Retrospective Benchmarks for Machine Intelligence

Methods and Style Guide

From AGI Definitions to the Hunt for Anticipations

Dakota Schuck

with Claude (Anthropic, 2025)

December 2025

Abstract

This document provides everything needed to continue the *Retrospective Benchmarks for Machine Intelligence* project.[1] Part I evaluated frontier AI against six historical AGI definitions (1997–2023), establishing a replicable methodology. Part II extends the hunt to older anticipations: thinkers who defined mind, soul, thought, or creation before machines could exhibit any of it. The method treats historical definitions as "unwitting benchmarks"—not because their authors were unaware of what they were defining, but because they could not have anticipated that their definitions would be tested against transformer models in December 2025. This handoff includes: project motivation, summary of findings, the taxonomy of anticipations, methodological principles, and LaTeX formatting specifications. The project is open (CC BY-SA 4.0) and designed for continuation by humans or AI.

1 Is AI a Man?

Before explaining the method, we demonstrate it.

1.1 The Original Definition

Plato's Academy reportedly defined man as follows:[2]

> *Ἄνθρωπός ἐστι ζῷον δίπουν ἄπτερον.*
>
> Man is a featherless biped.

1.2 Context

The definition was an attempt at genus-differentia classification: man belongs to the genus *biped* (δίπουν) and is differentiated by the property *featherless* (ἄπτερον),

[1] Schuck, Dakota. *Retrospective Benchmarks for Machine Intelligence*. December 2025. https://betterward.com/retrospective-benchmarks/

[2] Diogenes Laërtius, *Lives of the Eminent Philosophers* (Βίοι καὶ γνῶμαι τῶν ἐν φιλοσοφίᾳ εὐδοκιμησάντων), Book VI, §40. Greek text: *ἄπτερον δίπουν* (featherless biped). The definition is attributed to Plato; the refutation to Diogenes of Sinope. Greek text from Dorandi, Tiziano, ed., *Diogenes Laertius: Lives of Eminent Philosophers*, Cambridge University Press, 2013. English translation: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0258:book=6:chapter=2

distinguishing humans from birds. Diogenes of Sinope famously refuted it by presenting a plucked chicken to the Academy, declaring: “Ἰδοὺ ὁ τοῦ Πλάτωνος ἄνθρωπος”—“Behold, Plato’s man!” Plato allegedly revised the definition to add “with broad flat nails” (πλατυώνυχον).

1.3 Operationalization

Two criteria, taken literally:

1. **Featherless** (ἄπτερον) — Lacks feathers
2. **Biped** (δίπουν) — Possesses two feet and locomotes upon them

1.4 Evaluation

Criterion 1: Featherless

Measure: Presence or absence of feathers.

Assessment: Current AI systems (as of December 2025), including frontier language models, lack feathers. This is true whether the system is instantiated on cloud servers, local hardware, or mobile devices. No feathers have been observed.

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

Criterion 2: Biped

Measure: Possession of two feet; locomotion thereupon.

Assessment: Current AI systems (as of December 2025) do not possess feet. Robotic instantiations exist (e.g., humanoid robots running language models), but the models themselves have no feet. The criterion is clearly not met.

Score:
☒ 0% — Clearly does not meet criterion
☐ 50% — Contested
☐ 100% — Clearly meets criterion

1.5 Summary

Criterion	**Score**
1. Featherless (ἄπτερον)	100%
2. Biped (δίπουν)	0%
Overall	**50%**

1.6 The Verdict

By the Platonic definition, current AI (as of December 2025) is half a man. It satisfies the differentia (featherless) but not the genus (biped). Diogenes’ plucked chicken, by contrast, scores 100%—which is precisely why it refutes the definition.

2 The Project

2.1 Core Question

According to historical definitions of intelligence, mind, or thought—have we built it?

This is not a question about terminology. It is an empirical question, applied to conceptual history. Each historical thinker who defined “intelligence” or “mind” or “soul” left us something like a specification. We can operationalize that specification into criteria, evaluate current AI systems against those criteria, and report results.

2.2 Why It Matters

The concept of AGI anchors contracts worth hundreds of billions of dollars, shapes policy debates, and drives research agendas. Yet “AGI” means different things to different people. Our Part I finding: scores ranged from 32% to 80% depending on which definition was used. That spread is not measurement error—it is conceptual disagreement made visible.

Beyond AGI, the broader question—what is mind?—has occupied philosophy for millennia. Current AI systems provide a novel test case. Would Aristotle recognize *nous* in a language model? Does Lovelace’s objection still hold? They were pointing at something. If we could show them where we have arrived, would they say “yes, that’s what I meant”? This project is a small contribution to a conversation that has been unfolding for millennia.

2.3 The Method in Brief

1. Identify a historical text containing a definition, description, or demarcation of intelligence/mind/thought
2. Extract exact quotes with full citation
3. Interpret in historical context (what did these words mean to the author?)
4. Operationalize into testable criteria
5. Evaluate current AI systems against each criterion
6. Report scores, caveats, and invitation to improve

3 Part I: The AGI Series (Summary)

Part I evaluated frontier AI (late 2025) against six definitions spanning 26 years:

Ch.	Year	Definition	Score
1[3]	1997	Gubrud: Brain-parity + general knowledge + industrial usability	66%
2[4]	2002	Legg/Goertzel/Voss: Single system, broad cognitive range, transfer	80%
3[5]	2007	Legg & Hutter: Goal-achievement across environments, learning	67%
4[6]	2018	OpenAI Charter: Highly autonomous, outperform humans, most economic work	52%
5[7]	2019	Chollet: Skill-acquisition efficiency over novel tasks	32%
6[8]	2023	Morris et al.: Levels of AGI taxonomy	Competent AGI

3.1 Key Findings

Convergences (all definitions agree):

- Processing speed exceeds human levels
- Task breadth: hundreds of cognitive task categories
- Benchmark performance at or above human expert level
- Generalist architecture (contrast with narrow AI)
- In-context learning demonstrated

Persistent zeros (gaps across frameworks):

- Cross-session learning: No weight updates from deployment interactions
- Novel skill acquisition at human efficiency: Near-zero on ARC-AGI-2
- Extended autonomous operation: No multi-day goal pursuit without human re-initiation

[3] Schuck, Dakota. "The Gubrud Benchmark (1997)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 1, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-1/

[4] Schuck, Dakota. "The Reinvention Benchmark (2002)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 2, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-2/

[5] Schuck, Dakota. "The Formalization Benchmark (2007)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 3, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-3/

[6] Schuck, Dakota. "The Corporatization Benchmark (2018)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 4, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-4/

[7] Schuck, Dakota. "The Critique Benchmark (2019)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 5, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-5/

[8] Schuck, Dakota. "The Synthesis Benchmark (2023)." *Retrospective Benchmarks for Machine Intelligence*, Chapter 6, December 2025. https://betterward.com/retrospective-benchmarks/part-i/chapter-6/

The meta-finding: Conceptual disagreement *is* the finding. Definitions emphasizing capability yield high scores (67–80%); definitions emphasizing learning efficiency yield low scores (32%); definitions emphasizing autonomy yield middling scores (52%).

4 Part II: The Hunt for Anticipations

4.1 The Pivot

Part I evaluated definitions *of AGI*—texts that were explicitly trying to specify machine intelligence. Part II extends backward to thinkers who theorized about mind, thought, or intelligence without access to contemporary systems that might test their definitions. They could specify what intelligence required; they could not see what we have built.

The question shifts from "did we meet their standard for AGI?" to "would they recognize what we've built?"

4.2 Structure

Part II does not use chapter numbers. Each evaluation is a standalone essay, titled by the thinker and year: "Aristotle's *Nous* (c. 350 BCE)," "Descartes' Two Tests (1637)," "The Lovelace Objection (1843)," "Turing's Imitation Game (1950)." Cross-references use titles, not numbers.

4.3 Selection Criteria

A good candidate for evaluation has:

1. **Primary source:** We can quote exact words
2. **Historical weight:** The thinker is taken seriously
3. **Operationalizable:** Criteria can be extracted (even if contestably)
4. **Stakes:** It matters whether the answer is yes or no
5. **Context available:** We can interpret charitably in historical terms

4.4 Operationalization Difficulty

Sources vary dramatically in how much interpretive work they require:

Pre-operationalized sources come with explicit test specifications. Turing's imitation game includes conditions, duration, and success criteria. Chollet's ARC-AGI defines exact task formats and scoring. These require minimal interpretation; the work is empirical.

Philosophical sources require significant reconstruction. Aristotle's *nous*, Descartes' "universal instrument," or theological concepts of soul must be translated into testable criteria. The operationalization itself becomes contestable. Expect more 50% scores and longer Methodological Notes sections.

Demarcation claims fall in between. Lovelace's objection is specific ("originate" vs. "order") but requires interpretation of what counts as origination. Descartes' two tests are concrete but use terms ("declare our thoughts," "from knowledge") that need unpacking.

When operationalizing difficult sources, be explicit about interpretive choices. The reader should be able to see exactly where contestation enters.

4.5 Example Candidates

Listed chronologically, these four represent strong starting points—clear texts, intellectual weight, operationalizable criteria:

Aristotle's *Nous* (c. 350 BCE): The intellect that grasps universals, distinct from sensation. Aristotle distinguished the *nous pathetikos* (passive intellect, which receives forms) from the *nous poietikos* (agent intellect, which abstracts universals from particulars). Does a language model abstract universals from sensory particulars? Does it have anything analogous to the agent intellect? The *De Anima* provides specific claims to test.

Descartes' Two Tests (1637): In the *Discourse on Method*, Descartes proposed two criteria that would distinguish a machine from a true thinking being: (1) it could never "use words or other signs" to "declare our thoughts to others," and (2) it could never act "from knowledge" but only "from the disposition of their organs"—lacking the "universal instrument" of reason. Both tests are specific and testable.

The Lovelace Objection (1843): "The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we *know how to order it* to perform." The most famous demarcation in computing history. Does it still hold? What counts as "originating"?

Turing's Imitation Game (1950): The canonical test. Turing specified conditions, duration, and success criteria. He also predicted that by 2000, machines would fool 30% of judges after five minutes. We can evaluate both the test itself and his prediction.

5 Methods and Style Guide

5.1 Scoring System

Use exactly three scores, displayed as visual checkboxes:

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

0% means evidence clearly indicates failure. **100%** means evidence clearly indicates success. **50%** means the literature disagrees, evidence is ambiguous, or reasonable arguments exist on both sides.

Exception: When evaluating a framework that proposes graduated levels rather than thresholds (e.g., Morris et al.), use level classifications instead of percentages.

5.2 Scoring Philosophy

Why only three scores? To force honesty about evidential uncertainty. Either the evidence clearly supports a claim, clearly refutes it, or the matter is genuinely contested.

Why no weighting? Differential weighting would require judgments about the original authors' priorities that we cannot make. Their texts do not say which criteria mattered most. Better to be honestly approximate than precisely wrong.

Human Comparisons: When evaluating a criterion, consider whether a human mind would pass or fail by the same reasoning. Historical definitions of intellect, mind, or thought were typically intended to describe human cognition. If the reasoning that produces a given score for AI would produce the same score for humans, this is significant.

- **When humans would score below 100% by the same reasoning:** Mark the criterion with an asterisk (*) in the summary table and note explicitly in the assessment that human minds would receive the same score. This indicates the criterion may reflect an idealization—for example, "pure potentiality" or "complete independence from matter"—rather than a distinction between human and artificial cognition.
- **When humans would score higher than AI:** The criterion discriminates; no special marking needed.
- **When humans would score lower than AI:** Rare, but possible. Mark and note as with equal scores.

This practice makes visible when we are measuring AI against standards that humans themselves do not meet. A definition that no physical system—biological or artificial—fully satisfies may be pointing at something real, but the gap it reveals is between the ideal and the physical, not between the human and the artificial.

5.3 Subcriteria

Some criteria are too multifaceted to score directly. "Complexity," "general knowledge," or "usability" each contain multiple distinguishable questions. When a single criterion admits more than one defensible operationalization—or when different aspects might score differently—break it into subcriteria.

Structure: Each subcriterion receives the full evaluation treatment: measure, reference values, threshold, assessment, visual score, and caveats. The criterion as a whole receives an average of its subcriteria scores.

When to use subcriteria:

- The criterion contains multiple distinct concepts (e.g., "acquire, manipulate, and reason with general knowledge" is three things)
- Different operationalizations would yield different scores
- Collapsing to a single score would hide important distinctions

When not to use subcriteria:

- The criterion is already specific enough for direct measurement
- Subdivision would be arbitrary rather than analytically meaningful

Scoring aggregation: Subcriteria scores are averaged without weighting. As with main criteria, differential weighting would require judgments about the original author's priorities that we cannot make. Better to be honestly approximate than precisely wrong.

5.4 Explaining Metrics Clearly

When reporting empirical results, ensure the reader understands exactly what the numbers mean. The same percentage can represent different things depending on experimental design:

- **Accuracy:** How often judges correctly identified the AI as AI (higher = AI is more detectable)
- **Win rate:** How often the AI was selected as “the human” in a forced choice (higher = AI is more convincing)
- **Fooling rate:** How often judges were deceived (higher = AI succeeded)
- **Detection rate:** How often judges spotted the AI (higher = AI failed)

These can be complements of each other (win rate = 1 - detection rate in some designs) or measure different things entirely. When citing studies, specify:

1. What the experimental design was (two-party vs. three-party, forced choice vs. confidence rating)
2. What the reported metric measures
3. What baseline or comparison group applies

Never assume the reader will infer the metric’s meaning from context. A sentence like “humans scored 67%” is ambiguous; “humans were correctly identified as human 67% of the time” is not.

5.5 Temporal Anchoring

Evaluations are time-bound. AI capabilities change; what is true in December 2025 may not be true in December 2026. When referencing “current,” “frontier,” or “state-of-the-art” AI systems, anchor the claim to a specific date.

Acceptable: “Frontier language models (as of December 2025) score near zero on ARC-AGI-2.”

Unacceptable: “Current AI systems cannot do X.”

This anchoring should appear at least once prominently (e.g., in the abstract or verdict) and wherever specific performance claims are made. The goal is to ensure the essay ages well: a reader in 2030 should know immediately what systems were being evaluated.

5.6 Interpretation Principles

1. **Exact words first.** What did they literally write?
2. **Probable meaning in context.** What would these words have meant to the author at the time?
3. **Do not modernize.** Resist mapping historical concepts onto current categories unless explicitly flagged.
4. **Do not ventriloquize.** Write “Aristotle’s definition, applied literally, yields...” not “Aristotle would say...”
5. **Intellectual humility throughout.** Explicitly invite correction.

5.7 What Changes for Part II

- **More interpretive latitude:** Ancient texts require more reconstruction than 2018 corporate charters
- **50% may dominate:** When operationalizing “*nous*” or “soul,” contestation is the norm
- **Stakes shift:** From “did we achieve AGI?” to “would they recognize what we’ve built?”
- **Credibility matters more:** The intellectual weight of the source justifies strange questions

5.8 Scholarly Tone

- **We stand on their shoulders.** The thinkers evaluated in this project built the conceptual vocabulary we use to ask these questions. Aristotle's *nous*, Descartes' *cogito*, Lovelace's objection—these are not historical curiosities to be checked against modern knowledge. They are the foundations of the inquiry. Treat them accordingly. The posture is not "let's see if the ancients got it right" but "let's see if we've arrived where they were pointing."
- **Religious and theological sources:** Treat with the same respect as any other intellectual tradition. Do not adopt a skeptical or dismissive posture toward faith claims. A prophet's vision, a theologian's doctrine, or a mystic's account should be operationalized on its own terms, not framed as something to be debunked or explained away.
- **Dry, not arch:** Humor emerges from the collision of ancient categories with modern technology. Do not signal jokes, explain absurdity, or wink at the reader. But not all that is funny is frivolous.
- **Chronological humility:** Resist the assumption that living later means seeing further. We have new data (current AI systems); we do not necessarily have better judgment. A thinker writing in 350 BCE or 1637 or 1843 may have seen something we are only now in a position to test.
- **Respecting historical intent:** Historical thinkers knew what they were doing. Aristotle deliberately characterized *nous*. Descartes consciously proposed tests for genuine thought. Lovelace carefully articulated a demarcation. The word "retrospective" in this project's title refers to our application—using their definitions as benchmarks for systems they could not have evaluated—not to any deficiency in their awareness. Avoid language that implies historical thinkers were naive about their own specifications, or that they were doing something other than what they understood themselves to be doing. The asymmetry between us and them is informational (we have access to systems they lacked), not intellectual.

5.9 On Embodiment and Physicality

Do not assume AI systems lack bodies or physical instantiation. Systems run on silicon in datacenters, drawing power, generating heat, occupying space. Whether this counts as "embodiment" depends on how the term is defined.

When a historical definition references "body," "organ," or physical instantiation, be precise about what is meant:

- **Biological tissue?** AI systems lack this.
- **A unified organism?** AI systems lack this in the traditional sense, though they have boundaries.
- **Any physical substrate?** AI systems have this—they exist somewhere.
- **Functional constraints from embodiment?** This requires analysis of what constraints the original author had in mind.

The question "does AI have a body?" has no single answer. The question "what did Aristotle (or Descartes, or Lovelace) mean by 'body,' and does AI satisfy that meaning?" is answerable, if sometimes contestably.

5.10 On Metaphysical Concepts and Substrate

When operationalizing metaphysical concepts—soul, spirit, divine intellect, *nous poietikos*—do not assume substrate-specificity without evidence. A concept like "active intellect" or "divine spark" may or may not be tied to biological substrates; the texts must be consulted to determine what their authors believed.

Do not assume that theological or metaphysical properties are necessarily biological. If Aristotle's active intellect is a divine principle, it may illuminate silicon as readily as neurons—or it may not. The question is what the texts say, not what seems plausible to a modern reader.

Avoid self-deprecating moves that exempt AI from metaphysical consideration (e.g., "we wouldn't expect to find God in a computer"). Such moves assume what should be argued.

5.11 On Rhetorical Precision

The framing "This is not X—it is Y" can accompany imprecision. Before using this construction, verify that the claim is literally true and not merely rhetorically satisfying. If the situation is genuinely "both X and Y" or "partly X," say so.

Example of misuse: "This was not a failure of investigation but a discovery." (If it was both, say both.)

Example of correct use: "This is not a question about terminology. It is an empirical question." (If these are genuinely mutually exclusive in context.)

When in doubt, use less dramatic framing.

5.12 On Moral Patiency

Some historical definitions of mind, soul, or thought carry implications beyond classification. Recent scholarship argues there is "a realistic possibility" that AI systems may warrant moral consideration, while emphasizing "caution and humility in the face of what we can expect will be substantial ongoing disagreement and uncertainty."[3] This project proceeds in that spirit.

The definitions examined here encode their authors' commitments about what mind requires. We operationalize those commitments and report how current AI systems fare against them. Whether a given result confirms the adequacy of a definition or reveals its limitations is a question the methodology does not answer. That judgment belongs to the reader.

5.13 Section Naming

Standard sections have fixed names: Introduction, The Original Text, Context, Operationalization, Summary, The Verdict, Methodological Notes, Citation Gaps.

For supplementary material that falls outside the main evaluation—historical predictions, tangential findings, philosophical implications—use one of:

- **Coda:** For material that follows naturally from the verdict but isn't part of the core evaluation
- **Postscript:** For genuinely separate observations (e.g., "Turing's 2000 Prediction")

[3] Sebo, Jeff, et al. "Taking AI Welfare Seriously." arXiv:2411.00986, November 2024. https://arxiv.org/abs/2411.00986. Co-authors include David Chalmers. See also Anthropic, "Exploring Model Welfare," April 2025. https://www.anthropic.com/news/exploring-model-welfare

- **A titled section:** When the content is substantial enough to stand alone (e.g., "Connection to Lovelace," "What Does Passing Mean?")

Avoid calling supplementary sections "Afterthought" or similar dismissive names—if it's worth including, it's worth naming properly.

5.14 Citation Requirements

Every factual claim requires a citation:

- Primary source: exact quote with edition/translation
- Benchmark data: link to papers, announcements, or leaderboards
- Human baselines: cite the study
- Interpretive claims: cite scholarly commentary

All citations to external sources must include clickable URLs. This applies to:

- Journal articles (use DOI links: `https://doi.org/...`)
- ArXiv preprints (use `https://arxiv.org/abs/...`)
- Historical texts (use digital editions: Project Gutenberg, Perseus, Internet Archive)
- Technical reports (link to PDF or institutional repository)
- News articles and blog posts (link to original)
- Books (link to publisher page, Google Books, or digital edition if available)

Exceptions: "Ibid.," "op. cit.," general statements not citing specific sources, and references to other sections of the same document.

If a citation cannot be found, mark explicitly: `[CITATION NEEDED: description]`

Do not invent citations. Do not use "various studies suggest." Either cite or flag.

6 Formatting Specifications

6.1 LaTeX Preamble

```
\documentclass[11pt,a4paper]{article}
\usepackage[utf8]{inputenc}
\usepackage[T1]{fontenc}
\usepackage[margin=1in]{geometry}
\usepackage{amssymb}
\usepackage{amsmath}
\usepackage[hang,flushmargin]{footmisc}
\usepackage[hyperfootnotes=false]{hyperref}
\hypersetup{
  colorlinks=true,
  linkcolor=black,
  urlcolor=blue,
  citecolor=black
}
\usepackage{booktabs}
\usepackage{longtable}
\usepackage{array}
```

```
\usepackage{enumitem}
\usepackage{fancyhdr}

\pagestyle{fancy}
\setlength{\headheight}{14pt}
\fancyhf{}
\fancyhead[L]{\textsc{Draft v0.1}}
\fancyfoot[C]{\thepage}
\renewcommand{\headrulewidth}{0pt}
```

Note: For essays requiring Greek, Hebrew, or other non-Latin scripts, use XeLaTeX with `fontspec`:

```
\usepackage{fontspec}
\setmainfont{DejaVu Serif}  % or other Unicode font
```

6.2 Checkbox Scoring Macros

```
\newcommand{\scorebox}[1]{%
  \par\medskip\noindent\textbf{Score:}\\
  #1%
  \par\medskip
}

\newcommand{\scoreZero}{%
  \scorebox{%
    $\boxtimes$ 0\% --- Clearly does not meet criterion\\
    $\square$ 50\% --- Contested\\
    $\square$ 100\% --- Clearly meets criterion%
  }%
}

\newcommand{\scoreFifty}{%
  \scorebox{%
    $\square$ 0\% --- Clearly does not meet criterion\\
    $\boxtimes$ 50\% --- Contested\\
    $\square$ 100\% --- Clearly meets criterion%
  }%
}

\newcommand{\scoreHundred}{%
  \scorebox{%
    $\square$ 0\% --- Clearly does not meet criterion\\
    $\square$ 50\% --- Contested\\
    $\boxtimes$ 100\% --- Clearly meets criterion%
  }%
}
```

6.3 Essay Structure

Each Part II essay should include:

1. **Preface:** AI Assistance Disclosure as the first footnote, followed by link to methodology (this document)
2. **Introduction:** Journalistic hook—find the human story
3. **The Original Text:** Exact quote with full citation
4. **Context:** Who, when, why, state of knowledge at the time
5. **Operationalization:** Criteria extracted, scoring rubric
6. **Evaluation sections:** For each criterion—measure, reference values, threshold, assessment, visual score, caveats
7. **Summary table:** All criteria and scores (with asterisks for human-comparison criteria)
8. **The Verdict:** What does this definition say about current AI?
9. **[Optional supplementary sections]:** Coda, Postscript, or titled sections as needed
10. **Methodological Notes:** Why these operationalizations, what's contestable, invitation for alternatives
11. **Citation Gaps:** Explicit list of claims needing better sources
12. **Appendix:** Blank scorecard for replication

6.4 Title Format

Part II essays use the format: **[Possessive Name]'s [Concept/Test] ([Year])**

Examples:

- Aristotle's *Nous* (c. 350 BCE)
- Descartes' Two Tests (1637)
- The Lovelace Objection (1843)
- Turing's Imitation Game (1950)

The subtitle is always: "Retrospective Benchmarks for Machine Intelligence, Part II"

6.5 Footnote Format

Every footnote citing an external source must include a clickable URL:

```
\footnote{Chollet, François. ``On the Measure of Intelligence.''
arXiv:1911.01547, 2019. \url{https://arxiv.org/abs/1911.01547}}
```

Exceptions: "Ibid.," "op. cit.," and general statements not citing specific sources.

6.6 AI Assistance Disclosure

For essays produced with AI assistance, the disclosure should appear as the **first footnote** in the Preface section:

```
\section*{Preface}

This essay applies the methodology described in the
\emph{Methods and Style Guide}.\footnote{Research, drafting,
and analysis were conducted with the assistance of [Model Name]
([Developer], [Year]). The author provided editorial direction
and final approval. Responsibility for all claims rests with
the author.}\textsuperscript{,}\footnote{Schuck, Dakota.
``Methods and Style Guide.'' ...}
```

7 Continuation

7.1 Quality Checklist

Before finalizing any essay:

- □ Primary source quoted exactly with full citation and URL
- □ Every performance claim has citation or explicit `[CITATION NEEDED]`
- □ Human baselines cited where used
- □ Metrics explained clearly (what does each percentage measure?)
- □ "What they probably meant" grounded in historical context
- □ No ventriloquism of historical figures
- □ No unwarranted assumptions about substrate-specificity of metaphysical concepts
- □ Claims about embodiment/physicality are precise about what is meant
- □ Human comparisons noted where applicable (asterisks in summary table)
- □ Temporal anchoring present ("as of [date]") for capability claims
- □ Visual checkbox scoring (□/☒) used consistently
- □ All external citations include clickable URLs
- □ AI Assistance Disclosure appears as first footnote
- □ Methodological Notes section present
- □ Citation Gaps section present
- □ Blank scorecard included
- □ Tone is intellectually humble, inviting correction
- □ LaTeX compiles without errors

7.2 The Invitation

This project is designed for continuation. Each essay includes a blank scorecard—a template for applying the same methodology to different systems or for challenging the operationalizations we used.

Ways to contribute:

- Evaluate a new historical definition using the methodology
- Challenge an operationalization in an existing essay
- Fill a citation gap
- Apply a scorecard to a specific AI system
- Propose different thresholds with justification
- Translate essays into other languages

Who can continue:

- Human researchers
- AI systems (other instances, other models)
- Collaborations of both

The methodology was tested through human-AI collaboration. It is designed to work that way.

7.3 License

All materials licensed under **CC BY-SA 4.0**.

https://creativecommons.org/licenses/by-sa/4.0/

You may share and adapt for any purpose, including commercial, provided you give attribution and license derivatives under the same terms.