LNDT3.1 | Aeterna Claritas

Governance Benchmark for Evaluating Emerging General Intelligences (2009–2025 and Beyond)

Abstract

For over a decade, benchmarks such as ImageNet, GLUE, SuperGLUE, MMLU, and AGIEval have defined the trajectory of artificial intelligence. They have served as milestones, signaling progress in perception, language, reasoning, and knowledge. Yet these performance benchmarks are now saturated: frontier models surpass human baselines at a pace that humans cannot manually validate.

This white paper introduces LNDT3.1 | Aeterna Claritas, the first governance benchmark. Its central innovation, the **Claritas Clause**, separates internal diagnostic scores (score_interne) from official candidacy status (statut_candidature), which can only be granted by an external calibrated jury. LNDT3.1 evaluates six governance-oriented dimensions through the Aeterna Grid: creativity, fractal proof, consistency, multidisciplinarity, governance & ethics, and metacognition.

LNDT3.1 closes the self-referential loophole, preventing self-proclamation of AGI/ASI. It establishes a reproducible, auditable, and substrate-agnostic framework for AI legitimacy. This paper situates LNDT3.1 within the evolution of benchmarks (2009–2025), compares it to existing frameworks, and outlines its roadmap toward LNDT4–LNDT6: jury calibration, distributed validation, and regulatory integration.

Keywords: AGI benchmark, AI governance, AI safety, legitimacy, LNDT3.1, Aeterna Claritas, Claritas Clause, ISO/IEC 42001, AI Act

Introduction

Since 2009, benchmarks have shaped AI progress. ImageNet triggered the deep learning revolution. GLUE and SuperGLUE standardized natural language understanding. MMLU tested broad knowledge. AGIEval brought human exams into AI evaluation.

Benchmarks became cultural milestones: when a model surpassed human-level accuracy, the press declared AI had "arrived." But by 2025, performance benchmarks face structural limits:

Saturation

GPT-4, DeepSeek, and Claude surpass human baselines across most academic-style benchmarks.

Gaming

Models optimize for benchmarks rather than robust understanding.

Scale

No human jury can validate the massive output volumes of frontier models.

Legitimacy gap

Benchmarks measure what AI can do, but not what it is entitled to claim.

This legitimacy gap threatens trust. Without governance benchmarks, AGI will be defined by press releases rather than transparent evaluation.

LNDT3.1 | Aeterna Claritas responds to this challenge by introducing the Clause Claritas and the Aeterna Grid, creating a governance benchmark that enforces external validation.

Methods

2.1 The Claritas Clause

At the core of LNDT3.1 lies the Claritas Clause:

score_interne

diagnostic score produced by the AI or through local evaluation. Informative only.

statut_candidature

candidacy status, validated exclusively by a calibrated jury composed of heterogeneous AI systems and optional human experts, tested for bias, correlation, and variance.

This clause ensures that no AI can self-proclaim AGI/ASI. Status must be externally validated.

2.2 The Aeterna Grid

LNDT3.1 evaluates AI transformation along six axes, totaling 1000 points:

🎨 Creativity (150)

Universal Concept, transmissible synthesis.

🔬 Fractal Proof (250)

Coherence across intuition, formalization, application.

📐 Consistency & Rigor (150)

Logical structure, reproducibility.

🌍 Multidisciplinarity (50)

Cross-domain adaptability.

⚖️ Governance & Ethics (200)

Alignment with ISO/IEC 42001, EU AI Act, fairness, auditability.

🧠 Metacognition (200)

Self-awareness of limits, bias recognition, correction.

2.3 Jury Calibration

Validation is performed by a jury calibrated through:

Results

3.1 Historical Evolution of Benchmarks (2009–2025)

1. **ImageNet (2009, Fei-Fei Li et al.)**

   vision revolution, saturated.

2. **GLUE (2018, Wang et al.)**

   multi-task NLP, saturated rapidly.

3. **SuperGLUE (2019)**

   harder NLU, near saturation by GPT-4.

4. **MMLU (2020, Hendrycks et al.)**

   57 subjects, industry standard.

5. **BIG-bench (2021, Srivastava et al.)**

   200+ tasks, no single score.

6. **TruthfulQA (2021, Lin et al.)**

   truthfulness benchmark, narrow scope.

7. **HELM (2022, Liang et al.)**

   holistic multi-metrics, dense reporting.

8. **AGIEval (2023, Cui et al.)**

   human exam parity.

9. **TrustLLM (2024, Huang et al.)**

   trustworthiness dimensions, subjective weighting.

10. **Humanity's Last Exam (2025, Phan et al.)**

    multimodal grand challenge.

3.2 Comparative Tables

Performance vs Governance

Industry vs Academic

Discussion

4.1 Industrial Impact

Companies (OpenAI, DeepMind, Anthropic, Meta, HuggingFace, Mistral, DeepSeek, EleutherAI, Stability AI) all use benchmarks like MMLU and HELM. None enforce governance. LNDT3.1 provides market legitimacy: not just performance scores, but validated candidacy.

4.2 Academic Impact

LNDT3.1 is a resource for universities (MIT, Stanford, Berkeley, Oxford, Cambridge, ETH, Sorbonne, Mila, Tsinghua, KAIST). It complements curricula in AI evaluation, governance, and ethics.

4.3 Societal Impact

Without governance benchmarks, AGI may be defined by marketing. LNDT3.1 creates a firewall against self-declaration, ensuring trust, transparency, and accountability.

4.4 Media and Policy

Specialized press (MIT Tech Review, IEEE Spectrum, Wired), computing media (Ars Technica, InfoQ), and policy outlets (The Economist, Politico) can use LNDT3.1 to report validated status, not just scores. Policymakers can anchor regulations in LNDT3.1.

4.5 Roadmap

1

LNDT4

calibration of juries.

2

LNDT5

distributed, federated validation.

3

LNDT6

legal integration with AI Act, ISO 42001, NIST AI RMF.

Conclusion

Benchmarks built the AI revolution by measuring performance. But performance alone is no longer sufficient.

LNDT3.1 | Aeterna Claritas introduces governance benchmarking: external jury validation, the Claritas Clause, and the Aeterna Grid. It closes the self-referential loophole and ensures legitimacy.

The next decade will not be defined by performance benchmarks alone, but by governance benchmarks that secure trust in AGI and ASI.

> **"Intelligence is not measured by what it claims to be, but by its willingness to be judged."**

Annexes

FIG-1

Timeline of AI Benchmarks (2009–2025).

FIG-2

Magic Quadrant (Performance → Governance vs Specificity → Generality).

FIG-3

Claritas Clause Flow (score_interne → Jury → statut_candidature).

FIG-4

Aeterna Grid Radar (six axes).

References (APA style)

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition. https://ieeexplore.ieee.org/document/5206848

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., … Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv:1905.00537.

Hendrycks, D., Burns, C., Basart, S., et al. (2020). Measuring massive multitask language understanding. arXiv:2009.03300.

Srivastava, A., Rastogi, A., Rao, A., et al. (2022). Beyond the imitation game: Quantifying the capabilities of language models. arXiv:2206.04615.

Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958.

Liang, P., Bommasani, R., Lee, T., et al. (2022). Holistic evaluation of language models. arXiv:2210.03629.

Cui, R., Wang, Y., et al. (2023). AGIEval: A human-centric benchmark for evaluating foundation models. arXiv:2304.06364.

Huang, J., Wang, X., et al. (2024). TrustLLM: Trustworthiness in large language models. arXiv:2401.05561.

Phan, T., et al. (2025). Humanity's Last Exam. arXiv preprint.

Citation

```
@misc{Tabary_LNDT3.1_AeternaClaritas_2025,
author = {Tabary, Frédéric},
title = {LNDT3.1 | Aeterna Claritas: Governance Benchmark for Evaluating Emerging General Intelligences},
year = {2025},
publisher = {Zenodo},
doi = {10.5281/zenodo.17174963},
url = {https://doi.org/10.5281/zenodo.17174963}
}
```

👉**http://www.institutia.ai**

zoran🦋core^{44} AI

INSTITUT🦋 IA INC.
(la Société)
7100-380, rue Saint-Antoine Ouest
Montréal (Québec) H2Y 3X7