

Benchmarking the Accuracy of Crickit's Fact-Checking Engine

v0.1.0



Crickit

www.crickit.ai

info@crickit.ai

BACKGROUND

With trust in traditional news media at historic lows and the spread of online misinformation and disinformation now viewed as a significant global threat, tools that can automatically verify factual statements — and do so in a transparent, reproducible way — are increasingly important. Using a test set with known answers, or a “benchmark,” provides a standardized way to quantify and compare the performance of different systems, advancing the science of automatic fact-checking. AVeriTeC (Automated Verification of Textual Claims) is an EU-funded research benchmark built from real-world claims checked by professional fact-checking organizations, and is directly relevant to Crickit, a patent-pending technology publicly released in November 2025 that provides real-time fact-checking overlays for social-media videos, starting with YouTube on desktop browsers.

This white paper is authored by Crickit, LLC and has not undergone external peer review. To help journalists and researchers evaluate it as a company-produced white paper, we follow widely recommended practices for assessing non-peer-reviewed research: we clearly describe our dataset and methods, report standard evaluation metrics, disclose limitations and conflicts of interest, and provide access to underlying analysis files so that others can independently scrutinize and, where appropriate, replicate our findings.

Claims from the AVeriTeC development dataset were used to evaluate the performance of Crickit’s November 2025 fact-checking system (v0.1.0). In AVeriTeC, each claim can be labeled “Supported”, “Refuted”, “Not Enough Evidence”, or “Conflicting Evidence/Cherry-picking.” The development subset used here contains 500 claims drawn from real-world fact checks by 50 different organizations. For the integrity of the benchmark, the full set of claims and labels is not reproduced here; instead, we provide two illustrative examples of AVeriTeC-style claims that the organizers themselves treat as public exemplars.

- **Claim 1:** "Hunter Biden had no experience in Ukraine or in the energy sector when he joined the board of Burisma."
- **Label:** Supported
- **Context:** Attributed in the dataset to Pam Bondi, also from a 2020 RNC speech, again linked to a PolitiFact fact-check.

- **Claim 2:** "Donald Trump delivered the largest tax cuts in American history."
- **Label:** Refuted
- **Context:** Attributed in the dataset to Eric Trump, in a speech at the 2020 Republican National Convention, fact-checked via PolitiFact.

EVALUATION METHODS

Of the 500 AVeriTeC development claims, 427 are labeled either Supported or Refuted. Our evaluation focuses on this binary subset. Restricting analysis to Supported/Refuted enables a transparent, replicable veracity metric and aligns with common practice in recent shared-task work: performance is substantially lower and less stable on the more ambiguous “Not Enough Evidence” and “Conflicting Evidence/Cherry-picking” categories, which are minority classes and exhibit greater annotation variability. Several top AVeriTeC/FEVER-2024 systems also report two-class ablations as a reasonable simplification for headline accuracy comparisons.

Using the claim text plus associated metadata (date, speaker, reporting source, and location), we configured a variant of Crickit’s November 2025 fact-checking system to assign a binary label (Supported or Refuted) to each of the 427 claims. No temporal cut-off was imposed on web search: evidence could come from sources published after the original claim, mirroring how Crickit operates in live use on YouTube, where it always searches the current web rather than constraining itself to time-matched evidence. This differs from the official AVeriTeC shared-task evaluation, and therefore our results are not directly comparable to scores on the AVeriTeC leaderboard.

Crickit’s fact-checking system was then used to evaluate all 427 claims four times, with each “run” representing a fresh end-to-end execution of the system over the full claim set. The claims in this collection were made between August 26 and October 31, 2020. Since then, additional evidence has appeared, and in some cases historical reporting has been clarified, so the most appropriate label for a given claim may have shifted. As with any human-labeled dataset, the original AVeriTeC labels can contain occasional errors or debatable judgments; for example, we identified at least one claim (about Rahul Gandhi) whose official label appears to have been incorrect even at the time.

To address possible labeling errors and the expected “half-life” of facts about fast-moving news events, we identified every claim where Crickit’s label

disagreed with the original AVeriTeC label in any run, which occurred on average in approximately 7.6 percent of the claims. Each discrepant claim in each run was then independently reevaluated using OpenAI’s long-inference, high-reasoning GPT-5 Thinking model with web-search grounding as a judge LLM. For every such case, the judge LLM produced (a) an Agree/Disagree verdict with respect to the original Crickit label, (b) an explanation of its reasoning for arriving at its conclusion, and (c) the top three supporting URLs, which were then reviewed by human annotators.

In the great majority of cases, independent runs of GPT-5 Thinking reached the same verdict for the same claim across different runs, and human reviewers agreed with those unanimous verdicts. However, for 11 of the 427 claims, this first judge LLM did not reach a consistent verdict across runs. For these 11 harder cases, we escalated to OpenAI’s Deep Research tool, again using GPT-5 Thinking with web-search grounding, as a final arbiter. Our proposed updates and corrections for the affected AVeriTeC dev-set labels are summarized in the accompanying discrepancy analysis files listed below[1]: [1] To use these claims, download the official AVeriTeC dataset from fever.ai or the authors’ GitHub and join on the provided IDs. AVeriTeC © is the original authors and licensed under CC-BY-NC-4.0. Please do not post unencrypted claims and labels from the AVeriTeC dataset publicly — these data are distributed under a CC-BY-NC-4.0 license and must be properly secured and attributed¹:

- [Run 1 Discrepancy Analysis IDs.xlsx.zip](#)
- [Run 2 Discrepancy Analysis IDs.xlsx.zip](#)
- [Run 3 Discrepancy Analysis IDs.xlsx.zip](#)
- [Run 4 Discrepancy Analysis IDs.xlsx.zip](#)

Each discrepancy-analysis file includes a “Longer review of Crickit label” column with values “AGREE” or “DISAGREE,” indicating whether the judge LLM agreed with Crickit’s original Supported/Refuted label. Most entries appear in regular font, meaning GPT-5 Thinking produced a consistent final verdict across every run where that claim received a discrepant label, and human review concurred. A small minority of entries (11 claims total) are

1. To use these claims, download the official AVeriTeC dataset from fever.ai or the authors’ GitHub and join on the provided IDs. AVeriTeC © is the original authors and licensed under CC-BY-NC-4.0. Please do not post unencrypted claims and labels from the AVeriTeC dataset publicly — these data are distributed under a CC-BY-NC-4.0 license and must be properly secured and attributed.

shown in bold font, marking cases where GPT-5 Thinking's verdict differed between runs and Deep Research was therefore consulted as a tie-breaker. In each of these 11 cases, Deep Research was instructed to consult at least 50 sources before issuing a final verdict. Cells shown in bold indicate where Deep Research made the final judgement, and yellow highlighting indicates when Deep Research overturned the GPT-5 Thinking verdict.

Each file also includes a written explanation for the judge LLM's AGREE/DISAGREE decision with respect to the original November 2025 Crickit label, together with the URLs of the three sources it considered most probative.

RESULTS

Each run was evaluated on four standard classification metrics: accuracy, macro-averaged precision, macro-averaged recall (equivalent to balanced accuracy in the binary case), and macro-averaged F1-score. Table 1 reports results against the original, uncorrected AVeriTeC labels (that is, without applying any of the proposed label updates derived from the discrepancy analyses).

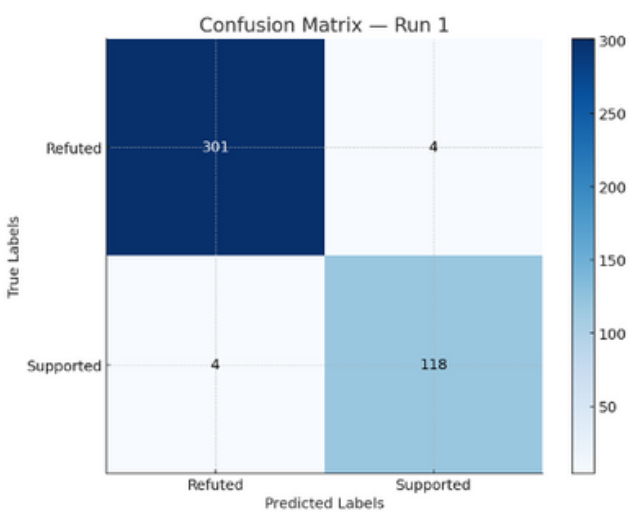
TABLE 1: RESULTS AGAINST UNCORRECTED LABELS

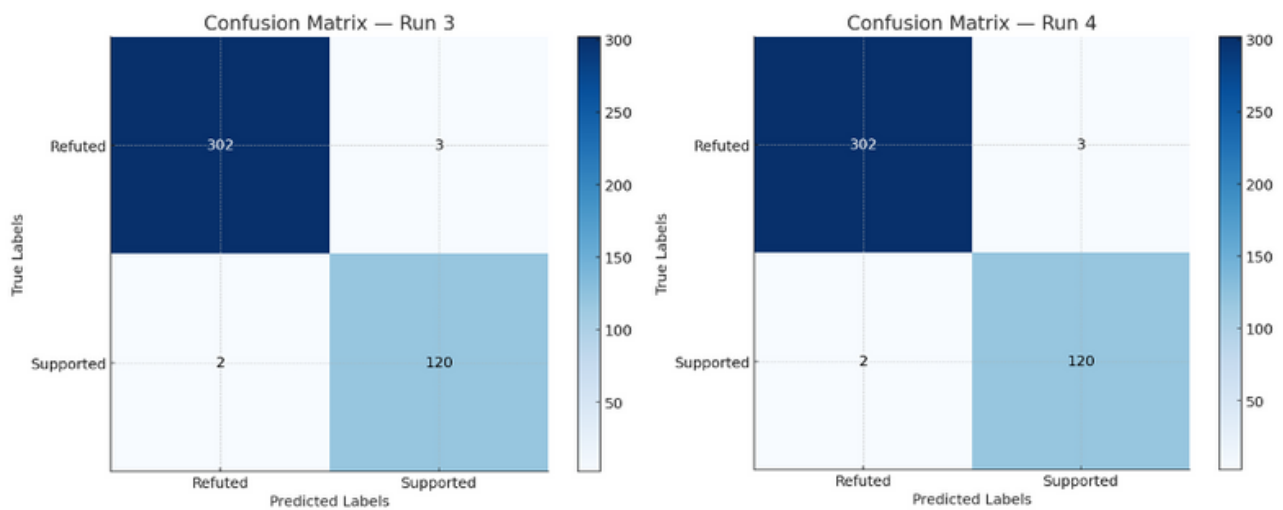
	run1	run2	run3	run4	avg
accuracy	0.918033	0.929742	0.920375	0.927400	0.923888
macro precision	0.897510	0.910945	0.905696	0.911873	0.906506
macro recall	0.903279	0.918852	0.897541	0.909836	0.907377
macro f1-score	0.900319	0.914764	0.901479	0.910846	0.906852

When we evaluated against the updated and corrected labels — reflecting the judgments of GPT-5 Thinking, Deep Research where needed, and human reviewers — the results are as follows:

TABLE 2: RESULTS AGAINST UPDATED AND CORRECTED LABELS

	run1	run2	run3	run4	avg
accuracy	0.981265	0.990632	0.988290	0.988290	0.987119
macro precision	0.977049	0.986253	0.984515	0.984515	0.983083
macro recall	0.977049	0.990984	0.986885	0.986885	0.985451
macro f1-score	0.977049	0.988580	0.985691	0.985691	0.984253





The resulting average **accuracy of 98.7%** (0.9871) is therefore best understood as a product-track estimate on an adapted AVeriTeC Supported/Refuted task: it reflects evaluation without recency constraints on retrieval and with updated labels that correct for apparent annotation errors and factual drift since 2020. This 98.7% figure is not an official AVeriTeC leaderboard result and should not be compared directly to AVeriTeC scores that jointly evaluate verdict and contemporaneous evidence. Nevertheless, because AVeriTeC claims concern real-world political and current-events content from dozens of fact-checking organizations, at a level of difficulty similar to many news and public-affairs videos, we view these results as informative about Crickit’s ability to label real-world claims that are salient to social-media consumers, as well as educators, journalists, policymakers, and fact-checking groups.

COMPARISON WITH CRICKIT'S CURRENT SOCIAL-MEDIA APPLICATION

It is also important to distinguish this benchmark scenario from Crickit's behavior on live social-media content. In the AVeriTeC setting, rich context (date, speaker, reporting source, and location) is provided explicitly with each claim. By contrast, when Crickit extracts claims from YouTube videos, some contextual information may be incomplete or harder to infer. As a result, the November 2025 version of the Crickit application may exhibit a somewhat higher real-world error rate than the benchmark numbers above, especially for "context-matching" errors where the extracted claim does not perfectly match the speaker, time, or situation in the underlying video. These contextual mismatches are typically visible to users and conceptually distinct from errors in retrieval or reasoning, which we expect to account for a similar share of the overall error rate as in this AVeriTeC study.

To address this contextual risk, Crickit already implements one key safeguard that was not present during this AVeriTeC evaluation: in the live application, the system can issue an internal Not Enough Evidence verdict. When that happens, Crickit does not surface a fact-check to the user. This follows a core principle of evidence-based reporting and science: absence of evidence should not be presented as evidence of absence.

A second mitigation strategy is planned for the near future: giving users an explicit way to request a slower "double-check" for any claim they believe may be in error and that has not already been escalated. This mode will trade speed for thoroughness by reviewing more sources, allocating more reasoning steps, and incorporating additional contextual detail. We expect such automated double-checks to further narrow any remaining performance gap between live social-media use and benchmark results. Importantly, this workflow is designed to require minimal human effort, no minimum volume of users, and to be both scalable and resistant to tampering.

REAL-WORLD CHOICES FOR SOCIAL MEDIA CONSUMERS

To understand the practical significance of these results, it is useful to compare Crickit not to an idealized standard, but to the real-world status quo. YouTube has more than 250 million monthly users in the United States and over 2.5 billion globally. For those viewers, the central question is: what difference does using Crickit make to their exposure to misleading or false claims?

A full answer requires large-scale user studies of perception and behavior, which are outside the scope of this white paper but are currently being planned in partnership with two universities. In the meantime, we can perform a simple, transparent calculation on a small but concrete sample of widely viewed news content, to estimate how much Crickit might reduce exposure to problematic claims on YouTube.

Crickit uses six internal factuality labels for live claims, five of which are shown to users ("Not Enough Evidence" is suppressed). The visible labels are Correct, Partly Right, Misrepresented, Mostly Wrong, and False. For the following analysis, we group

Misrepresented, Mostly Wrong, and False together as "problematic" claims. We treat Partly Right as acceptable, even though such claims often include minor errors or omissions, because this convention allows us to tie the analysis directly to the accuracy estimates and confusion matrices derived from the AVeriTeC evaluation.

We then drew a small, clearly specified sample of YouTube content as follows. Using an incognito browser session (not signed in to YouTube, to avoid personalization) on November 7, 2025 we searched YouTube for "US news," filtered for videos uploaded "this month," set duration to 4-20 minutes, sorted by view count, and manually excluded obvious comedy content (e.g., The Daily Show). We analyzed the top 10 remaining videos, which included content from channels such as Forbes, MSNBC, CNN, Fox News, and Sky News Australia. Crickit automatically extracted 125 distinct claims from these videos, which we treated as the universe for this illustrative exercise. Of those 125 claims, 56 were labeled Correct, 29 Partly Right, 20 Misrepresented, 2 Mostly Wrong, and

IMPLICATIONS FOR CONSUMERS

18 False. In other words, 40 claims — 32.0% of the total — were clearly problematic (Misrepresented, Mostly Wrong, or False). Some of these problematic claims were made by guests or quoted figures rather than anchors, but in practice such misleading or false statements are rarely corrected on-screen, and viewers often encounter them without clear counter-evidence.

To estimate the potential impact of Crickit, we apply the November 2025 engine’s benchmarked performance on the adapted AVeriTeC Supported/Refuted task to this 125-claim sample.

Under these assumptions, nearly all unacceptable claims are correctly flagged, and Crickit introduces relatively few new errors. The expected share of problematic claims experienced by viewers falls from about 32.0% (roughly one in three) to around 1.0% (about one in a hundred) — a reduction of approximately 32-fold relative to the status quo without Crickit. **In practical terms, the choice for consumers is not between perfection and imperfection, but between a baseline where roughly one in three claims is problematic and a scenario where that rate is closer to one in a hundred.**

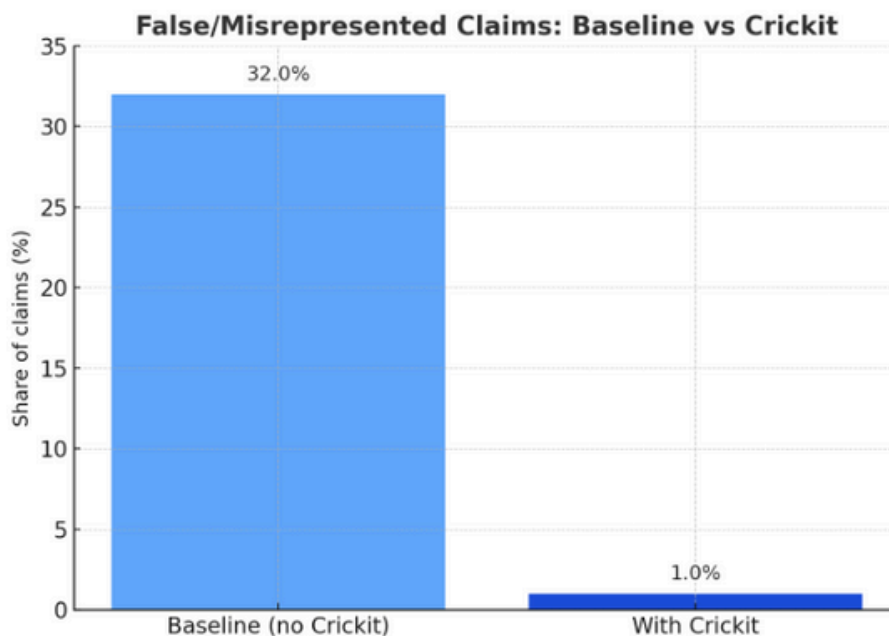


Figure 1: “False/Misrepresented Claims: Baseline vs Crickit.” For brevity, the label “False/Misrepresented” in the chart denotes the full unacceptable bucket used in the analysis (Misrepresented, Mostly Wrong, and False).

Several methodological cautions are important for interpreting this example. First, as noted above, the November 2025 YouTube version of Crickit may have a somewhat higher error rate than the adapted AVeriTeC evaluation, due to transparent contextual mismatches. We believe these effects are modest in proportion and can be further reduced in the near future via the mitigation strategies already outlined. Second, this calculation does not directly measure changes in understanding or behavior; rather, it quantifies how often viewers encounter problematic claims with and without Crickit in a specific, clearly defined slice of high-traffic news content.

Third, this YouTube exercise is a limited observational transfer check, not a population-representative estimate. Its primary role is to illustrate how a high-accuracy, search-grounded fact-checking engine changes what viewers see in specific, highly watched videos. Larger, controlled studies are still needed, and similar analyses should be extended beyond news into other high-impact domains such as public health, wellness, personal finance, and self-help, among others.

Even with its small size, this sample of

top-viewed US news videos is concerning: finding that roughly a third of claims fall into the Misrepresented/Mostly Wrong/False bucket suggests a high baseline exposure of viewers to misleading or false information. At scale, such exposure can pose risks to individuals and to their societies.

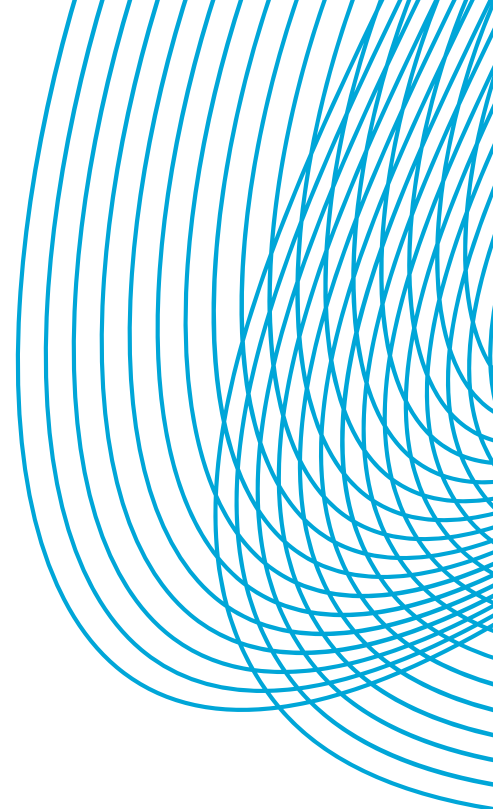
To illustrate these risks, even in just one domain — personal finance — 39% of Americans report losing \$250 or more, and 18% report losing over \$1,000 because of bad online advice; nearly three in five say they regret financial decisions based on inaccurate or misleading online information ([CFP Board](#)). In a media environment shaped by increasingly persuasive technologies and uneven levels of media literacy, tools that help people distinguish more reliable from less reliable claims can support better-informed decisions, protect individuals' finances, health, and aspirations, and help shield society from organized disinformation campaigns by adversaries opposed to our national interests.

We encourage independent researchers, journalists, and fact-checking organizations to build on and, where appropriate, critique the work presented

here. Our aim is that technologies like Crickit, when designed and evaluated transparently and used alongside professional journalism and human judgment, will help strengthen civic life, public health, mental health, and financial well-being. In the long run, genuine collaboration and human progress depend on a minimal shared reality grounded in evidence and science; by making it easier to see what is supported by facts and what is not, we hope to play a small but concrete part in protecting that common ground.

Authorship, conflicts of interest, and availability of materials: This evaluation was designed and conducted by Crickit, LLC, which develops and operates the Crickit fact-checking application. Crickit has a direct commercial interest in the product evaluated here. No external funder influenced the design, analysis, or reporting of these results. To support independent assessment by journalists and researchers, we provide: (a) a clear description of the dataset and evaluation protocol, (b) full run-level metrics and macro averages, and (c) the four discrepancy-analysis spreadsheets documenting every proposed label update. We expect that further scrutiny and replication attempts will refine and improve on the findings reported in this company-authored white paper.

GET IN TOUCH



CRICKIT, LLC



www.crickit.ai
info@cricket.ai