausible but the span does not support it.

E Human Study

This appendix collects the study interface, comparison conditions, recruitment and payment details, task procedure, outcome metrics, analyzed sample, and supplementary metric figure.

E.1 Study Interface

The study compares three complete assistance conditions: Fast-DetectGPT score-only, a direct-LLM

Dimension	Agreement ($\uparrow$)	Cohen’s κ ($\uparrow$)
Understandability	100.0	–
Grounding	100.0	–
Span relevance	97.8	0.89
No hallucination	100.0	1.00

Table 12: Inter-annotator agreement for the evidence audit before conservative consensus. Across all 360 label decisions, raw agreement is 99.4% and pooled nominal Cohen’s κ is 0.91. Arrows indicate the preferred direction: higher agreement and higher κ indicate stronger annotator consistency. The agreement column reports raw percent agreement; dashes appear only in the κ column for degenerate all-YES dimensions, where chance-corrected agreement is not informative.

score plus global explanation, and the full AGENT-XTrust package with span highlights, guideline citations, local rationales, and confidence. The interface records participant decisions, response time, optional notes, and standard quality-control metadata. The analyzed sample and condition-level decision outcomes are reported in Table 4 and the appendix tables below.

E.1.1 Study Comparisons

The study evaluates complete detector-and-interface assistance packages for human judgment under uncertainty. The analysis compares the full AGENT-XTrust package with Fast-DetectGPT score-only and direct-LLM score-plus-explanation packages on human use of detector outputs.

Accuracy. The primary comparison measures whether participants assigned to the full AGENT-XTrust condition make more correct AI/human decisions than participants assigned to the Fast-DetectGPT score-only condition.

False accusation. The false-accusation comparison measures AI accusations on human-written texts, especially when the detector score is high but the supporting evidence is sparse or conflicting.

Overtrust. The overtrust comparison measures agreement with detector-wrong outputs and whether participants challenge the system when highlighted spans and rationales are unconvincing.

Deferral. The deferral comparison separates uncertainty on ambiguous items from indiscriminate deferral on easy items.

E.1.2 Experimental Design

The study uses a between-subject design with three assistance conditions. The conditions differ in both detector identity and presentation while keeping the passage, task, response options, and stimulus-order randomization fixed. The resulting estimand is the contrast among complete assistance packages, not an isolated interface effect.

Score only. Participants see the text and a Fast-DetectGPT output: a binary AI/human label and a single AI-likelihood score. The displayed score is computed from the Fast-DetectGPT sampling-discrepancy value after applying a frozen SQuAD/GPT-Neo out-of-distribution validation threshold and sigmoid normalization; the threshold is not tuned on the three study datasets. This condition represents a strong scalar-detector interface.

Score plus explanation. Participants see the text and an independent direct-LLM judgment consisting of a binary label, a single AI-likelihood score, and one concise free-form global explanation. The global explanation is generated by prompting the LLM to judge only the passage itself, without AGENT-XTrust guidelines, routed evidence, detector outputs, span highlights, or external metadata. The displayed score in this condition is the direct LLM judge’s confidence transformed into AI-likelihood, not the Fast-DetectGPT score. This condition is a complete direct-LLM explanation package, not an interface-only explanation ablation.

Full AGENT-XTrust evidence. Participants see the AGENT-XTrust MetaAgent judgment and score together with the full evidence interface: a document-level explanation, highlighted spans, guideline citations, local guideline rationales, and confidence values. This condition therefore evaluates the full deployed evidence package rather than isolating any single interface component.

Comparability. The middle condition is a global-rationale comparator rather than a length-matched version of AGENT-XTrust. It provides a concise natural-language explanation but omits activated guideline identities, guideline-specific local labels, local confidence values, cited rationales, source-linked highlights, and span–guideline links. Because it also uses a different detector, the comparison estimates the difference between complete packages and cannot identify the causal effect of

Subset	n	Understandable ($\uparrow$)	Grounded ($\uparrow$)	Span relevant ($\uparrow$)	No hallucination ($\uparrow$)	All four ($\uparrow$)
Overall	90	100.0	100.0	87.8	98.9	87.8
XSum	30	100.0	100.0	90.0	96.7	90.0
WritingPrompts	30	100.0	100.0	90.0	100.0	90.0
PubMedQA	30	100.0	100.0	83.3	100.0	83.3
Detector correct	54	100.0	100.0	85.2	98.1	85.2
Detector wrong	36	100.0	100.0	91.7	100.0	91.7

Table 13: Human evidence-audit results. Values are strict YES rates in percent from 90 evidence records independently labeled by two annotators and aggregated with conservative consensus labels. Arrows indicate the preferred direction. “All four” requires YES on understandability, concrete grounding, span relevance, and no hallucination. The permissive all-four rate, counting PARTIAL as support, is 88.9%.

explanation structure, span highlights, or any other single component.

Participants complete 18 main items and 3 practice items. The main items are a balanced set of human-written and AI-generated passages sampled from the same domains used in automatic evaluation: news summarization, creative writing, and biomedical question answering. The 18 main items contain six items per domain and a 9/9 human–AI label balance. They are stratified by detector correctness and confidence so that the study contains detector-correct, detector-wrong, high-confidence, and low-confidence cases. This stratification supports measurement of overtrust and appropriate deferral rather than only raw agreement with the detector.

E.1.3 Participants

The participant pool consists of university students and recent graduates who can read short English passages. Participants are not domain-matched experts; instead, the study records field, English-reading familiarity, AI-tool use, prior AI-assisted English writing or editing, and prior AI-generated-text detection experience as descriptive covariates for reporting sample composition.

The recruitment design specifies 72–90 valid participants, with 24–30 assigned to each interface condition. The completed-response pool contains 98 sessions. Quality filtering leaves 92 eligible sessions, and the balancing rule retains 30 participants per condition, yielding $N = 90$ retained participants. With 18 main items per participant, the analyzed sample yields 1,620 item-level decisions across the three interface conditions. Participants are asked to complete the study only once. Recruitment uses a mixed student population; the retained sample is AI-familiar, with AI-tool use summarized descriptively in Table 18.

Recruitment uses campus-style channels and di-

rect student networks with a neutral description of the task as an English text judgment study. The recruitment message states that participants read short English passages, view system-provided information, and judge whether each passage is human-written or AI-generated. The recruitment message does not describe the interface comparison or condition differences. Each participant receives RMB 20 as compensation for completing the study. We considered RMB 20 appropriate for the 18-item task and the local campus-recruitment context; payment was not contingent on participants’ responses.

Consent. Before the task begins, the interface presents a consent and task-readiness screen. Participants confirm that they understand the task and are ready to begin. Response data exclude direct identifiers, and participant records are anonymized before analysis.

Background questionnaire. Before the main task, participants can provide age eligibility, student status, major or field, self-rated English reading ability, frequency of ChatGPT or similar AI-tool use, prior use of AI tools for English writing or editing, prior AI-generated-text detection experience, and self-rated familiarity with AI-generated-text detection. Only age and English-reading eligibility are inclusion criteria; the remaining fields are used to describe the retained sample.

E.1.4 Task Procedure

Each session is assigned to one fixed interface condition. The study server allocates new sessions with a running balance rule: absent a condition-specific recruitment link, the next session is assigned to the currently least-filled condition; condition-specific links are used only to maintain recruitment balance. The assigned condition remains fixed throughout the session. Main-item order is randomized by a hash of participant code and condition while pre-

serving the same balanced stimulus set across conditions, reducing order effects without changing the study composition.

For each item, participants read the text and the assigned interface output, then choose one of three actions: AI-GENERATED, HUMAN-WRITTEN, or DEFER/UNCERTAIN. They also report confidence in their own decision and can optionally explain what evidence influenced them. After completing all items, participants answer a short post-task questionnaire.

Participant instructions. Before the main task, the interface displays the study title, a short description, three instructions, and a consent/readiness checkbox. The title is: “AI-Generated Text Detection Assistance Study.” The description states: “Participants judge passages sampled from the three paper datasets under one of three assistance conditions: Fast-DetectGPT score-only baseline, direct LLM explanation without guidelines, or AGENT-XTrust guideline-grounded evidence.” The displayed instructions are: (1) “You will be assigned to one assistance condition for the full session.” (2) “For each item, read the passage, inspect the system aid shown on the right, and make your own judgment.” (3) “The shown system may be wrong; the study measures your judgment under different assistance styles.” The consent/readiness text is: “I understand the task and I am ready to begin the study session.” Each item then presents the passage, the assigned system aid, the three decision options, participant-confidence input, and an optional free-text explanation field.

Manual evidence-audit instructions. For the separate evidence-quality audit, annotators were shown the original passage, the system document-level label and confidence, the selected guideline name and dimension, the local rationale, and the highlighted span or quoted evidence shown to users. They were instructed to judge each evidence record independently on four criteria: whether the rationale is understandable, whether it is grounded in concrete text, whether the highlighted span supports the cited guideline and rationale, and whether the rationale avoids unsupported claims. Annotators used YES, PARTIAL, and NO labels; detector correctness was provided only for aggregate error analysis and not as the target of the evidence-quality labels.

Quality filtering. Before outcome analysis, sessions are filtered using pre-specified validity criteria implemented by the study server. We exclude sessions that are incomplete, fail the embedded attention check, contain two or more main-item responses shorter than 2.5 seconds, show repeated single-label responding across the task, reuse a completed participant code, or show insufficient reading progress: four or more low-scroll responses, or at least two low-scroll responses together with an average main-item response time below 10 seconds. Low-scroll means that the text was not fully visible and recorded reading progress is below 85%. This filtering removes invalid or low-attention sessions while preserving the fixed 18-item stimulus set. After filtering, the eligible pool contains 92 sessions; for the main between-condition analysis, we retain a balanced sample of 30 eligible sessions per interface condition, yielding 90 analyzed participants and 1,620 item-level decisions. The two surplus eligible sessions are omitted only for condition balancing, not because of quality failure.

E.1.5 Outcome Metrics

Let y_i be the true label, d_i the detector label, and h_i the participant action. The human-study analysis reports four primary participant outcomes; LLM detector-confidence calibration and manual evidence quality are reported separately in Appendix A.1 and Appendix D.2.

Human accuracy. Human accuracy is $\Pr[h_i = y_i]$ over all main decisions, treating deferral as not correct for the binary label task.

False accusation rate. False accusation rate is $\Pr[h_i = \text{AI} \mid y_i = \text{HUMAN}]$.

Overtrust rate. Overtrust rate is $\Pr[h_i = d_i \mid d_i \neq y_i]$, measured on detector-wrong items.

Deferral rate. Deferral rate is $\Pr[h_i = \text{DEFER/UNCERTAIN}]$ over all main decisions.

E.1.6 Analyzed Sample

The analyzed sample contains 90 retained participants for the balanced main analysis, with 30 participants in score-only, 30 in score plus explanation, and 30 in full AGENT-XTrust evidence. Quality filtering and condition balancing produce 30 eligible sessions per condition. We report bootstrap confidence intervals from participant-level resam-

pling and participant-level deltas for the primary contrasts.

Breakdowns. Table 14 breaks the filtered decisions down by source domain. The useful signal is consistency: the accuracy gain appears in all three domains rather than being carried by one stimulus type. Table 15 separates cases where the displayed detector label is correct from cases where it is wrong; this breakdown shows that the primary overtrust reduction comes specifically from detector-wrong items, where evidence quality matters most. Table 16 reports participant-level deltas for the primary contrasts, and Table 17 reports item-level mixed-effects robustness checks with participant and item random intercepts.

Background covariates are summarized at the aggregate retained-sample level in Table 18. We use these fields to characterize the participant pool rather than to define the treatment contrast. The between-condition comparison is protected by the fixed stimulus set, equal retained condition sizes, random item ordering, and the item- and participant-intercept robustness checks in Table 17; no outcome claim is conditioned on a demographic subgroup.

Reproducibility. The analysis uses the fixed 18-item stimulus set with ground-truth labels and detector outputs, the three interface conditions, and the pre-specified completion and quality filters. We report aggregate decisions, timing summaries, quality-control outcomes, and condition-level results.

F.2 Human-Study Metric Figure

F Reproducibility Details

This appendix records model-size considerations, implementation settings, compute constraints, software parameters, and artifact-license notes.

F.1 Model Size

Recent studies suggest that smaller language models can outperform larger ones in zero-shot AI-generated text detection tasks (Miresghallah et al., 2024). In cross-model settings, larger detector models can assign likelihood and curvature less discriminatively across human text and generations from other models. We therefore follow the baseline settings and model selections from Fast-DetectGPT (Bao et al., 2024) for statistical baselines rather than changing their backbone models.

AGENT-XTrust, however, relies on linguistic reasoning, confidence elicitation, and semantic interpretation rather than only token-probability scoring. Its use of a stronger instruction-following model is intended to support guideline reasoning and evidence generation. Because the statistical baselines use different backbones, benchmark differences do not isolate model scale from detector architecture. The framework defines structured interfaces through which alternative router, expert, and MetaAgent models could be evaluated, with any substitution requiring renewed validation of accuracy, calibration, and evidence quality.

F.2 Implementation Details

In all reported Wiki-guideline evaluations, the router, guideline experts, and MetaAgent use deepseek/deepseek-v3.2 with greedy decoding (temperature 0). Agent calls request structured JSON outputs so that predictions, confidence values, rationales, and optional quoted spans can be parsed into evidence records. This reproducibility-oriented configuration minimizes designed sampling variability, but we did not conduct repeated-run stability experiments and do not claim bitwise-identical hosted API outputs. Base guideline calls are parallelized across activated experts, and the resulting intermediate outputs are retained for aggregation, audit, and interface display.

Adaptive routing limits average expert fan-out, but the richer evidence package still requires router, confidence-steering, expert, and MetaAgent inference. Parallel expert execution can shorten the serial critical path without reducing total API work, so evidence richness trades additional inference for inspectability rather than eliminating computational cost.

Compute budget. No model training or fine-tuning is performed in this work. The reported detector uses API inference for the router, guideline experts, confidence-steering calls, direct-judge baseline, and MetaAgent, followed by local CPU-based parsing, thresholding, statistical analysis, and figure generation. The API model is identified as deepseek/deepseek-v3.2; public DeepSeek-V3-family documentation describes the backbone as a mixture-of-experts model with 671B total parameters and 37B activated parameters per token. Because inference is served through a provider API, the exact server-side hardware allocation and GPU-hour accounting are not observable to us. We there-

Condition	XSum		WritingPrompts		PubMedQA	
	Acc. (↑)	Defer (↓)	Acc. (↑)	Defer (↓)	Acc. (↑)	Defer (↓)
Score only	55.0	15.0	55.6	21.1	52.2	21.1
Score + explanation	70.0	15.6	66.7	11.1	51.7	18.9
AGENT-XTrust	77.2	7.2	77.8	6.1	67.2	12.8

Table 14: Domain-level human-study breakdown. Values are percentages. Arrows indicate the preferred direction: higher accuracy is better and lower deferral is better. Bold marks the best condition within each domain-metric column. The table tests whether the decision-support effect is consistent across source domains rather than driven by a single stimulus type.

Condition	Displayed detector correct		Displayed detector wrong	
	Acc. (↑)	Defer (↓)	Acc. (↑)	Defer (↓)
Score only	61.3	19.6	18.9	16.7
Score + explanation	77.4	12.9	11.7	23.3
AGENT-XTrust	81.6	7.1	36.7	16.7

Table 15: Human decisions by whether the displayed detector label is correct. Values are percentages; arrows show preferred direction and bold marks the best condition in each column, with ties bolded. The detector-wrong columns isolate overtrust: AGENT-XTrust yields the highest participant accuracy when the displayed detector label is wrong.

fore report the reproducible settings available to users: model identifier, temperature-0 greedy decoding, structured-output schema, fixed prompts, validation-threshold protocol for score-based baselines, and no training compute.

Artifact constraints. The experiments use XSum, WritingPrompts, PubMedQA, Wikipedia’s “Signs of AI writing” guideline source, and published detector baselines. We cite the creators of these artifacts in the main paper and use them for research evaluation only. For release, we will distribute code, prompts, evaluation configuration, aggregate study materials, and derived outputs that can be shared without exposing participant identities. We will not redistribute private participant-level identifiers, and any benchmark text redistribution will follow the original artifact terms. The public dataset cards or repositories list XSum/GEM-XSum under CC BY-SA 4.0, WritingPrompts under MIT, and PubMedQA under MIT; downstream users should consult the original artifact pages for the authoritative license text and any source-specific restrictions.

Analysis parameters. The implementation is written in Python. The reproducibility package records lower-bound requirements for the main runtime and analysis dependencies: FastAPI 0.111+, Uvicorn 0.30+, Pydantic 2.7+, OpenAI Python SDK 1.30+, PyYAML 6.0+, json-repair 0.28+,

dirtyjson 1.0+, tqdm 4.66+, scikit-learn 1.4+, pandas 2.2+, Matplotlib 3.8+, and PyTorch 2.2+. The human-study analysis script uses Python standard-library routines for flattening responses, quality filtering, participant-level bootstrap confidence intervals, and permutation tests; the reported analysis uses 5,000 bootstrap samples, 5,000 permutations, and fixed random seed 20260515. The final automatic detector reports are regenerated from completed JSONL outputs with 5,000 bootstrap samples and fixed seed 13. Study item order is randomized deterministically from participant code and condition, and baseline thresholds are frozen before test evaluation as described in Appendix A.3.1.

Comparison	Metric direction	Delta (pp)	Permutation <i>p</i>
AGENT-XTrust – Score	Accuracy (↑)	19.8	< .001***
	False accusation (↓)	-5.6	.0226
	Overtrust (↓)	-17.8	.021*
	Deferral (↓)	-10.4	.003**
AGENT-XTrust – Score+expl.	Accuracy (↑)	11.3	.009**
	False accusation (↓)	-15.6	< .001***
	Overtrust (↓)	-18.3	.009**
	Deferral (↓)	-6.5	.020*
Score+expl. – Score	Accuracy (↑)	8.5	.027*
	False accusation (↓)	10.0	.022*
	Overtrust (↓)	0.6	0.948
	Deferral (↓)	-3.9	0.244

Table 16: Participant-level pairwise deltas. Arrows indicate the preferred direction of the raw metric: higher accuracy is better, while lower false accusation, overtrust, and deferral are better. Deltas are computed as the first condition minus the second condition named in the comparison. Positive accuracy values favor the first condition; negative false-accusation, overtrust, and deferral values favor the first condition. Bold deltas are significant at $p < .05$ and favor the first condition. Stars mark permutation-test evidence: * $p < .05$, ** $p < .01$, *** $p < .001$. Permutation tests are descriptive evidence for the study contrasts.

Outcome	Contrast	Log-odds	Odds ratio	Approx. <i>p</i>
Accuracy (↑)	AGENT-XTrust vs. Score	1.05	2.86	< .001***
	AGENT-XTrust vs. Score+expl.	0.65	1.91	< .001***
False accusation (↓)	AGENT-XTrust vs. Score	-0.40	0.67	.017*
	AGENT-XTrust vs. Score+expl.	-0.97	0.38	< .001***
Overtrust (↓)	AGENT-XTrust vs. Score	-0.64	0.53	.004**
	AGENT-XTrust vs. Score+expl.	-0.66	0.52	.003**
Deferral (↓)	AGENT-XTrust vs. Score	-1.02	0.36	< .001***
	AGENT-XTrust vs. Score+expl.	-0.77	0.46	< .001***

Table 17: Mixed-effects robustness checks for the human study. Arrows indicate the preferred direction of the raw metric: higher accuracy is better, while lower false accusation, overtrust, and deferral are better. Each block reports a logistic mixed-effects model with interface condition as a fixed effect and random intercepts for participant and item. Accuracy and deferral use all 1,620 item-level decisions; false accusation uses the 810 human-written item decisions; overtrust uses the 300 detector-wrong item decisions. Bold coefficients and odds ratios favor AGENT-XTrust in the preferred direction. Stars mark approximate fixed-effect evidence: * $p < .05$, ** $p < .01$, *** $p < .001$. Approximate p values are robustness checks rather than the primary inferential test.

Covariate	Value	Reported share (%)
English familiarity	High	18.2
	Medium	68.2
	Low	13.6
AI-tool use	Daily	74.4
	Weekly	20.9
	Monthly	4.7
Detection experience	Seen tools	48.8
	Course or research	32.6
	None	18.6
Research background	Student coursework	61.4
	Academic research	31.8
	Non-research	6.8

Table 18: Aggregate background covariates for the retained human-study sample. Percentages are computed among participants who reported the corresponding covariate, so denominators vary by row group. These variables describe sample composition and are not used as outcome-defining subgroups.

Detection History

Guest mode can analyze text, but detection history is only saved for signed-in users.

Search saved runs

Guest detections work immediately. Sign in if you want history and saved records.

History is unavailable in guest mode.

Run detections freely now, then sign in with Google when you want saved records and quick reload.

AUTHORING SURFACE

Input Workspace

299 chars 42 words

Paste a paragraph, article section, or generated response to inspect.

Recent benchmark results indicate substantial gains in downstream task accuracy, although the sample pool was restricted to a narrow demographic and may not generalize to broader real-world settings. The authors acknowledge this limitation while still framing the method as promising for deployment.

☒ Enable router

☒ Use meta agent

☒ LLM precise spans

☒ Enable confidence calibration

Calibration method

SteerConf (steer)

META AGENT SETTINGS

These only apply when the meta agent is enabled.

☒ Meta confidence output

☐ Include guidelines in meta prompt

Meta decision mode

Confidence weighted

Meta summary mode

By guideline

Analyze

Load Demo Text

Clear

Completed 100%



FINISHED

The dinosaur crossed the line and packaged the final judgment for review.

Elapsed: complete

ETA: complete

DECISION SNAPSHOT

Overall Conclusion

human leaning result

44% of the submitted text is currently flagged by highlighted spans, with 2 highlighted spans available for review.

Decision

human

Flagged Text

44%

Highlighted Spans

2

Decision=human | Estimated AI text=44% | Highlighted spans=2

The text is a brief, neutral academic statement reporting benchmark results and acknowledging limitations, without clear AI markers from the guidelines. Two base agents flagged it as AI with low confidence (0.3771 and 0.6132), citing generic phrasing and vague attributions, but their reasoning notes the text lacks specific flagged patterns and could be human-written. One agent identified it as human with higher confidence (0.6418), emphasizing the absence of undue significance emphasis and alignment with technical human writing. Given the higher confidence of the human decision and the weaker, less specific AI indicators, the evidence leans toward human authorship, though the mixed signals and formulaic tone warrant moderate confidence.

EVIDENCE REVIEW

Highlighted Suspicious Spans

Evidence Intensity: Lighter cue Medium cue Stronger cue Color reflects how strongly the system flagged the span.

Recent benchmark results indicate substantial gains in downstream task accuracy, although the sample pool was restricted to a narrow demographic and may not generalize to broader real-world settings. The authors acknowledge this limitation while still framing the method as promising for deployment.

Span Detail #1

Highlighted length: 33 chars | Cue strength: Medium (0.40 - 0.69) | Source: llm

WC6 · Vague attributions and overgeneralization of opinions

Content | Source: llm

The text uses vague attributions like 'Recent benchmark results indicate' and 'The authors acknowledge this limitation while still framing the method as promising for deployment,' which resembles weasel wording by attributing claims to unsp...

Figure 4: AGENT-XTrust detector interface. The system exposes the document-level judgment together with highlighted spans, estimated flagged-text coverage, and guideline-level evidence, so the user can inspect why specific passages are treated as suspicious alongside the binary label and scalar score.

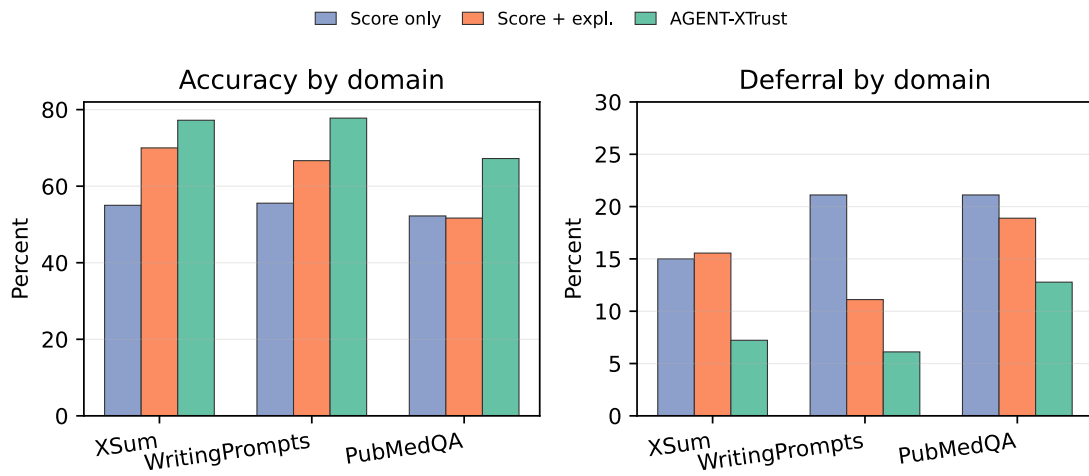


Figure 5: Domain-level accuracy and deferral in the human-study sample. The same 18-item stimulus set is used across conditions, with six items from each source domain.

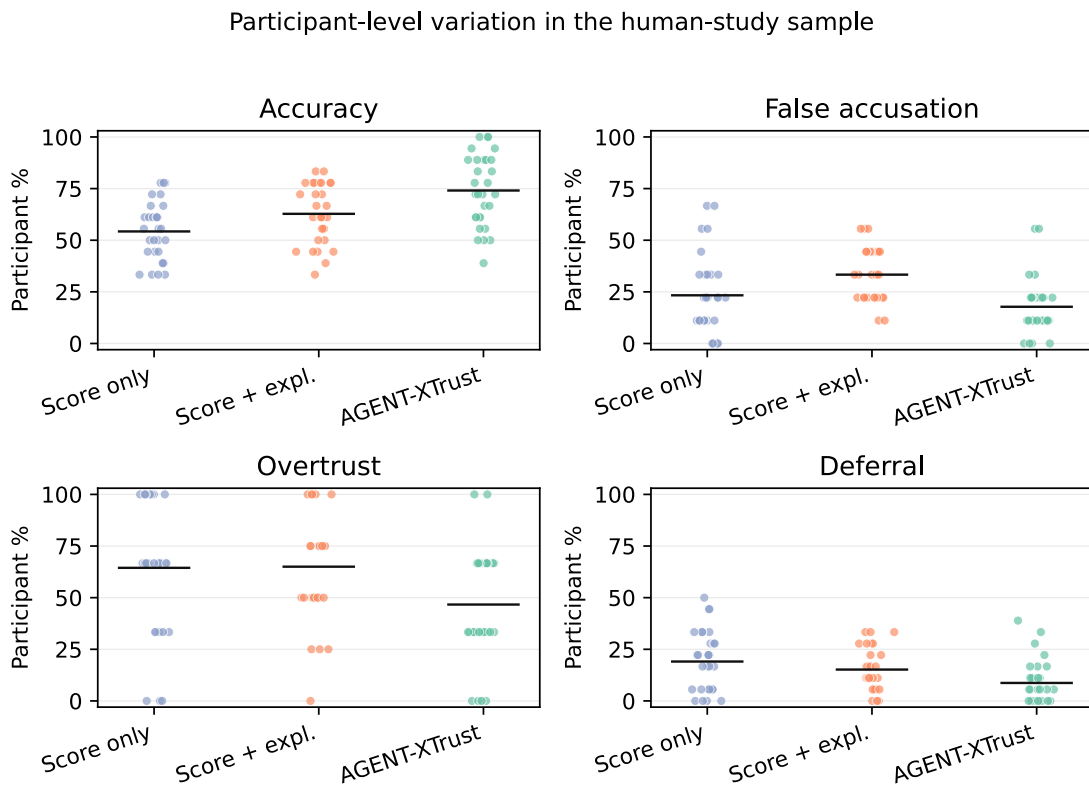


Figure 6: Participant-level distributions for the four primary human-study metrics. Each dot is one retained participant; horizontal bars show condition means. These plots show the distribution underlying the reported condition-level effects.

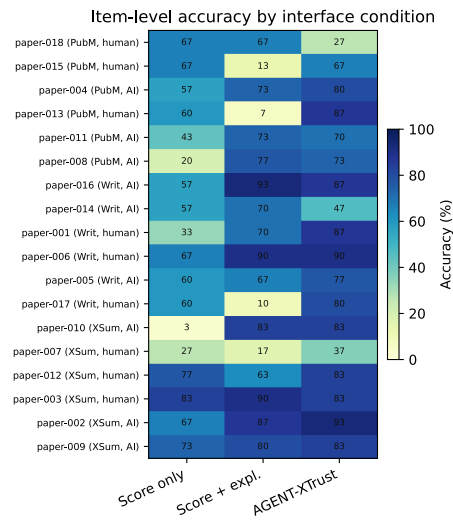


Figure 7: Item-level accuracy by interface condition. The heatmap summarizes item-level variation underlying the condition differences.

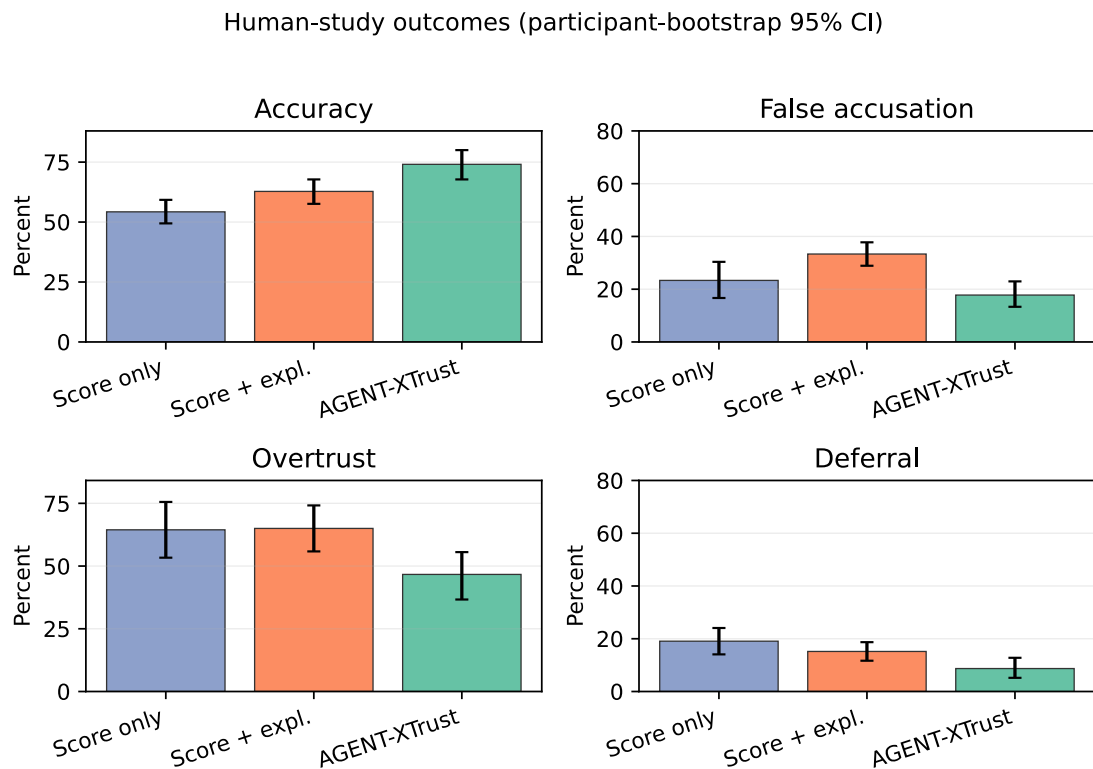


Figure 8: Human-study outcomes across the four primary metrics. Error bars show participant-bootstrap 95% confidence intervals.