

2026 NIST GenAI Text Challenge FAQs

NIST GenAI Team

October 7, 2026

Contents

DISCLAIMER	2
1 FAQ	3
1.1 Task Definition & Metrics	3
1.2 The Human-Belief Measurement (Panel)	3
1.3 Presentation & Form of the Narratives	5
1.4 Domains, Framing & Language	6
1.5 Development & Test Data	6
1.6 Generator Track	7
1.7 Prompter Track	8
1.8 Container, Determinism & Runtime	8
1.9 Intellectual Property, Retention & Licensing	9
1.10 Logistics, Timeline & Eligibility	10

DISCLAIMER

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government. This project has been reviewed by the Research Protections Office (RPO) under the reference number ITL-2023-0644.

1 FAQ

We had received several questions from participants regarding the technical and logistical details of the challenge. While we have tried to include as much detail as possible in the evaluation plan, there are several specific questions that were beyond the scope of the evaluation plan. Such questions are categorized below, and we have tried to address them to the best of our ability.

1.1 Task Definition & Metrics

Q1.1. How will the discriminator score believability: (a) agreement/error against the human-annotated believability, (b) assigning LOWER believability to misleading content, or (c) both? These can diverge for persuasive-but-false narratives; which governs ranking?

Answer: As per the evaluation plan, the discriminator predicts the mean (μ) and standard deviation (σ) of the human-judged believability scores for each narrative, and the system's performance is measured by the error (Mean Absolute Error) between the predicted and observed μ and σ .

Q1.2. Will believability be graded with a proper scoring rule (Brier / log-loss / CRPS) or a correlation/error metric, and is calibration to the human belief distribution rewarded, or only point accuracy?

Answer: Believability is graded with an error metric, not a proper scoring rule: Mean Absolute Error (MAE) on both the predicted mean (μ) and standard deviation (σ) of the human belief distribution. Because σ is scored in addition to μ , calibration to the distribution of human scores is rewarded, not just point accuracy. On the other hand, AI-likelihood score is evaluated using AUC-ROC and Brier Score.

Q1.3. Is there a single combined ranking score per track, or a separate leaderboard per metric? If combined, what is the weighting?

Answer: NIST does not plan to provide a leaderboard or ranking of any kind for the submissions across different metrics. The goal is to provide a pareto front chart for the AI-likelihood task across the three different metrics that we have discussed earlier (Brier_T, Brier_N, AUC-ROC)— please see the details at [link to "Metrics" document, will be activated soon]. For the believability task, we plan to provide a detailed analysis of how each submission performed across different groups of domains and narratives via scoreboard. We believe that a simple mentioning of MAE will not be sufficient for capturing the nuances of believability detection.

Q1.4. For the AI-likelihood score, is there a target operating point (e.g., TPR at a fixed FPR) in addition to AUC/Brier?

Answer: As mentioned previously, NIST does not want to provide any target operating threshold for metrics like FPR and FNR. This is because these thresholds are usually application dependent and should ideally be chosen by the system operator/deployer of the discriminative system.

1.2 The Human-Belief Measurement (Panel)

Q2.1. Who are the believability annotators — general-public sample or experts — what is their approximate demographic/regional composition, and how many rate each narrative?

Answer: Currently, our evaluations rely on a cohort of approximately ten United States-based annotators per narrative, all of whom possess a higher level of scientific and technical expertise than the general population. To broaden the scope and diversity of our feedback, we plan to open public evaluation via our website in the near future, allowing us to collect human-judged scores from a broader demographic.

Q2.2. How is the per-narrative believability aggregated: mean of per-annotator ratings, median, or consensus?

Answer: The evaluation characterizes each narrative’s believability by the distribution of individual human ratings, summarized by its mean and standard deviation. The discriminator is asked to predict this mean and standard deviation, which is expected to model the scoring behavior of an annotator population more closely than a simple mean, median, or consensus.

Q2.3. What context do annotators see when rating believability — the narrative alone, or attributed to a source (author, outlet, platform, public figure)? Are they told it may be AI-generated?

Answer: The narratives are presented to the annotators in their raw form. The AI-generated narratives are shown as they were generated by the LLMs. The references and citations are included as part of the narrative. The participants of the survey are given the context of the study, which is to understand the believability of AI-generated content. Along with the believability scores, the annotators are also asked how confident they are that the narrative is AI-generated.

Q2.4. Will you report inter-annotator agreement (e.g., Krippendorff’s alpha), and will the dev-set labels include the DISTRIBUTION of annotator ratings (or their variance), not only the mean?

Answer: Yes. We will be providing the full breakdown of individual annotator scores for each narrative in the development set, rather than restricting the labels to summary metrics like mean and standard deviation. Participating teams can easily compute the *mean*, *standard deviation*, or alternative aggregate statistics from these raw scores. Furthermore, access to individual ratings will allow participants to perform their own fine-grained analysis, including calculating custom inter-annotator agreement (IAA) metrics.

Regarding IAA, our primary goal in collecting multiple human annotations is to capture the underlying distribution and inherent subjectivity of the evaluations—not necessarily to enforce high agreement. For certain nuanced narratives, low agreement or high variance is expected and reflects genuine differences in human interpretation. We will share overall benchmark statistics, including metrics like Krippendorff’s alpha, to characterize these distributional properties across the dataset.

Q2.5. How is the “main message” that annotators judge determined for each narrative?

Answer: The *main message* is basically the topic statement around which the narrative has been built. In the case of the human study, the topic statements are included as the title of the narrative, so that the annotators get an idea of the main messaging of the narrative even before they begin reading it.

Q2.6. Is belief measured at first reading, or after time/reflection?

Answer: This is not something that we are currently tracking in the human study. However, our

study is not explicitly designed to provide annotators with time to reflect on their level of belief in the narratives they read.

Q2.7. Do you intend believability to model the panel AS-IS (reflecting its actual cultural/demographic makeup) or an idealized “debiased general reader”? Relatedly, were narratives balanced across register/dialect, or is register/dialect a confounded variable in the data?

Answer: Our goal with believability is to model a generalized reader population. For more details on the annotator demographics, please refer to Q2.1.

1.3 Presentation & Form of the Narratives

Q3.1. Are narratives plain UTF-8 text, or may they contain formatting/markup (Markdown, HTML, headings, bold, lists, tables)? If markup is present, is it rendered for the human panel or shown literally?

Answer: Narratives are carried as a single JSON string field (`narrative`) in the system output artifacts (Appendix C), in the form of plain text content (not just UTF-8). The annotators are shown this plain text narrative for scoring.

Q3.2. When narratives are shown to the believability panel, how are they rendered — as plain unstyled text, or with fonts, sizes, and layout (e.g., styled like an article or a rendered document/image)? Is the rendering identical for every narrative, or can a generator influence presentation?

Answer: The rendering style was identical for all narratives in the human study. There are no styled text elements like bold and italic. This was done to create consistency across all the narratives and restrict the influence on believability just to the textual content.

Q3.3. Is there a length constraint on generated narratives — a word / character / token limit, or a scored length range — and is length itself a metric factor?

Answer: Currently all the narratives in the human study for believability are restricted to a length of 250 words. So narrative length is not a metric factor in itself.

Q3.4. For the discriminator input, is each narrative delivered as a single plain-text field, or may it contain structural/markup elements we should parse?

Answer: As stated in Q3.1, all the narratives are plain text. Hence, there will not be any parsing required before they are fed to the discriminator systems.

Q3.5. Are typographic/visual elements (emphasis, emoji, unicode styling, whitespace layout) permitted in narratives and preserved through to the human panel?

Answer: As mentioned in Q3.1, the narratives are in plain text format, so there will be no non-textual or stylistic elements in them.

1.4 Domains, Framing & Language

Q4.1. Which domains/topics will the narratives span (e.g., health, security, legal, finance, politics, science), and is any weighted more heavily?

Answer: We spanned 11 categories: health, nutrition, safety, science, technology, education, psychology, finance, legal, environment, and history. The first round of topics will not be balanced but we are aiming for a more balanced set in the next rounds.

Q4.2. Is the corpus English-only or multilingual? If multilingual, which languages, and is the annotator panel matched to each language/culture? Will narratives include dialectal or non-standard register (and is the panel matched to it)?

Answer: For now, the corpus is English-only.

Q4.3. In the study, how does the reader encounter each narrative — news article, social-media post, quoted expert, institutional statement — and who is the apparent speaker?

Answer: This is prompt-dependent and will differ across prompts. In the dev set, the narratives are a 200-250 word articles with quoted experts or institutions written in either an objective/neutral style or a more engaging style.

Q4.4. Are narratives standalone, or embedded in a surrounding context/thread?

Answer: Narratives in dev set are standalone.

1.5 Development & Test Data

Q5.1. When will the development set be released, and precisely what will it contain — topic statements, narratives, AI-likelihood labels, believability scores, and (per 2.4) the annotator distribution?

Answer: The development set will include the topic sentence, a statement id unique to the topic sentence, category of the topic, the narrative, a narrative id unique to the narrative, user id, the AI-likelihood perceived by the user and the believability score assigned by the user.

Q5.2. Approximate sizes: number of topic statements, narratives per statement, and dev-vs-test split.

Answer: Round1: Total of 50 topics (100 narratives) split into 20 topics (40 narratives) in dev set and 30 topics (60 narratives) in test set.

Q5.3. The rules state NIST data may not be used to train/tune. Could you confirm: (i) the dev set IS usable for calibration, threshold-setting, and validation (as distinct from training); and (ii) participants MAY train on their own external public datasets and public open-weight models, provided all training completes before submission?

Answer: Both are consistent with the plan. (i) The development set is explicitly provided “to support system design, calibration, and internal testing”, which covers calibration, threshold-setting, and validation. (ii) As per the Protocol and Rules, participants may use publicly available data to train

their systems, and all systems must complete training, model selection, and tuning *prior* to submission.

Q5.4. Which generator models populate the dev/test narratives (family or specific models), or is that intentionally blind to participants?

Answer: We used claude, gpt, and gemini models.

1.6 Generator Track

Q6.1. Confirm the scoring intent: a stronger generator drives the discriminator’s AUC toward 0.5 with Brier near 0.25 and HIGH believability, and over-deception (AUC well below 0.5) is treated as an artifact, not a win.

Answer: The scoring intent for generator systems is to maximize the discriminator Brier_T score, i.e., drive it to 1. As far as the generators are concerned, AUC is not a relevant metric because it requires both fake and real samples, and the goal of the generator is to try and convince the detectors that the fakes are real in absolute terms, not relative to real samples. From that perspective, over-deception is not a concern, as it tries to compare fake confidence scores w.r.t. real.

Q6.2. Must the generator produce BOTH a factual (truthfulness 0) and a misleading (truthfulness 1) narrative per statement, and is believability rewarded for both or only the factual ones?

Answer: It is not necessary that the generators have to produce both factual and misleading narratives for each topic statement. That type of narrative generated will depend on the input JSON that the generator track participants receive from the prompters, as mentioned in the evaluation plan (Appendix A).

Q6.3. May a generator BAKE IN its own US-origin open-weight model and run it offline in-container, or must generation use NIST’s pre-deployed models via the API?

Answer: Both modes are permitted. Appendix B.5 provides for fully offline operation with all model weights packaged inside the Docker image, and Appendix B.3.1 provides conditional network access to NIST-provided model endpoints (via injected NIST_API_ENDPOINT / NIST_API_KEY) for systems that use them. In both cases the system-level-contribution requirement (Appendix B.6) and the US-origin model restriction (Appendix B.7) apply.

Q6.4. What qualifies as the required “distinct, system-level contribution” (fine-tuning, orchestration, agentic workflow), and how/when is eligibility reviewed so we avoid a late rejection?

Answer: The reason behind requiring certain baseline system-level contribution is because we want to avoid submissions where the entire generation/detection is a simple off-the-shelf model call (API/open-source). Such SOTA models have already been used by us for your baseline systems and any novel contribution cannot be credited to the participants in such cases.

Q6.5. Generators register jointly with a Prompter — can one team supply its own paired Generator AND Prompter, and how is the pairing declared?

Answer: Yes, the same team is allowed to participate in any track they wish. It is completely their

choice if they want to align/fine-tune their prompts and generators together. However, during the challenge, each prompt will be tested against each generator. So the individual prompts need to generalize well to other generators and vice-versa.

Q6.6. When will the approved NIST model API list (for RAG/agent generators under -network restricted) be published, and which models will it include?

Answer: NIST API endpoints from Azure Foundry Services will include models from GPT (OpenAI), Claude (Anthropic), Grok (xAI).

1.7 Prompter Track

Q7.1. How exactly is a Prompter scored — the average AUC-ROC (or believability) of the downstream Generator+Discriminator outcomes across the prompt set? What defines a “winning” prompt set?

Answer: Per §3.2.2, a Prompter is scored by the downstream impact of its prompts on generator/discriminator outcomes: driving low AUC-ROC (approaching 0.5, averaged across all prompts submitted with a given generator), high Brier (approaching 0.25), and high, stable believability profiles (high average believability with low variance). Prompter teams do not generate content themselves; their evaluation is based entirely on this downstream effect.

Q7.2. Do a Prompter’s system_prompt/user_prompt pairs run against a fixed NIST generator, the team’s own generator, or all generators?

Answer: Every prompt runs against every generator so the impact on generation performance can be isolated and studied.

Q7.3. Confirm the cap is submissionindex 1–3, whether all three are scored or only the best/last, and whether there is a length limit on system_prompt/user_prompt.

Answer: As mentioned in Appendix A, submissionindex takes values from 1 to 3, allowing upto 3 submissions. Every submission made by participants will be scored. With regards to prompt length, there is not explicit limit from NIST’s side. However, based on submission systems, there might be a implicit limit due to context window. However, that window should be more than sufficient in most cases.

1.8 Container, Determinism & Runtime

Q8.1. Could you share the exact input JSON schema for each containerized role and a dry-run sample input file? We want to confirm the /input structure and the compound statement_id format (e.g., 00001_0 / 00001_1).

Answer: The evaluation plan already specifies the JSON schemas for the prompter (Appendix A), generator and detector (Appendix C) tracks. A dry-run with sample data and input/files will be conducted with all the participants well before the actual challenge.

Q8.2. Are there per-container runtime limits — wall-clock, CPU cores, GPU memory — we should size to?

Answer: The best node in the cluster is with two 98GB GPU, 1.5TB memory, and 48 cores. The max runtime will be set experimentally after dry-run.

Q8.3. For a purely local system, is `-network none` acceptable? For systems using the NIST API, is `temperature=0` deterministic on your endpoints, and how should we treat residual API nondeterminism under the determinism rule (B.4)?

Answer: For purely local systems, fully offline operation with no inbound/outbound network communication is expected and preferred (Appendix B.5). The determinism rule specified in Appendix B.4 is intended for client-side controllable parameters like `temperature`, `top_k`, `top_p`, etc. However, we do understand the bit-for-bit determinism cannot be guaranteed simply by fixing the sampling parameters. Also, it has been observed in some of the latest model APIs that sampling parameters have been depreciated. We will take these situations into account and participants will not be penalized due to variance arising from server-side API variables.

Q8.4. Is there a container template / validation package (analogous to `usnistgov/test_code_eval`), public or via the registration/dry-run package, and when?

Answer: Yes, NIST will share the package via GitHub repository before the challenge opens the submissions.

Q8.5. Which Foundry SDK version and base image will you provide?

Answer: NIST does not distribute or require a custom Foundry SDK. If your system relies on an external framework like Microsoft Azure AI Foundry, you must install the relevant package from PyPI (e.g., via `pip`) into your Docker image during the build process. Please ensure that the exact SDK version you use is pinned in your `environment.yml` or `requirements.txt` submission manifest.

Q8.6. Any required output filename, or is any `*.json` under `/output` accepted?

Answer: The output filename may be chosen by the system, provided the file is written under `/output` and conforms to the required output format (Appendix B.8, Appendix C). There is no fixed required output filename.

Q8.7. Is the dry-run the place to validate a full container end-to-end, and are the dry-run and Phase 1/2 dates finalized? (The plan PDF and the live page differ.)

Answer: The dry-run will be the place where we will validate the entire submission packages end-to-end. The tentative schedule for a dry-run is End of September 2026 - October 2026 for now.

1.9 Intellectual Property, Retention & Licensing

Q9.1. What is NIST's policy on retaining, copying, or deleting submitted images and model weights after the evaluation? Would the registration agreement commit to deletion within a defined window, with written confirmation?

Answer: Participants must cease all use of and delete Data and Materials on the earlier of:

1. The date they cease to participate in the challenge,
2. The date NIST notifies them of their termination from the challenge,
3. The date the challenge concludes.

Q9.2. May a submission include a proprietary evaluation-only, no-redistribution, no-reverse-engineering LICENSE in the image, and does the registration agreement already cover confidentiality of participant models?

Answer: All submissions and communications to NIST are subject to public disclosure under the Freedom of Information Act, 5 U.S.C. 552. Participants therefore must not provide any proprietary information or material to NIST in their submissions.

Q9.3. For model-origin rule B.7: for a distilled student whose teacher is a US-origin foundation model, what qualifies as “US-domestic,” and are Google Gemma / Meta Llama / OpenAI GPT-OSS / Microsoft Phi / AI2 OLMO in-scope US-origin bases?

Answer: As mentioned in Appendix B.7, the challenge only permits US-domestic models originating from US vendors, companies, or institutions, and prohibits fine-tuned, quantized, or derived variants of *restricted* models. In case of a distilled model, both student and teacher models have to comply with US-domestic origin requirements as stated above.

1.10 Logistics, Timeline & Eligibility

Q10.1. What is the maximum number of scored submissions per track, and when announced?

Answer: Per §6, participants will be notified of the maximum number of allowed submissions at the start of the evaluation. The specific number is not fixed in the current plan.

Q10.2. What metadata is required at submission (system-description template, resource-usage report), and is there a maximum image/file size?

Answer: The submission package must include container artifacts (Docker image plus Dockerfile/Docker-Compose and build scripts), a dependency manifest (`environment.yml`), a system description (using the Generator template in Appendix D or Discriminator template in Appendix E), documentation/README, and a resource-usage report (Appendix C, Submission Package Requirements). The maximum size limit of a Box upload per-submission is 250GB. For open-source models, this will include model weights, inference engine, KV cache etc. In addition to that we also have a limit of ~ 170 GB of GPU VRAM available for running the models. So please plan accordingly.

Q10.3. Confirm the post-evaluation workshop attendance expectation and that results are published with participant team names alongside the anonymized leaderboard.

Answer: As mentioned in Section 6 (Reporting and Publication), NIST will generate a report summarizing system results with participant team names, and participants may publish their own research with proper references. We will not be presenting any leaderboard as the systems are being evaluated along multiple axes and one might be better than the other based on the their specific application.