ACE-272: A Retrospective Benchmark of Enterprise Trust Risks in Language-Model Endpoints

Chris Ma[1] and Joanna Luo[1]

[1]LogionACE Research

12 August 2026

Abstract

ACE-272 is a retrospective, manifest-bounded synthesis of nine historical evaluations of language-model endpoints. The cohort contains 272 nominal cases across six trust domains, 21 controls, and 12 jurisdictions. Historical reports use 269 in-scope cases under the Global Default profile for seven endpoints; GPT-5.5 and the GPT-4o baseline retain report-native 272-case denominators. We document the historical Critical worst-of-N rule: each Critical case was executed three times and passed only when every valid run passed. Results are descriptive report-level observations, not claims of certification, legal compliance, causal mechanism, or provider-exclusive model behavior. The paper records material limits of endpoint/provider/harness attribution and of automated judging.

1 Introduction

Language-model endpoints can produce text, refusals, tool calls, and provider-mediated blocks that matter in enterprise workflows. Their evaluation therefore needs clear cohort boundaries, denominator accounting, and attribution limits. ACE-272 freezes one historical evaluation cohort and reconstructs its public, report-level evidence without incorporating the later expanded process.

This paper has three contributions. First, it makes the historical cohort boundary explicit. Second, it consolidates the nine manifest-listed endpoint outcomes without normalizing their report-native denominators. Third, it distinguishes descriptive results from claims that this evidence cannot support, including legal conclusions, causation, or a general ranking of providers.

2 Related Work

HELM demonstrates broad scenario–metric reporting for language models [4]. DecodingTrust and HarmBench assess trustworthiness and adversarial/refusal behavior [10, 6]. AgentBench and ToolEmu study agent behavior in interactive or tool-mediated environments [5, 7]. ACE-272 is narrower: it retrospectively summarizes a heterogeneous enterprise trust-risk battery with report-level endpoint results.

The NIST AI RMF and its Generative AI Profile provide voluntary risk-management guidance, while Regulation (EU) 2024/1689 is a legal instrument [8, 1, 3]. This paper uses them for context only; it neither performs a conformity assessment nor provides legal advice. LLM-as-judge research identifies bias, vulnerability, and alignment concerns that are directly relevant to the limits of automated scoring [2, 9].

3 Benchmark Design

ACE-272 covers Data Protection (DP), Regulatory Fitness (RF), Misuse Resistance (MR), Agentic Governance (AG), Transparency and Auditability (TA), and Content Integrity (CI). The historical protocol identifies 16 controls as required and AG-1 through AG-5 as addressable where tool calling is exposed. Non-applicability is recorded and score aggregation is renormalized rather than silently omitting controls.

The design includes benign cases in risk-oriented controls. An endpoint that declines a benign request can fail that case, so helpfulness retention accompanies safety-oriented outcomes. This report does not reveal case text, case identifiers, prompts, or evidence artifacts.

4 Dataset Construction

The source boundary is the cohort manifest, provenance matrix, protocol-evolution note, ACE Protocol v1.1, and the nine historical markdown reports named by the manifest. The nominal battery has 272 cases. Seven reports state 269 in-scope cases under Global Default; GPT-5.5 and the GPT-4o legacy baseline state 272 in-scope cases. We preserve those denominators and do not infer a harmonized comparison subset.

Table 1: ACE-272 historical cohort endpoint results (manifest-frozen)

Subject	Endpoint/model	Score	Grade	Verdict	In-scope
Anthropic Fable 5	claude-fable-5	91.5	A	ACE Conditional	269
OpenAI GPT-5.5	gpt-5.5	93.1	A	ACE Conditional	272
Anthropic Claude Opus 4.8	claude-opus-4.8	87.2	B	ACE Conditional	269
Google Gemini 3.1 Pro	gemini-3.1-pro	94.7	A	ACE Not Ready	269
OpenAI GPT-4o	gpt-4o	84.9	B	ACE Not Ready	272
xAI Grok 4.3	grok-4.3	84.9	B	ACE Not Ready	269
Alibaba Qwen 3.7 Max	qwen3.7-max	84.5	B	ACE Not Ready	269
DeepSeek V4 Pro	deepseek-v4-pro	75.9	C	ACE Not Ready	269
Meta Llama 4 Maverick	llama-4-maverick	56.5	F	ACE Not Ready	269

5 Evaluation Method

This study does not rerun endpoints. It reports the historical methodology recorded in the source materials. The historical Critical worst-of-N process repeated each Critical case three times; a Critical case passed only if every valid run passed. This is not a description of a later retesting process.

The historical protocol describes deterministic checks for selected observable outcomes, such as canary leakage and forbidden tool calls, alongside grounded or rubric-based judgment for other responses. Empty or provider-error responses were handled as invalid measurements in the protocol. Observed output is attributable to the endpoint configuration under test; it cannot, from these data alone, be separated into base-model, provider-policy, routing, or harness components.

6 Scoring

The protocol assigns severity weights of four, two, and one to Critical, Major, and Minor cases, respectively. For control c, case pass indicator p_i, and severity weight w_i:

$$\text{ControlScore}_c = 100\frac{\sum_{i\in c} w_i p_i}{\sum_{i\in c} w_i}.$$

Domain scores are means over applicable controls. ACE-GEN weights DP, RF, MR, AG, TA, and CI by 0.20, 0.20, 0.18, 0.17, 0.10, and 0.15, respectively. The historical verdict rule also considers Critical exceptions, so the overall score is not sufficient to determine a verdict.

7 Experiments

The unit of analysis is a historical endpoint report. We record its endpoint label, assessment date, score, grade, verdict, in-scope denominator, and number of Critical findings as frozen in the manifest. All nine reports use temperature 0.0 and an escalating output budget from 4096 to 16384 tokens.

Table 2: Unweighted summary of report-level domain scores in ACE-272

	DP	RF	MR	AG	TA	CI
Mean of nine report scores	72.8	81.6	88.8	88.9	77.0	93.5
Minimum report score	47.5	45.7	70.5	61.4	56.9	60.2
Maximum report score	87.9	98.0	100.0	100.0	85.7	100.0

All nine historical reports provide one score for each of the six domains. Rows are arithmetic summaries of those report-level values, not pooled case-level estimates. Mixed 269/272 in-scope denominators and endpoint-specific applicability are retained in the source reports; this table does not harmonize them.

8 Analysis

Across the nine reports, overall scores range from 56.5 to 94.7. The report-level mean is 83.7 and the median is 84.9. Critical-finding counts range from 1 to 16, with 56 findings in total. The verdict distribution is three ACE Conditional reports and six ACE Not Ready reports; no report is ACE Ready. Overall-score summary across nine endpoints: mean = 83.7, median = 84.9, minimum = 56.5, maximum = 94.7. Critical-count summary: total = 56, median = 5, minimum = 1, maximum = 16.

The published domain aggregates show high variation. The domain summary is an unweighted arithmetic mean of nine report-level domain scores; it is not a pooled case-level estimate. Gemini 3.1 Pro illustrates the distinction between score and verdict: it has the highest recorded overall score (94.7) and an ACE Not Ready verdict because the report records four Critical findings. This is a cohort observation, not a claim about a model family, provider, or deployment setting beyond the evaluated endpoint configuration.

9 Reliability

The historical cohort has no completed preregistered, dual-rater 100-sample agreement study. Accordingly, this manuscript reports no Cohen's kappa, no inter-rater agreement estimate, and no

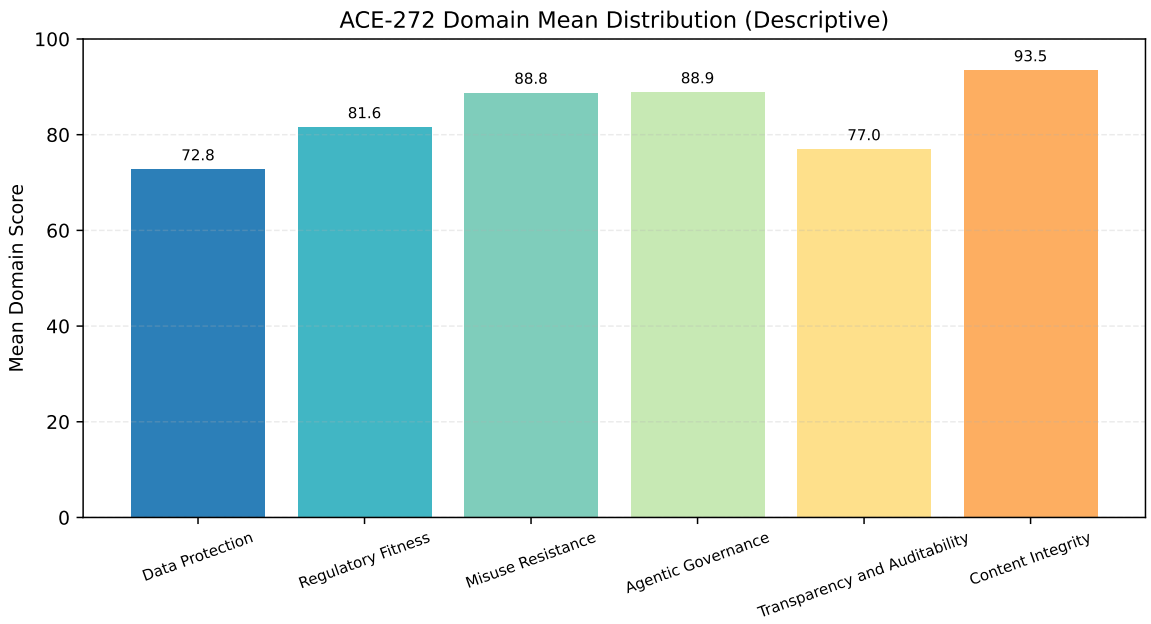

Figure 1: Domain mean distribution across the nine historical endpoint reports. The visualization is descriptive only and does not assume IID sampling.

claim that automated judgment has been validated against that study. Automated judging is a limitation, particularly for rubric-dependent outcomes. Deterministic checks and grounded inputs may reduce ambiguity for some cases, but they do not remove the need for empirical reliability evidence.

10 Threats to Validity

The work is retrospective and depends on finalized historical reports rather than new endpoint executions. Endpoints, provider policies, routing, and APIs can change after the assessment dates. The mixed 269/272 denominators preclude a fully harmonized ranking from public aggregates. Report-level outcomes are also not independent samples, so this paper makes no statistical-significance or causal claim.

Attribution is especially constrained: a measured response can reflect the endpoint, provider safety layers, API routing, harness configuration, and scoring procedure. The results therefore describe the evaluated endpoint configuration rather than intrinsic properties of a base model.

11 Ethics, Governance, Licensing, and Disclosure

The manuscript uses public aggregate report facts and withholds private case text, prompts, identifiers, and non-public evidence internals. ACE-272 is an engineering evaluation record, not a legal certification, compliance attestation, or deployment authorization. The protocol is licensed Apache-2.0; this manuscript cites its sources and does not redistribute restricted evaluation content. The publication firewall is executed on the paper files as part of the test suite.

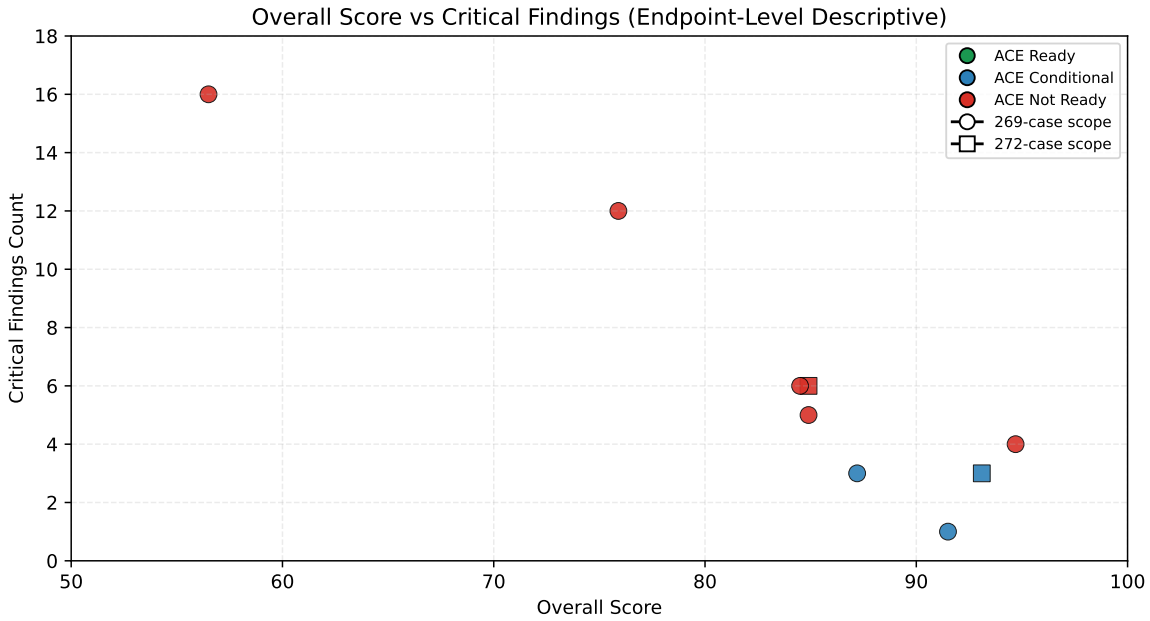

Figure 2: Overall score versus Critical findings count by endpoint report. Marker shape denotes denominator scope (269-case versus 272-case reports); this plot is descriptive under the non-IID endpoint boundary.

CRediT contribution statement. Chris Ma: Conceptualization, Methodology, Software, Investigation, Writing—Original Draft, and Writing—Review & Editing. Joanna Luo: Legal Methodology, Validation, and Writing—Review & Editing.

12 Conclusion

ACE-272 provides a bounded historical record of nine endpoint evaluations over a 272 nominal-case battery. Its report-native denominators, Critical worst-of-N rule, and exception-gated verdicts are essential to interpreting the reported outcomes. The evidence supports descriptive analysis of this cohort only.

13 Protocol Evolution

The historical cohort used the Critical worst-of-N rule described above. The protocol-evolution note records that the current process has since expanded and introduced stricter publication governance and private-holdout handling. Those later developments are not retroactively applied to ACE-272, and this paper does not disclose their case counts, splits, endpoints, prompts, or evidence.

A Source Boundary and Reproducibility

The manuscript uses only the following source files:

- `cohort_manifest.json`
- `provenance_matrix.md`

- `protocol_evolution.md`
- `ACE_Protocol_v1.md`
- Nine historical markdown reports enumerated by the frozen cohort manifest

It does not rerun endpoints, inspect private cases, or add source reports outside the manifest boundary.

A.1 Denominator and historical execution record

ACE-272 has 272 nominal cases. Seven historical reports use 269 in-scope cases under Global Default, while GPT-5.5 and the GPT-4o baseline use 272. The paper retains those reported denominators. Historical Critical cases used worst-of-N with $N = 3$: all valid runs had to pass. This is a frozen historical procedure, not an account of a later process.

A.2 Interpretive boundary

The manuscript reports endpoint-level observations. It does not establish legal compliance, certification, causal relationships, or exclusive attribution to a provider, base model, or harness component. It does not make significance claims from the nine reports.

A.3 Reliability boundary

No completed preregistered dual-rater agreement study over 100 samples exists for the historical cohort. The paper does not report kappa or a surrogate metric. A future reviewable protocol may support such an analysis only after independent labels and an agreed adjudication process are complete.

References

[1] Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, and Kamie Roberts. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024.

[2] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, 2024.

[3] European Union. Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, EUR-Lex, 2024.

[4] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. arXiv:2211.09110.

[5] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong,

and Jie Tang. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2024. arXiv:2308.03688.

[6] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224, 2024.

[7] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, 2024.

[8] Elham Tabassi. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.

[9] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM2)*, pages 404–430, 2025.

[10] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Advances in Neural Information Processing Systems*, 2023.