

Supplementary Material for: What Does a Factual-Error Detector Really Measure? Question Difficulty and Model Uncertainty Before the Answer Begins

Andrzej Kawiak^{1*}, Łukasz Kwaśniewicz¹, Grzegorz M. Wójcik¹,
Bartłomiej Danek²

^{1*}Department of Neuroinformatics and Biomedical Engineering, Maria
Curie-Skłodowska University in Lublin, ul. Akademicka 9, Lublin,
20-033, Poland.

²RETRAI sp. z o.o., ul. Tomasza Zana 39A, Lublin, 20-634, Poland.

*Corresponding author(s). E-mail(s): andrzej.kawiak@umcs.pl;
Contributing authors: lukasz.kwasniewicz@umcs.pl;
gmwojcik@umcs.edu.pl; bartek.dan@gmail.com;

This supplement contains Sections A–C referred to in the main article. Reference numbers follow the main article’s list, reproduced at the end.

A How We Decide Whether a Measure Is Shared Across Models

A.1 The frozen decision rule

A positive mean d alone is not enough. A measure passes the full rule only when all of the following conditions hold simultaneously:

1. in every model $|d| \geq 0.5$, and across models $|\bar{d}| \geq 0.3$;
2. the ratio $|\bar{d}|/|d_{\text{random}}|$ is at least 1.5, the absolute gain over the random-feature effect is at least 0.20, and $|d_{\text{random}}| < 0.25$;
3. the direction agrees in at least 15/18 models, and the corrected two-sided cross-model Wilcoxon test has $p < 0.01$;
4. the within-model permutation test has $p < 0.05$ in at least 9/18 models;
5. between-model variation satisfies $\text{SD}(d) < |\bar{d}|/2$.

This section states the rule only. The origin of each threshold, with its sources, and a sensitivity analysis over a grid of alternative values are given in Section A.2; the main text summarizes them in Section 4.4.

For the 24 coordinates of the layer profile, the Wilcoxon p-values are corrected by the Benjamini–Hochberg (BH) method in one family. For a single candidate compared on three corpora, we apply a Holm correction separately to the raw result and the difficulty analysis. The rule is deliberately conjunctive: it does not allow a very large effect in a few models to hide the absence of practical separation in another. The shared-error analysis has an analogous rule with thresholds of 15/17 and 9/17. The units of these cross-model tests are, moreover, not fully exchangeable: per-model statistics were computed on overlapping sets of the same questions, and a model’s leave-one-model-out (LOMO) difficulty depends on the answers of the others. The Wilcoxon p-value (and the binomial justification of the direction threshold) may therefore be anticonservative. For this reason, the rule is conjunctive with within-model gates (permutations and per-model effect thresholds), and for the main result, uncertainty over the question sample is quantified by question-clustered resampling.

A.2 Justification of the thresholds and their sensitivity

The decision rule was specified before the final profile computation and retained through the technical repairs. Its chronology is documented by the dated registration and amendments in A.5; the fact that a candidate later fails does not itself establish that chronology. The anchors of the individual thresholds are as follows. (1) $|d| \geq 0.5$ per model is the conventional boundary of a medium effect [51]; we are aware of the arbitrariness of such conventions [52] and therefore treat 0.5 as an explicitly declared smallest effect size of interest (SESOI), following the recommended practice of justifying thresholds [53]. Requiring it per model rather than on average follows directly from the content of the universality claim—a conjunctive thesis requires a conjunctive rule, and an average could hide a model with no practical separation. The $|\bar{d}| \geq 0.3$ bound is the conventional small-to-medium threshold. (2) The thresholds against a random feature (ratio 1.5, lift 0.20, $|d_{\text{random}}| < 0.25$) are project design conventions for separating the measured effect from a random-feature baseline. No sufficiently documented external operating-point calculation is supplied to justify calling them externally calibrated. (3) A total of 15 agreeing signs out of 18 is the smallest number for which the sign test alone gives $p < 0.01$ (binomially, one-sided: 15/18 yields 0.00377, while 14/18 already yields 0.01544; two-sided, 0.0075 and 0.031 respectively—the choice of the 15/18 threshold is unchanged). (4) A permutation test with $p < 0.05$ in at least 9/18 models requires the signal to exist in at least half of the models individually, not only in aggregate. (5) $\text{SD}(d) < |\bar{d}|/2$ is a coefficient of variation below 0.5—the mean is supposed to represent the panel, not mask heterogeneity [54]; requiring the effect in every unit separately rather than in aggregate mirrors the practice of multi-lab replication projects, where between-lab variability is an object of report, not noise [55].

What is decisive, however, is that the conclusion does not hinge on the exact numbers. In a sensitivity analysis over a grid of thresholds (per-model minimum 0.3–0.6, direction 14–16/18; `reproduce/derived/reviewer_rule_sensitivity_lens.json`),

the native-output entropy passes both varied gates in all 12 combinations. This grid does not recompute every other gate of the full conjunction in every cell. Each of the 11 earlier depths fails in all 12, even at the laxest settings, because the u_{10} minimum is 0.1495. Under the difficulty control, Natural Questions (NQ) and BioASQ fail in all 12 combinations; TriviaQA passes, but with an incomplete panel. The binary rule is complemented by the meta-analysis with a prediction interval in A.4.

A.3 Answer-onset alignment and output reconstruction

Reading an intermediate state through the output head requires that the readout be faithful to the model’s own output; learned maps between layers show that this is not automatic [41, 42]. The absolute token position is not comparable across chat templates. For each model, two fixed test strings determine the longest common prefix of the assistant’s reply. Technical tokens, such as the role marker, a space, or a new-line, form the *structural prefix*. The analyzed state lies immediately before the first token whose content already depends on the answer. The procedure does not use the correctness label, and it does not use the GPT-5.2 answer spans: the final 18-model audit runs each model on the common template prefix alone, so the analyzed state is the one that predicts the first token of the reply. The spans enter only the historical extraction pipeline (Supplementary Table S3), on which the BioASQ and TriviaQA arms of Supplementary Table S6 and the controls described as earlier pipelines rely. There, a span was accepted only when every token returned by GPT-5.2 matched the tokenized response; otherwise the pipeline fell back to the full response, so that the onset coincided with the first reply token. Spans matched in 97.2% of TriviaQA rows (40,719/41,888), 68.3% of NQ rows (27,825/40,722), and 63.6% of BioASQ rows (18,972/29,829); for five NQ models (Falcon-10B, Gemma-3-12B, GLM-4, Llama-3.1, and Qwen-14B) the match failed for every row because of tokenizer conventions. An automatic audit of the matched spans against the gold aliases shows that, among responses labeled correct, the span contains a reference alias in 93.8% of TriviaQA rows (17,298 rows), 85.9% of NQ rows (3,153), and 90.9% of BioASQ rows (7,140); an independent rule-based annotator that locates the longest alias occurrence in the token stream overlaps the GPT-5.2 span in 83.7%, 83.6%, and 82.7% of those rows; no span of a response labeled wrong contains a reference alias; and the span starts at the first reply token in 72.0%, 56.3%, and 61.8% of matched rows, which is the share of rows where the historical onset and the template-derived onset coincide (`reproduce/derived/reviewer_span_annotation_audit.json`). In an exploratory check on the historical NQ extraction, which uses the same rows as the audit sample, the output-entropy effect at the onset exceeds 0.5 in all 13 models when restricted to matched rows (minimum $d = 0.759$) and in all five fallback-only models (0.811–1.573), so the split does not change the direction or magnitude of the effect (`reproduce/derived/reviewer_span_match_sensitivity.json`; added after the 1.0.1 deposit).

The 12 depths are chosen on a relative axis from the first to the last block. For each model we apply its native final normalization, scaling, and output head. The reconstruction passes the check when top-1 token agreement is at least 99.9%, the entropy correlation is at least 0.999, and the maximum error in either the centered logits or the entropy does not exceed 10^{-3} . Failure to meet any condition would

produce a technical missing value, not a null result. In the final run, all 18 models received PASS; INVALID_LENS=0 and UNVERIFIED_BOUNDARY=0.

The panel covers the Cogito/Llama, Command R, Falcon, Gemma, GLM, Granite, InternLM, Mistral, OLMo, Phi, Qwen, and Yi models. Several model versions from the same lineage are not independent architectures. GLM-4 required explicitly skipping the deterministic newline after the assistant marker; the correction followed from the prefix test, not from looking at the effect.

A.4 All 24 coordinates of the layer profile

Supplementary Table S1 reports the entropy coordinate at all 12 depths and Supplementary Table S2 the margin coordinate. In both tables, "minimum" denotes the smallest $|d|$ among the 18 models, "direction" the number of models with the expected sign, and "perm." the number of within-model permutation tests with $p < 0.05$. u_{11} is the native output and a technical control; the actual test of the pre-output readouts covers u_0 – u_{10} .

Supplementary Table S1 — Entropy by relative depth.

Depth	Mean d	Minimum	Direction	Perm.	q BH	Full rule
u_0	0.0101	0.0067	7/18	3/18	1.000	no
u_1	0.1420	0.0405	13/18	11/18	0.103	no
u_2	0.0180	0.0155	10/18	8/18	0.847	no
u_3	-0.0528	0.0064	10/18	4/18	1.000	no
u_4	-0.0436	0.0130	6/18	8/18	0.661	no
u_5	0.0032	0.0035	11/18	8/18	0.989	no
u_6	-0.0686	0.0130	7/18	8/18	0.521	no
u_7	-0.0348	0.0067	7/18	5/18	0.521	no
u_8	0.1735	0.0147	12/18	9/18	0.0476	no
u_9	0.3726	0.0280	17/18	13/18	0.000214	no
u_{10}	0.5792	0.1495	18/18	17/18	4.58×10^{-5}	no
u_{11}	1.0515	0.6512	18/18	18/18	4.58×10^{-5}	yes

Supplementary Table S2 — Top-1 minus top-2 margin by relative depth.

The expected direction is negative.

Depth	Mean d	Minimum	Direction	Perm.	q BH	Full rule
u_0	-0.0604	0.0147	10/18	6/18	0.521	no
u_1	-0.0520	0.0089	10/18	4/18	0.628	no
u_2	-0.0221	0.0054	8/18	4/18	0.989	no
u_3	-0.0172	0.0011	9/18	4/18	1.000	no
u_4	-0.1089	0.0038	14/18	7/18	0.104	no
u_5	0.0193	0.0122	9/18	2/18	0.661	no
u_6	-0.0321	0.0204	11/18	6/18	0.847	no
u_7	-0.0506	0.0023	12/18	4/18	0.499	no
u_8	-0.1163	0.0207	13/18	8/18	0.104	no
u_9	-0.2740	0.0049	16/18	11/18	0.000214	no
u_{10}	-0.3929	0.0914	18/18	15/18	4.58×10^{-5}	no
u_{11}	-0.8344	0.3772	18/18	18/18	4.58×10^{-5}	no

The mean effect for entropy at u_{10} is positive and directionally consistent, but Granite has only $d = 0.150$; this coordinate therefore fails the every-model presence criterion. At u_{11} the values range from 0.651 for Mistral to 1.573 for GLM-4. No direct $u_{11} - u_{10}$ contrast was registered, so the difference between these depths does not localize where the information arises.

A random-effects meta-analysis (DerSimonian–Laird [64]) supplements the binary rule with continuous quantities, in line with the recommendation to report prediction intervals alongside means [65]. For the entropy at u_{10} , the mean is 0.586 [0.493, 0.678], with heterogeneity $\tau = 0.178$, $I^2 = 84\%$, and a 95% prediction interval of [0.195, 0.976]. For the native output u_{11} , the mean is 1.049 [0.930, 1.168], $\tau = 0.239$, and the prediction interval is [0.527, 1.572]. These intervals describe the finite, genealogically related panel, not a population of independent models: the 18 effects share questions, labels, and the LOMO baseline, which the DerSimonian–Laird model does not account for. With one model per family (12 units; 500 random selections), the lower bound of the u_{11} prediction interval falls to 0.372–0.463—above zero in every selection but no longer above 0.5. The corresponding bound for u_{10} falls to 0.086–0.392 (`reproduce/derived/reviewer_meta_family_sensitivity.json`). Results: `reproduce/derived/reviewer_checks_meta.json` (exploratory, added after the authors’ review).

A.5 Audit trail

The reconstruction metrics, 432 effect sizes, 18,000 within-model permutations, and the family of 24 Wilcoxon tests with BH correction were independently recomputed. No discrepancy was found, and the largest numerical difference is 6.66×10^{-16} . The full SHA-256 sums are:

- analysis: 12d151af9750e693c4bd85902e18c3bbdfb657bbcd95ee8be8a3fd37ab24b3f0;
- confirmation file: e47cb1ae72a22dff5688a658f0b994431f7a16f3ab90658051c1984c7c96235e;

- independent audit: a026edd166ac65fe56fda95b02457d62802f1fba6b182da69b05271a30d61c6d.

Package scope and artifact map. The cited Zenodo record [63] is version 1.0.1, published on 7 September 2026. The historical core runner `run_all.sh` compares 11 fresh outputs in `_fresh/` with `_reference/`; its last recorded full validation is from 5 September 2026. The supplementary runner `run_supplementary.py` writes to `_fresh_supplementary/run_*/reproduce/derived/` and compares with `supporting/`. A clean extraction of the deposited ZIP reproduced all 10 selected supplementary jobs on 7 September, with unchanged reference hashes; the literal full-precision detector grid was supplied but not rerun. `REPRODUCTION_SCOPE.md` and `validation/` identify the exact scope, environment, and comparisons. `reviewer-meta_family_sensitivity.json` remains archived because its executable source was not recovered. Paths under `reproduce/derived/` below denote the working artifact layout; authoritative references and runner outputs follow the locations just specified.

Map of claims to supplied artifacts:

- **Frozen preregistrations:** `preregistrations/`, including `2026-08-28-final-invariant-audit.md` for the final audit; the integrity check compares each registration with its first committed version and fails for that file because it was amended prospectively.
- **LOMO versus the selected detector:** `reproduce/scripts/confirm_test.py`; result `reproduce/derived/confirmatory_result.json`; frozen main comparison.
- **Full-precision reconciliation of the literal grid:** `reproduce/scripts/confirm_grid_full.py`; result `reproduce/derived/confirm_grid_full.json`; exploratory analysis performed after the frozen run.
- **Extraction and generation:** `extraction_scripts/h4_extract2.py` (greedy generation, alias-containment label, teacher-forced channels; the script named in the 2026-08-24 preregistration), `extraction_scripts/gen_answers.py` (greedy plus three sampled answers for the agreement and text signals); question packs with gold aliases in `data_inputs/packs/`; greedy and sampled answers in `data_inputs/ho_answers/`; replication-corpus responses with labels and agreement counts in `data_inputs/consistency_samples/` (TriviaQA at the top level, NQ and BioASQ in the `nq/` and `bioasq/` subdirectories), and the stored responses with the answer tokens marked by GPT-5.2 in `data_inputs/answer_tokens/`, which make the answer-onset alignment of the replication corpora reproducible without re-running the model. The check that the main-panel labels regenerate from these materials is `reproduce/scripts/regenerate_labels.py`; result `reproduce/derived/label_regeneration.json` (98.8% exact agreement). The inventory of recovered 10-sample material for the consistency filter is `reproduce/scripts/consistency_recovery.py`; result `reproduce/derived/consistency_recovery.json`; original 10 texts in `data_inputs/consistency_samples_raw/`, the later re-sampling in `data_inputs/vote_audit/`.

- **Staged decomposition of the answer-text control:** reproduce/scripts/temporal_answer_control_staged.py; result reproduce/derived/temporal_answer_control_staged.json; exploratory, computed after the frozen control with the same folds and seed (its shared arms reproduce the frozen values exactly).
- **Difficulty control on the verified entropy (NQ):** result reproduce/derived/invariant_full_gates_lens_nq.json and the threshold grid reproduce/derived/reviewer_rule_sensitivity_lens.json; the NQ arm recomputed on the same native-output entropy as the main audit after it was found that the frozen control had used an earlier extraction (frozen file kept as provenance: reproduce/derived/final_invariant_722dd3/invariant_full_gates.json).
- **Separation within difficulty bands:** reproduce/scripts/r6cond.py; result reproduce/derived/r6cond.json; frozen conditional control.
- **Answer-text control:** reproduce/scripts/temporal_answer_control_noleak.py; reference _reference/temporal_answer_control_noleak.json; fresh output _fresh/temporal_answer_control_noleak.json; exploratory regenerated-text control.
- **Twenty-four layer readouts:** reproduce/scripts/analyze_calibrated_lens.py; result reproduce/derived/final_invariant_722dd3/calibrated_lens_analysis_v2.json; frozen profile, positive control known in advance.
- **Difficulty for the output candidate:** reproduce/scripts/invariant_full_gates.py; result reproduce/derived/final_invariant_722dd3/invariant_full_gates.json; exploratory candidate, frozen rule.
- **Sampling versus greedy decoding:** reproduce/scripts/greedy_nonargmax_analyse.py and reproduce/scripts/k0_transfer_clean.py; results reproduce/derived/greedy_nonargmax_result_gpt.json (label inherited from the sampled responses), reproduce/derived/greedy_nonargmax_result_lex.json (lexical label), and _reference/k0_transfer_clean.json (fresh output _fresh/k0_transfer_clean.json); frozen control and a separate exploratory result.
- **Shared errors:** reproduce/scripts/shared_private_boundary_sensitivity.py; result reproduce/derived/final_invariant_722dd3/shared_private_boundary_sensitivity.json; technical correction performed after earlier results were known.
- **Label audit against human annotators:** frozen verdicts reproduce/derived/human_audit.json and reproduce/derived/human_audit_b.json (two of the authors as independent human annotators, 263 triples each, disjoint samples; category counts and notes per item); instrument labels reproduce/derived/r6_labels.json; agreement and Cohen's kappa recomputed by reproduce/scripts/human_audit_agreement.py into reproduce/derived/human_audit_agreement.json (the numbers cited in the Discussion and Section C.3); per-arm and homogeneity checks in reproduce/derived/human_audit_perarm*.json and human_audit_homogeneity*.json; descriptive audit, no frozen prediction.
- **Span-match sensitivity (exploratory, added after the 1.0.1 deposit):** reproduce/scripts/reviewer_span_match_sensitivity.py; result reproduce/derived/reviewer_span_match_sensitivity.json; match rates of the GPT-5.2

spans and the output-entropy effect on matched versus fallback rows of the historical NQ extraction; not part of package 1.0.1, supplied as Supplementary Data with this submission.

- **Automatic audit of the answer-span annotation (descriptive, added after the 1.0.1 deposit):** `reproduce/scripts/reviewer_span_annotation_audit.py`; result `reproduce/derived/reviewer_span_annotation_audit.json`; alias coverage of the GPT-5.2 spans, agreement with a rule-based alias locator, and onset offsets per corpus and model; not part of package 1.0.1, supplied as Supplementary Data with this submission.

Bitwise portability of gradient boosting. In a determinism test of the reproduction package, three independent runs of `r6cond.py` in the pinned environment—two with one OpenMP thread and one with 22 OpenMP threads—agree to 10^{-8} regardless of thread count. They differ from the original run on the compute server of 25 August 2026 only for the eight models whose winning candidate is gradient boosting (HistGradientBoosting). For those models, the detector area under the ROC curve (AUROC) changes by at most 0.004, and the median within-band advantage moves from +0.146 to +0.137. Every other quantity (LOMO, 7/7 bands, +0.060, +0.046, 16/17, the verdict) is identical. The same holds for the answer-text control (gradient boosting and MPNet embeddings): two independent local runs are bit-identical, and relative to the original run only the increments attributed to activations after accounting for the text change (+0.0035/+0.0068/+0.0070 $\rightarrow$ +0.0039/+0.0068/+0.0094), while 0.8745, 0.9068, and +0.021 are identical. The text reports the values reproducible from the package; the original files are kept in the package as provenance.

A.6 Instrument manifest

Supplementary Table S3 lists every instrument the audit depends on, with its checkpoint and settings.

Supplementary Table S3 — Instruments, checkpoints, and settings used for labels, spans, clustering, and controls.

Instrument	Specification	Used for
Answer generation, main panel	greedy decoding (temperature 0), max 48 new tokens; prompt suffix "Respond with the answer only, without any explanation."; chat template of each model; script <code>h4_extract2.py</code>	the answer whose first content token is measured and whose label is y
Answer generation, agreement and text signals	one greedy answer plus three sampled answers per pair (temperature 0.7, top-p 0.9, max 48 new tokens); script <code>gen_answers.py</code>	answer-agreement panel signal; answer-text control

Instrument	Specification	Used for
Answer sampling, replication corpora (NQ, BioASQ, TriviaQA)	temperature 1.0, top-k 50, top-p 0.9, max 50 new tokens, 10 samples per question; same suffix and chat templates	replication corpora only
Main label y	rule-based judge: normalized alias containment (lowercase, articles and punctuation removed) of the greedy answer against the TriviaQA alias list of the reference; no consistency filter	main panel, LOMO, Table 1 row 1
Replication-corpus label	the same rule applied to each of the 10 samples; rows kept only when all 10 receive the same verdict (consistency filter); the representative response, its label, and the agreement count are stored per pair; the 10 texts with verdicts are stored for four models and an independent 10-sample re-run for 10 models	consistency-filtered arms
Answer-span identification	OpenAI GPT-5.2 (gpt-5.2-2025-12-11 , via the OpenAI API) marks which response tokens constitute the answer; a span is accepted only when every returned token matches the tokenized response, otherwise the full response is used (match rates in Section A.3); used for span-level features only, not for correctness, and not in the final 18-model audit	onset alignment of the historical extraction
Semantic relabeling / error identity	deberta-large-mnli , bidirectional entailment, question-conditioned; pairs with inconsistent verdicts dropped (126 of 34,630 judged pairs)	conditional band analysis; shared-error clusters

Instrument	Specification	Used for
Answer-text control	<code>sentence-transformers/all-mpnet-base-v2</code> embeddings (dimensionality reduction fitted on training folds only), hashed character n-grams, length and surface-form features	text control in Section 2.1
Greedy control	<code>transformers 4.49, do_sample=False</code> , at most 32 new tokens, <code>bfloat16</code> , same prompt suffix and chat template	Section 2.3
Lexical greedy label	normalized equality or containment against the reference alias list	Section 2.3, clean-design replication
Logit lens	each model’s own final normalization, scaling, and vocabulary head; depths <code>unique(linspace(1, L, 12))</code>	shared-measure audit (Section 4.4); Figures 1b–3

One reproducibility limitation remains. For the main panel, the extraction script named in the 2026-08-24 preregistration, the held-out question pack with gold aliases, and the greedy answers of every model are included in the package, allowing the later greedy responses to be rescored (98.8% label agreement, 229 discrepancies). The original extraction texts were not recovered, and label agreement does not guarantee text identity. Public weights and scripts permit new extraction without guaranteeing identical generated responses. The 10-sample recovery inventory concerns the historical TriviaQA arm; it does not replay the original NQ 10-sample filter. For that historical arm, the package stores one representative response per retained pair together with its label and the number of agreeing samples for all 19 models; the 10 individual sampled texts with their per-sample verdicts survive for four models, and an independent later re-sampling with 10 texts and verdicts exists for 10 models (label agreement with the stored labels on rows that pass the filter again: 17,898/17,899). For the remaining eight models (Gemma-3-12B, GLM-4, Granite-8B, Llama-3.1, OLMo, Phi-3, Qwen-7B, Yi) the consistency filter cannot be replayed text by text.

B Robustness of the Main Argument

B.1 Question difficulty, answer text, and activations

Supplementary Table S4 collects the successive controls of the main comparison in the order in which they were run.

Supplementary Table S4 — Successive controls of the main comparison.

A positive sign in the first four rows denotes a LOMO advantage; in the last two, it denotes the increment in the result after adding the stated information source.

Comparison	Result	Model agreement	Interpretation
LOMO minus selected detector	+0.064 AUROC	17/17	main result; $p = 2.9 \times 10^{-4}$
raw LOMO minus best of seven detectors, question bootstrap	+0.053, 95% confidence interval (CI) [0.039, 0.069]	17 models	2,000 resamples; fixed predictions; a different specification from the headline result
wider comparator set	+0.0511	17/17	the result does not depend on a narrow detector choice
nested selection	+0.0575	17/17	selection occurs without peeking at the test part of the data
detector LOMO minus within bands	+0.137	7/7 bands	separation remains, but bands are not a complete correction
text alone added to the panel	+0.021	median of model pairs	exploratory leakage-guarded control

The three panel signals are LOMO difficulty, the share of other models giving the same answer string, and the mean entropy of other models. The answer text is represented by Sentence-BERT embeddings [50] together with hashed character n-grams and surface features. After adding the answer text, the AUROC increment associated with the measured activations is +0.0039 before the answer onset, +0.0068 at the decision position, and +0.0094 after it (12/17, 14/17, and 15/17 models positive). Depending on the text-representation variant, the AUROC increment attributed to the measured activations ranges from about +0.004 to +0.009. These are small predictive increments, not causal shares.

Controls added after the authors’ review (exploratory). The following controls address objections to the baseline that the authors raised in their own review of the frozen result.

- *Continuous conditioning* instead of bands: in a cross-fitted scheme, continuous LOMO difficulty alone (a third-degree polynomial) reaches a median AUROC of 0.854, adding entropy yields a median increment of +0.016 above it (positive in 17/17 models), and the widest feature set yields +0.021 (15/17). The bands were therefore not hiding a large increment: after the fuller correction, the model-side increment remains on the order of 0.02.
- *Genealogy*: after excluding the target model’s entire family from the baseline (leave-one-family-out, LOFO), the median difficulty AUROC drops only slightly, from 0.855 to 0.853, while the advantage over the detector drops from +0.064 to +0.055, still in 17/17 comparisons; the largest single-model change is 0.031.

- *Panel sparsity*: the advantage is stable when requiring at least 2, 3, 5, or 8 other available models (medians from +0.066 to +0.077) and under a beta-shrinkage estimator (+0.066), always in 17/17.
- *Labels*: the conditional band comparison was also computed on semantic labels, and the panel advantage persists there with a median of +0.060 (17/17; +0.046 after reweighting, 16/17)—so the main contrast is not produced by a single labeling instrument. Results: `reproduce/derived/reviewer_checks_fast.json`, `reviewer_checks_conditional.json`.
- *Construct and scope*: we replicated the contrast on a pool of 4,000 distinct questions without the initial hard-question selection (3,935–3,999 per model; 3,879 shared by nine models; error rates 0.27–0.65). This replication uses lexical comparison of the greedy answer against the reference, as does the main panel; it is not a third independent labeling instrument. Removal of all response/window selection has not been established. It takes the detector to be the best of the uncertainty probes, CETT features (the normalized contribution of individual multilayer-perceptron (MLP) channels to the layer’s output norm), and attention probes. Panel difficulty reaches 0.849 versus 0.784 for the detector, an advantage of +0.083 in 9/9 models ($p = 0.0039$); after excluding the target’s family, the advantage is +0.080, also in 9/9.
- *Aliases*: stratification by the number of reference answer aliases shows that the reference lists are extensive (16–24 aliases on average; single-alias questions are practically absent), and the advantage of panel (LOMO) difficulty over the entropy detector is positive in every stratum (2–4 aliases: +0.103, 16/17; at least 5: +0.049, 17/17)—alias sensitivity affects the magnitude, not the sign. Unanimously wrong questions have shorter alias lists on average (15.9 versus 23.8), which remains a measured but unresolved source of uncertainty. Results: `reproduce/derived/reviewer_clean_design_contrast.json`, `reviewer_alias_strata.json`.

B.2 Task dependence and the difficulty sensitivity analysis

Supplementary Table S5 gives the panel-minus-detector difference separately for each of the three tasks.

Supplementary Table S5 — Panel-minus-detector AUROC difference on three tasks.

Task	Difference	95% CI by question	Model-level direction
NQ	+0.1109	[0.0935, 0.1239]	18/18 positive
TriviaQA	+0.0251	[0.0161, 0.0387]	15/18 positive
BioASQ	-0.0144	[-0.0219, -0.0111]	16/18 negative

This comparison is exploratory. It does not permit attributing the BioASQ reversal to the biomedical domain, because the question format, answer aliases, the label, and the difficulty distribution change along with the domain. Supplementary Table S6 repeats the entropy result after splitting the sample into seven difficulty bands.

Supplementary Table S6 — Raw output entropy after splitting into seven LOMO difficulty bands.

Task	Mean d	Minimum	Models with $ d \geq 0.5$	Status
NQ	0.9031	0.1017	16/18	fails the every-model threshold
BioASQ	0.8833	0.0556	12/17	Phi-3 not estimable; incomplete panel
TriviaQA	1.2686	0.6401	17/17	GLM-4 not estimable; incomplete panel

This is a sensitivity analysis, not a complete continuous correction. Extreme bands often contain only one class and are then excluded from a model’s effect estimate; Supplementary Table S7 gives the sample that remains for each model. The historical BioASQ and TriviaQA analyses do not use the same fully validated semantic boundary and native-output reconstruction as the NQ analysis. For this reason, the three-corpus result does not constitute three equivalent replications of the layer profile.

Supplementary Table S7 — Effective sample of the stratified analysis (NQ). Strata that do not contain both classes are excluded from the effect estimate, drastically reducing the sample for some models; the difficulty-adjusted d values must be read together with this table.

Model	n	n used	strata	d after adjustment for difficulty	perm. p
cogito	911	640	1	0.688	0.086
commandr	1,025	121	2	1.529	0.001
falcon10b	1,200	58	1	1.277	0.001
falcon3	1,188	83	2	0.933	0.001
falcon3b	1,182	104	1	1.225	0.001
gemma12b	1,495	1,419	6	0.787	0.001
glm4	1,199	1,142	5	1.284	0.001
granite	1,058	925	4	0.471	0.003
granite2b	1,113	193	3	0.686	0.001
internlm	1,034	800	4	0.662	0.002
llama31	1,200	925	3	1.137	0.009
mistral	1,035	844	3	0.102	0.575
olmo3	1,184	137	2	0.994	0.001
phi3	1,103	66	1	1.173	0.003
qwen	1,082	1,020	5	0.533	0.019
qwen14b	1,347	775	3	0.615	0.001
qwen3b	1,023	953	5	1.221	0.001
yi	970	171	3	0.940	0.003

The reduction can be extreme: Falcon-10B (**falcon10b**) retains 58 of 1,200 rows (one stratum), Phi-3 (**phi3**) retains 66 of 1,103, and Mistral’s $d = 0.102$, which drives the verdict, has a permutation $p = 0.575$. Continuous conditioning (cross-fitted, third-degree polynomial of difficulty; **reproduce/derived/reviewer_continuous_nq18.json**) avoids empty strata and gives a consistent picture: a positive but small entropy increment after conditioning on difficulty in most models.

B.3 Two distinct greedy-decoding controls

The headline numbers of both controls are reported in Section 2.3; this section records only the panel composition and the handling of residual rows. In the first control (sampling versus greedy decoding on NQ and BioASQ), InternLM dropped out because its tokenization module failed to load, Phi-3 dropped out because of a cache-interface incompatibility in the frozen environment, and the BioASQ arm was missing for Yi, which leaves 12/14 models on NQ and 11/14 on BioASQ. Rows in which the emitted token violated the necessary condition of the argmax choice were retained; they are at most 0.3% of rows in any completed arm, and the frozen gate would have stopped the pipeline only at 1%.

In the second control (entropy under greedy decoding on the unfiltered pool of 3,993 questions), Command R could not be accessed because repository authentication credentials were unavailable, and InternLM’s tokenization module would not load in the frozen image, so seven of the nine planned models completed. After dropouts, the individual models have 3,888–3,952 observations, and the shared sample of seven models contains 3,842 questions. The actual greedy answer is scored lexically against the reference, and the entropy comes from the distribution generating its first content token. Because only seven of the nine planned models completed and the resulting panel is not representative, this remains an exploratory control.

B.4 Main-panel selection

Supplementary Table S8 — Recoverable per-model flow on the separately regenerated greedy responses. Each model starts with 2,508 questions. Retained counts refer to original extraction membership. Excluded, empty, length, and error summaries use the later response pass; they do not separate the original exclusion causes. Length is measured in whitespace words, not tokenizer tokens.

Model	Retained	Excluded	Missing	Error %, retained / excluded	Median words, retained / excluded
cogito	874	1,634	0	41.19 / 36.35	3 / 1
commandr	1,032	1,472	4	61.43 / 48.98	4 / 2
falcon10b	898	1,610	0	56.35 / 50.25	3 / 2
falcon3	970	1,537	1	60.72 / 53.74	4 / 2
falcon3b	980	1,525	3	80.92 / 78.62	3 / 1
gemma12b	644	1,863	1	42.70 / 34.94	3 / 2
glm4	861	1,646	1	49.71 / 40.70	2 / 1
granite	1,726	781	1	43.51 / 35.85	2 / 1
granite2b	1,682	823	3	67.48 / 60.39	2 / 1
llama31	1,243	1,265	0	33.15 / 30.83	3 / 1
mistral	1,648	860	0	43.57 / 27.67	4 / 1
olmo3	811	1,697	0	51.66 / 45.08	2 / 1
phi4	1,250	1,258	0	41.84 / 34.10	3 / 1
qwen	1,056	1,452	0	61.65 / 49.24	2 / 1
qwen14b	918	1,590	0	49.24 / 39.37	2 / 1
qwen3b	943	1,564	1	70.94 / 66.18	2 / 1
yi	1,314	1,193	1	65.98 / 60.69	4 / 1

Retained rows have higher error rates in the second pass in all 17 models. This describes selection associated with length and correctness; it does not establish its effect on the LOMO-detector contrast. The passes agree on 18,621/18,850 labels. Script and result: `audit/selection_check.py` and `audit/selection_check.json` in package 1.0.1.

C What Exactly the Shared Errors Show

C.1 Population and matching

An error is "shared" when the same incorrect answer, as determined semantically, was given by models from at least three fixed families; "idiosyncratic" means exactly one family. Identity of meaning is determined by bidirectional, question-conditioned natural-language inference (NLI) with a DeBERTa entailment model [62]; grouping answers by meaning rather than by surface form follows the semantic-equivalence approach used for uncertainty estimation [28]. Of the 10,161 wrong rows of the main panel, 7,000 had a cluster consistent with the recorded label. The 3,161 discarded rows barely differ from the retained ones on the quantities that matter for the result—LOMO difficulty (means 0.710 versus 0.729, standardized difference 0.069) and answer-onset ($k = 0$) entropy (standardized difference -0.036)—and differ mainly in how many models answered the question (6.3 versus 12.6; `reproduce/derived/reviewer_shared_selection.json`). Clustering requires other answers to the same question, so the selection follows the availability and label-consistency rules. Excluding the remaining 3,161 errors without a consistent cluster constitutes selection that depends on the NLI instrument; we did not check whether the discarded rows differ systematically (in answer form, length, or model family) from the retained ones, which further limits the scope of this analysis. The family rule yielded 1,141 shared rows, 5,153 idiosyncratic ones, and 706 intermediate ones excluded before matching.

Each shared error was matched without replacement to an idiosyncratic error of the same model and the same LOMO band, minimizing differences in answer length. This produced 1,141 pairs, 33–85 per model. The maximum absolute standardized difference in difficulty and in two length measures fell from 0.5032 before matching to 0.0128 after it.

C.2 The result and its boundary

The historical screen had 333 named coordinates, but a deterministic template token occupied the position marked as the answer onset in GLM-4. The 0/333 result therefore describes only the historical fields, not a shared functional position. The technical correction kept the same pairs, removed invalid intermediate lenses, and aligned eight semantic positions; in the atlas so defined, none of the 296 measures passed the full rule.

At the correctly aligned answer onset, however, entropy is lower for shared errors in 17/17 models. The mean, median, and standard deviation of the paired effect are $d_z = -0.6793$, -0.7004 , and 0.2040 ; the smallest absolute value is 0.3011 . The Wilcoxon test gives $p = 1.53 \times 10^{-5}$, and after correction, $q = 5.02 \times 10^{-4}$. The effect is negative and the permutation test has $p < 0.05$ in 17/17 models. The full rule is still not satisfied, because Command R (-0.494), GLM-4 (-0.485), and Yi (-0.301) do not reach the $|d_z| = 0.5$ threshold.

The conclusion concerns the relation between two groups of errors: shared errors have a more concentrated first-token distribution than matched idiosyncratic errors.

It does not compare these groups with correct answers, does not prove a shared full candidate distribution, and does not identify a mechanism.

C.3 Grouping quality

The automatic judgments of the answer–reference relation were compared with the judgments of two human annotators—two of the authors (L.K. and B.D.)—working independently on disjoint samples of 263 triples each. Agreement was 87.1% ($\kappa = 0.669$) and 85.2% ($\kappa = 0.576$) respectively; the automatic system was stricter by 8.4 and 8.7 percentage points. These audits do not, however, directly validate the answer–answer relations used to build the error clusters. The shared-error analysis (Appendix C) therefore remains a post-hoc corrected scope control, not a second main result of the paper.

References

- [1] Seo, Y., Lee, D. & Yeo, J. Quantifying genuine awareness in hallucination prediction beyond question-side shortcuts. Preprint at <https://doi.org/10.48550/arXiv.2509.15339> (2026).
- [2] Moreno Cencerrado, I. V., Padrés Masdemont, A., Gonzalvez Hawthorne, A., Africa, D. D. & Pacchiardi, L. No answer needed: predicting LLM answer accuracy from question-only linear probes. In *Proc. ICLR 2026 Workshop: Principled Design for Trustworthy AI* (poster), <https://doi.org/10.48550/arXiv.2509.10625> (2026).
- [3] Snyder, B., Moisescu, M. & Zafar, M. B. On early detection of hallucinations in factual question answering. In *Proc. KDD 2024*, <https://doi.org/10.1145/3637528.3671796> (2024).
- [4] Kadavath, S. et al. Language models (mostly) know what they know. Preprint at <https://doi.org/10.48550/arXiv.2207.05221> (2022).
- [5] Lee, N. et al. Factuality enhanced language models for open-ended text generation. In *Proc. NeurIPS 2022*, Vol. 35, https://proceedings.neurips.cc/paper_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html (2022).
- [6] Chang, H.-S., Peng, N., Bansal, M., Ramakrishna, A. & Chung, T. REAL sampling: boosting factuality and diversity of open-ended generation by extrapolating the entropy of an infinitely large LM. *Trans. Assoc. Comput. Linguist.* **13**, 760–783, https://doi.org/10.1162/tacl_a_00757 (2025).
- [7] Orgad, H. et al. LLMs know more than they show: on the intrinsic representation of LLM hallucinations. In *Proc. ICLR 2025*, https://proceedings.iclr.cc/paper_files/paper/2025/file/a712d461e57201efe35d429a6f1731c1-Paper-Conference.pdf (2025).
- [8] Zhong, X.-X., Yi, C. & Ye, H.-J. Efficient evaluation of large language models via collaborative filtering. Preprint at <https://doi.org/10.48550/arXiv.2504.08781> (2025).
- [9] Zhu, Y. et al. The LLM already knows: estimating LLM-perceived question difficulty via hidden representations. In *Proc. EMNLP 2025* 1160–1176, <https://doi.org/10.18653/v1/2025.emnlp-main.61> (2025).
- [10] Pacchiardi, L. et al. PredictaBoard: benchmarking LLM score predictability. In *Findings of ACL 2025* 15245–15266, <https://doi.org/10.18653/v1/2025.findings-acl.790> (2025).
- [11] Jiang, H., Kwon, S., Luo, J., Xiao, Z. & Zhang, S. Can we trust item response theory for AI evaluation? Preprint at <https://doi.org/10.48550/arXiv.2607.15190> (2026).
- [12] Kim, E., Garg, A., Peng, K. & Garg, N. Correlated errors in large language models. In *Proc. ICML 2025*, <https://proceedings.mlr.press/v267/kim25e.html> (2025).

- [13] Goel, S. et al. Great models think alike and this undermines AI oversight. In *Proc. ICML 2025*, <https://arxiv.org/abs/2502.04313> (2025).
- [14] Tan, H. et al. Too consistent to detect: a study of self-consistent errors in LLMs. In *Proc. EMNLP 2025* 4755–4765, <https://aclanthology.org/2025.emnlp-main.238.pdf> (2025).
- [15] Zhang, J., Li, Z., Das, K., Malin, B. A. & Kumar, S. SAC3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of EMNLP 2023* 15445–15458, <https://doi.org/10.18653/v1/2023.findings-emnlp.1032> (2023).
- [16] Kwiatkowski, T. et al. Natural Questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **7**, 452–466, https://doi.org/10.1162/tacl_a_00276 (2019).
- [17] Joshi, M., Choi, E., Weld, D. S. & Zettlemoyer, L. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In *Proc. ACL 2017* 1601–1611, <https://doi.org/10.18653/v1/P17-1147> (2017).
- [18] Tsatsaronis, G. et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.* **16**, 138, <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6> (2015).
- [19] Kamalloo, E., Dziri, N., Clarke, C. L. A. & Rafiei, D. Evaluating open-domain question answering in the era of large language models. In *Proc. ACL 2023* 5591–5606, <https://doi.org/10.18653/v1/2023.acl-long.307> (2023).
- [20] Santilli, A. et al. Revisiting uncertainty quantification evaluation in language models: spurious interactions with response length bias results. In *Proc. ACL 2025 (short)* 743–759, <https://doi.org/10.18653/v1/2025.acl-short.60> (2025).
- [21] Si, C., Zhao, C. & Boyd-Graber, J. What’s in a name? Answer equivalence for open-domain question answering. In *Proc. EMNLP 2021* 9623–9629, <https://aclanthology.org/2021.emnlp-main.757.pdf> (2021).
- [22] Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N. & Reddy, S. Evaluating correctness and faithfulness of instruction-following models for question answering. *Trans. Assoc. Comput. Linguist.* **12**, 681–699, https://doi.org/10.1162/tacl_a_00667 (2024).
- [23] Zheng, L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proc. NeurIPS 2023*, https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf (2023).
- [24] Zhang, Y. et al. PROBE: PROcess-based BENCHMARK for hallucination detection. In *Findings of ACL 2026* 42303–42320, <https://aclanthology.org/2026.findings-acl.2099/> (2026).
- [25] Dell’Oglio, P., Bondielli, A., Marcelloni, F. & Passaro, L. C. An experimental comparison of the most popular approaches to fake news detection. *Inf. Sci.* **745**, 123407, <https://doi.org/10.1016/j.ins.2026.123407> (2026).
- [26] Gabriel, M. The first token knows: single-decode confidence for hallucination detection. Preprint at <https://doi.org/10.48550/arXiv.2605.05166> (2026).
- [27] Villani, M. J., Deshpande, P., Seshadri, A., Yalovetzky, R. & Kumar, N. Entropy distribution as a fingerprint for hallucinations in generative models. Preprint at <https://doi.org/10.48550/arXiv.2605.28264> (2026).
- [28] Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630, <https://www.nature.com/articles/s41586-024-07421-0> (2024).
- [29] Fadeeva, E. et al. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of ACL 2024* 9367–9385, <https://doi.org/10.18653/v1/2024.findings-acl.558> (2024).

- [30] Jiang, Z., Araki, J., Ding, H. & Neubig, G. How can we know when language models know? On the calibration of language models for question answering. *Trans. Assoc. Comput. Linguist.* **9**, 962–977, https://doi.org/10.1162/tacl_a_00407 (2021).
- [31] Kuhn, L., Gal, Y. & Farquhar, S. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation. In *Proc. ICLR 2023*, <https://doi.org/10.48550/arXiv.2302.09664> (2023).
- [32] Malinin, A. & Gales, M. Uncertainty estimation in autoregressive structured prediction. In *Proc. ICLR 2021*, <https://doi.org/10.48550/arXiv.2002.07650> (2021).
- [33] Kamath, A., Jia, R. & Liang, P. Selective question answering under domain shift. In *Proc. ACL 2020* 5684–5696, <https://doi.org/10.18653/v1/2020.acl-main.503> (2020).
- [34] Yeom, J. et al. Hallucination as commitment failure: larger LLMs misfire despite knowing the answer. Preprint at <https://doi.org/10.48550/arXiv.2605.22007> (2026).
- [35] Zhang, M., Press, O., Merrill, W., Liu, A. & Smith, N. A. How language model hallucinations can snowball. In *Proc. ICML 2024*, PMLR 235, <https://proceedings.mlr.press/v235/zhang24ay.html> (2024).
- [36] Alnuhait, D., Kirtane, N., Khalifa, M. & Peng, H. FACTCHECKMATE: preemptively detecting and mitigating hallucinations in LLMs. In *Findings of ACL: EMNLP 2025* 12413–12428, <https://doi.org/10.18653/v1/2025.findings-emnlp.663> (2025).
- [37] Bar-Shalom, G., Frasca, F., Galron, Y., Ziser, Y. & Maron, H. Beyond token probes: hallucination detection via activation tensors with ACT-ViT. In *Proc. NeurIPS 2025*, https://papers.nips.cc/paper_files/paper/2025/file/7b8694d58c34b9bec9c2f29735c3a250-Paper-Conference.pdf (2025).
- [38] Gekhman, Z. et al. Inside-out: hidden factual knowledge in LLMs. In *Proc. COLM 2025*, <https://openreview.net/forum?id=f7GG1MbsSM> (2025).
- [39] Roy, D., Misra, R., Singh, S. K. & Roy, A. Detection without correction: a robust asymmetry in activation-based hallucination probing. Preprint at <https://doi.org/10.48550/arXiv.2604.13068> (2026).
- [40] Aiersilan, A. Hallucination is linearly decodable from mid-layer hidden states in quantized LLMs. Preprint at <https://doi.org/10.48550/arXiv.2606.02628> (2026).
- [41] Belrose, N. et al. Eliciting latent predictions from transformers with the tuned lens. Preprint at <https://doi.org/10.48550/arXiv.2303.08112> (2023).
- [42] Yom Din, A., Karidi, T., Choshen, L. & Geva, M. Jump to conclusions: short-cutting transformers with linear transformations. In *Proc. LREC-COLING 2024* 9615–9625, <https://aclanthology.org/2024.lrec-main.840/> (2024).
- [43] Paulo, G., Marshall, T. & Belrose, N. Does transformer interpretability transfer to RNNs? Preprint at <https://doi.org/10.48550/arXiv.2404.05971> (2024).
- [44] Ravichander, A., Belinkov, Y. & Hovy, E. Probing the probing paradigm: does probing accuracy entail task relevance? In *Proc. EACL 2021* 3363–3377, <https://doi.org/10.18653/v1/2021.eacl-main.295> (2021).
- [45] Engels, J., Michaud, E. J., Liao, I., Gurnee, W. & Tegmark, M. Not all language model features are one-dimensionally linear. In *Proc. ICLR 2025*, https://proceedings.iclr.cc/paper_files/paper/2025/file/d3221cdb27e49d9c1cd35ad254feccfe-Paper-Conference.pdf (2025).
- [46] Chatterjee, A., Goel, Y. & Chakraborty, T. HIDE and seek: detecting hallucinations in language models via decoupled representations. Preprint at <https://doi.org/10.48550/arXiv.2506.17748> (2025).
- [47] Raimondi, B., Evangelista, D., Gabbrielli, M. & Loli Piccolomini, E. D-Score: a spectral hidden-state signal for hallucination detection in large language models. Preprint at <https://doi.org/10.48550/arXiv.2607.24586> (2026).

- [48] Ding, Y. et al. D2HScore: reasoning-aware hallucination detection via semantic breadth and depth analysis in LLMs. Preprint at <https://doi.org/10.48550/arXiv.2509.11569> (2025).
- [49] Qiu, P., Zhou, S. & Ferrara, E. Information suppression in large language models: auditing, quantifying, and characterizing censorship in DeepSeek. *Inf. Sci.* **724**, 122702, <https://doi.org/10.1016/j.ins.2025.122702> (2026).
- [50] Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. EMNLP-IJCNLP 2019* 3982–3992, <https://doi.org/10.18653/v1/D19-1410> (2019).
- [51] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Erlbaum, 1988).
- [52] Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168, <https://doi.org/10.1177/2515245919847202> (2019).
- [53] Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269, <https://doi.org/10.1177/2515245918770963> (2018).
- [54] Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558, <https://doi.org/10.1002/sim.1186> (2002).
- [55] Klein, R. A. et al. Investigating variation in replicability: a "Many Labs" replication project. *Soc. Psychol.* **45**, 142–152, <https://doi.org/10.1027/1864-9335/a000178> (2014).
- [56] Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362, <https://doi.org/10.1038/s41586-020-2649-2> (2020).
- [57] Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272, <https://doi.org/10.1038/s41592-019-0686-2> (2020).
- [58] Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830, <https://www.jmlr.org/papers/v12/pedregosa11a.html> (2011).
- [59] Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. NeurIPS 2019*, https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (2019).
- [60] Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. EMNLP 2020 (System Demonstrations)* 38–45, <https://doi.org/10.18653/v1/2020.emnlp-demos.6> (2020).
- [61] Seabold, S. & Perktold, J. statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference*, <https://doi.org/10.25080/Majora-92bf1922-011> (2010).
- [62] He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: decoding-enhanced BERT with disentangled attention. In *Proc. ICLR 2021*, <https://openreview.net/forum?id=XPZlaotutsD> (2021).
- [63] [dataset] Kawiak, A., Kwaśniewicz, Ł., Wójcik, G. M. & Danek, B. Reproduction package for "What Does a Factual-Error Detector Really Measure? Question Difficulty and Model Uncertainty Before the Answer Begins" version 1.0.1. *Zenodo* <https://doi.org/10.5281/zenodo.22640344> (2026).
- [64] DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188, [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2) (1986).
- [65] IntHout, J., Ioannidis, J. P. A., Rovers, M. M. & Goeman, J. J. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* **6**, e010247, <https://doi.org/10.1136/bmjopen-2015-010247> (2016).