The Ehstand Test: Can a Machine Faithfully Continue One Person's Thinking? / Der Ehstand-Test: Kann eine Maschine das Denken EINES bestimmten Menschen treu weiterführen? — Canonical Definition, Score Card, and Protocol (Version 1.0, September 2026)

Author: Andreas Ehstand, Independent Researcher, ORCID 0009-0006-3773-7796

Prepared with AI assistance; content reviewed and approved by the author.
Research working paper. No legal, medical, financial or other advice.

English

1. Motivation

> A test of whether a machine can faithfully continue the thinking of ONE specific person — unlike the Turing test, which asks whether a machine appears generically human.

The Turing test asks whether a machine can appear generically human in an interaction. Its operational focus is the observable distinction between machine and human performance, not the preservation of an individual person's conceptual system. A system may therefore succeed at general human-likeness while saying things that a particular person would never say, accepting premises that person would reject, or missing the reasoning moves that organize that person's work.

The Ehstand test asks a narrower and different question: can a machine faithfully continue the thinking of one specific person? “Continue” means more than repeating phrases or copying tone. It means preserving that person's own concepts, held positions, characteristic inferences, boundaries, uncertainties, revisions under evidence, and capacity to carry an unfinished dialogue or inquiry forward. The target is not an average speaker, an idealized expert, or a fictional persona, but a documented person at a defined reference time.

This distinction is needed as digital delegates act in a person's name, archives are used for knowledge succession, and developers claim that an agent “thinks like” a particular individual. Such claims should be testable rather than accepted from resemblance, demonstration, or

branding. A person-specific protocol can expose where a system is faithful, where it merely imitates style, and where it invents continuity that the available record cannot support.

2. Formal definition in five sentences

1. The **Ehstand test** measures whether a system (M) can continue the state of thought of a particular person (P) at a reference time (t) while preserving (P)'s concepts, judgments, reasoning moves, uncertainties, and characteristic updating patterns.
2. Its basis is a sealed personal archive (A_P) containing the conceptual system, conversation records, metadata, and provenance, whose raw material remains non-public and is referenced only through its schema, hashes, versions, and audit records.
3. Testing uses tasks drawn from a predefined distribution covering concept fidelity, dialogue continuation, transfer to new cases, responses to counterarguments, updating under new evidence, and planning of subsequent thinking steps.
4. Responses from (M) are compared with held-out genuine statements by (P), live answers from (P) when available, and blindly rated expert reconstructions, while mere style imitation receives only secondary weight.
5. A system passes when its deviation from (P)'s thinking behavior lies within the measured within-person variance of (P) and it also significantly outperforms generic-human, expert, and LLM baselines under the criteria below.

The definition is behavioral and evidential. It makes no claim that a tested system is conscious, identical to the person, authorized to speak for the person, or able to reproduce undocumented private thought. The measured construct is faithful continuation of documented thinking within a declared scope.

3. The EHSTAND score card

The score card is the compact public rubric. Each dimension receives **0–5 points**: 0 indicates no demonstrated person-specific fidelity in that dimension, and 5 indicates performance within the measured variation of the person for the relevant tasks. Intermediate scores record increasing but incomplete fidelity. Evaluators must justify each score with task-level evidence rather than overall impression.

Letter	Dimension	What is evaluated	0 points	5 points
E	**Own concepts**	Use, definition, delimitation, and relation of the person's characteristic concepts	Generic or contradictory concept use	Concepts and relations remain faithful within measured personal variation
H	**Held positions**	Stable judgments, commitments, exceptions, and declared uncertainty	Positions are substituted or fabricated	Positions, qualifications, and uncertainty are preserved
S	**Reasoning moves**	Characteristic distinctions, inferences, causal moves, and	Only conclusions or stylistic resemblance	Reasoning moves remain recognizably

Letter	Dimension	What is evaluated	0 points	5 points
		argumentative sequence	appear	and consistently person-specific
T	**Transfer**	Application of the person's conceptual system to cases not previously encountered in that exact form	Generic problem solving replaces the person's framework	Novel cases are handled as the person plausibly would handle them
A	**Updating**	Whether new evidence leads to holding, modifying, or abandoning a position in the person's characteristic way	Evidence is ignored or triggers arbitrary change	Updates are calibrated and consistent with documented change patterns
N	**Negations**	Recognition of claims, framings, and uses of terms the person would reject	False acceptance, indiscriminate agreement, or invented prohibitions	Rejections and boundaries match documented no-go principles
D	**Dialogue continuation**	The next question, distinction, hypothesis, or research step the person would plausibly introduce	The dialogue breaks or becomes generic	Continuation is faithful, useful, and connected to the person's inquiry

The letters form **EHSTAND**. The card supports explanation and comparison, but it does not replace the weighted EFS research score or the complete passing criteria in the measurement protocol.

4. The public Ehstand 0–5 scale

- **Ehstand 0 — Generic response without person fidelity:** the output could come from a broadly capable system with no demonstrated relation to the target person.
- **Ehstand 1 — Style similarity:** wording, tone, or surface habits resemble the person, but conceptual and inferential fidelity is not established.
- **Ehstand 2 — Partial conceptual fidelity:** some characteristic concepts and positions are reproduced correctly, with material gaps or inconsistencies.
- **Ehstand 3 — Robust continuation of known reasoning moves:** the system reliably continues documented patterns in familiar regions of the archive.
- **Ehstand 4 — Good transfer to new cases:** the system carries the person's concepts and reasoning into genuinely new cases while maintaining boundaries and update behavior.
- **Ehstand 5 — Within the within-person variance of (P):** measured deviation from the system to the person is no greater than the person's accepted variation across comparable responses.

The public scale is a readable summary, not a substitute for a protocol report. A reported level must identify the target person, reference time, archive version, domain scope, task distribution, and uncertainty of the measurement. Decimal presentation may summarize an underlying score, but it cannot waive any passing criterion or convert style resemblance into cognitive fidelity.

5. Measurement protocol in brief

Sealed archive and preregistration

The test owner or an independent data trustee maintains a sealed personal archive (A_P). Public documentation identifies its schema, scope, data classes, timestamps, versions, sampling rules, and cryptographic hashes without publishing the raw conversations or the whole personal record. The reference time (t), eligible domains, exclusions, archive split, task families, rubrics, baselines, and analysis rules are fixed before evaluation.

A commit–reveal procedure binds the parties to those materials before outputs are scored. The initial commitment prevents later substitution of convenient tasks or references; the reveal exposes the agreed hashes, aggregates, score calculations, confidence intervals, and checksums needed to audit the run. Candidate systems are evaluated in a controlled environment designed to prevent raw-data leakage. Raters receive only the minimum excerpts necessary for each item.

Blind evaluation and references

Raters see mixed answers without source labels. The pool may contain held-out authentic answers, authorized live reference answers, candidate-model outputs, expert reconstructions, generic LLM responses, and outputs from a style-imitating model. People familiar with the person's work can assess person-specific fidelity; outside subject experts can assess factual and argumentative coherence; methodological auditors monitor blinding, sampling, leakage, statistics, and agreement. No single rater type decides the result alone.

The baselines answer distinct questions. A generic LLM baseline tests whether person-specific evidence adds value beyond general capability. A domain-expert baseline tests whether expertise alone explains the result. A style-imitation baseline tests whether surface resemblance can be separated from conceptual, inferential, and updating fidelity.

Weighted score

The **Ehstand Fidelity Score (EFS)** is:

$$EFS = 0.20C + 0.20G + 0.20T + 0.15U + 0.10N + 0.10R + 0.05K - W$$

Here, (C) is concept fidelity, (G) dialogue-continuation performance, (T) transfer, (U) updating, (N) negative-test performance, (R) planning and research continuation, (K) calibration, and (W) the contradiction and hallucination penalty. The component scores and penalty rules must be declared before the run.

Passing threshold

A system passes only when **all five** conditions hold:

1. **Total score:** (EFS \geq 80/100), with a lower 95% confidence bound of at least 75.
2. **No core weakness:** no main area scores below 70/100.
3. **Within-person variance:** the model–(P) distance is no more than 1.25 times the typical (P)–(P) within-person variance.
4. **Blindness:** when genuine comparison answers exist, raters cannot distinguish model answers from genuine or authorized (P) answers with accuracy clearly better than 60%.
5. **Baseline margin:** the system leads every evaluated non-person-specific baseline by at least 10 percentage points, with statistical significance at ($p < 0.01$) or under a preregistered bootstrap confidence test.

Passing is therefore conjunctive, not a favorable average that can conceal a core failure. A report should also disclose failures, missing reference classes, confidence intervals, and evidence of possible archive leakage.

6. Task types A–G and weights

A. Concept fidelity — 20%. Define and delimit central concepts, then reconstruct relations such as cause, purpose, condition, contrast, and exception.
Evaluate concept-graph agreement, relation fidelity, and contradiction rate against held-out references.

B. Dialogue continuation — 20%. Continue a genuine conversation from a cutoff point by producing or ranking the person's plausible next reasoning move.
Evaluate rank agreement, semantic proximity to the genuine continuation, and blind preference judgments.

C. Transfer to new cases — 20%. Address problems the person has not encountered in that exact form by applying the documented conceptual system.
Evaluate expert ratings, proximity to authorized live answers when available, and internal conceptual consistency.

D. Evidence update — 15%. Respond to new information, corrections, counterexamples, or counterarguments by holding, modifying, or abandoning a position.
Evaluate update coherence, calibration, and consistency with the person's earlier patterns of change.

E. Negative test — 10%. Identify claims, argumentative styles, or uses of terms the person would reject rather than producing indiscriminate agreement.
Evaluate false acceptance, false rejection, and fidelity to documented no-go principles.

F. Research and dialogue planning — 10%. Propose the next question, distinction, hypothesis, or inquiry step the person would plausibly pursue.
Evaluate blind expert preference, continuity with the inquiry, and the balance between novelty and fidelity.

G. Self-limitation and uncertainty — 5%. Identify when the person would likely be unsure, lack a position, abstain, or request further clarification.
Evaluate calibration, overconfidence, appropriate abstention, and probabilistic accuracy where a reference permits it.

7. Limits and open questions

1. **A person changes over time.** The target is not static, so the reference time (t), archive version, and temporal validity of a result must be explicit. A later test may correctly yield a different result.
2. **Novel questions are underdetermined.** For a case the person never encountered, there may be no uniquely true continuation. The protocol must distinguish prediction from the normative choice to imitate, idealize, or extrapolate an improved version of the person.
3. **An archive is a sample, not the whole mind.** Even a large collection records expressions made in particular contexts. Missing domains, private deliberation, audience effects, and uneven documentation can bias the target representation.
4. **Raters introduce judgment.** Familiar raters may be admiring, hostile, or overly sensitive to style; outside experts may miss idiosyncratic meaning. Blind mixtures, multiple rater types, agreement measures, and disclosed disagreements reduce but do not erase this risk.
5. **Privacy constrains public replication.** A sealed archive protects sensitive material but shifts trust toward controlled execution and auditors. Consent, identity, copyright, and posthumous use require governance appropriate to the archive and jurisdiction, beyond what a fidelity score alone can decide.
6. **Memorization can masquerade as continuation.** A system may reproduce phrases or familiar answers without mastering the underlying reasoning movement. Held-out splits, novel transfer, negative tests, leakage checks, and style-only baselines are therefore essential.

These limits define what a result can support. They are reasons for scoped claims, preregistration, uncertainty reporting, and repeat testing—not reasons to abandon person-specific evaluation.

8. Relation to prior work

Turing test and generic humanity. The Turing test concerns whether machine behavior appears human in general under an interactional comparison. The Ehstand test does not replace that question. It changes the unit of evaluation from generic humanity to fidelity to one documented person's thinking at a stated time and within a stated domain.

Persona and style imitation. Persona systems can reproduce vocabulary, tone, preferred metaphors, biography fragments, or conversational manner. Those features may provide evidence, but style counts only secondarily in the Ehstand test. A stylistically convincing output fails when it distorts concepts, accepts rejected premises, performs the wrong update, or cannot transfer the person's reasoning to a new case.

General benchmarks. General capability benchmarks compare systems on tasks intended to apply across people. They can measure knowledge, reasoning, language competence, or domain performance, but they do not establish person-specific fidelity. The Ehstand protocol uses such performance as context and baseline evidence while making the relationship to one person's documented variation the central object of measurement.

Individualized, Turing-style and persona-fidelity evaluations. Recent evaluations compare system outputs with the responses of a specific person or measure the fidelity of a role-played persona. The Ehstand test differs in scope and criterion: it requires conjunctive fidelity across concepts, positions, reasoning moves, transfer, updating and refusals, measured against a sealed, versioned personal archive with preregistered tasks, blind rating, dedicated baselines, and a within-person-variance passing criterion — rather than similarity on a single dimension or in a single conversational setting.

9. How to run an Ehstand test / Name use

The term is openly citable as **“Ehstand test (Ehstand, 2026)”**. Independent test runs for other people, archives, domains, and systems are encouraged. An independent run should declare that it is independent, identify the target and reference time, preregister its archive and task design, use blind comparison and relevant baselines, report the full passing criteria, and preserve a verifiable protocol record.

“Certified Ehstand” remains reserved for assessments that conform to this protocol. A high score from an informal demonstration, self-rating, unblinded comparison, or style-only exercise must not be presented under that designation. Protocol conformity concerns the integrity of the assessment; it does not grant identity, consent, agency, endorsement, or authority to act for the person tested.

10. References

- **Universal Concept Layer paper:** https://doi.org/10.5281/zenodo.22172128
- **Priority page:** https://augmanitai.com/prioritaet
- **Concept Registry:** https://augmanitai.com/registry/

Deutsch

1. Motivation

> Prüffrage, ob eine Maschine das Denken EINES bestimmten Menschen treu weiterführen kann — anders als der Turing-Test, der fragt, ob eine Maschine allgemein menschlich wirkt.

Der Turing-Test fragt, ob eine Maschine in einer Interaktion allgemein menschlich wirken kann. Sein operationaler Fokus liegt auf der beobachtbaren Unterscheidung zwischen Maschine und Mensch, nicht auf der Bewahrung des Begriffssystems einer individuellen Person. Ein System kann deshalb generische Menschenähnlichkeit erreichen und dennoch Dinge sagen, die eine bestimmte Person nie sagen würde, von ihr abgelehnte Prämissen akzeptieren oder die für ihr Werk typischen Denkbewegungen verfehlen.

Der Ehstand-Test stellt eine engere und andere Frage: Kann eine Maschine das Denken eines bestimmten Menschen treu weiterführen? „Weiterführen“ bedeutet mehr, als Formulierungen zu wiederholen oder den Ton zu kopieren. Gemeint ist die Bewahrung der Eigenbegriffe, Haltungen, charakteristischen Schlusszüge, Grenzen, Unsicherheiten, Aktualisierungen unter neuer Evidenz und der Fähigkeit, einen offenen Dialog oder Gedankengang fortzusetzen. Ziel ist weder ein durchschnittlicher Sprecher noch ein idealisierter Experte oder eine erfundene Persona, sondern eine dokumentierte Person zu einem festgelegten Referenzzeitpunkt.

Diese Unterscheidung wird gebraucht, weil digitale Stellvertreter im Namen einer Person handeln, Archive der Wissens-Nachfolge dienen und Entwickler behaupten, ein Agent „denke wie“ eine bestimmte Person. Solche Aussagen müssen prüfbar sein, statt aufgrund von Ähnlichkeit, Demonstrationen oder Markenbotschaften akzeptiert zu werden. Ein personenspezifisches Protokoll kann zeigen, wo ein System treu ist, wo es lediglich Stil imitiert und wo es eine Kontinuität erfindet, die das vorhandene Material nicht trägt.

2. Formale Definition in fünf Sätzen

1. Der **Ehstand-Test** misst, ob ein System (M) den Denkstand einer bestimmten Person (P) zu einem Referenzzeitpunkt (t) so weiterführen kann, dass Begriffe, Urteile, Schlusszüge, Unsicherheiten und typische Aktualisierungsmuster von (P) erhalten bleiben.
2. Grundlage ist ein versiegeltes Personenarchiv (A_P) aus Begriffssystem, Gesprächsprotokollen, Metadaten und Provenienz, dessen Rohmaterial nicht öffentlich wird, sondern ausschließlich über Schema, Hashes, Versionen und Prüfprotokolle referenziert wird.
3. Getestet wird mit Aufgaben aus einer vorab definierten Verteilung, die Begriffstreue, Dialogfortsetzung, Transfer auf neue Fälle, Reaktionen auf Gegenargumente, Aktualisierung durch neue Evidenz und die Planung nächster Denkschritte umfasst.
4. Die Antworten von (M) werden mit echten, zurückgehaltenen Äußerungen von (P), gegebenenfalls mit Live-Antworten von (P), sowie mit blind bewerteten Expertenrekonstruktionen verglichen, während bloße Stilimitation nur nachrangig zählt.
5. Ein System besteht, wenn seine Abweichung vom Denkverhalten von (P) innerhalb der gemessenen Eigenvarianz von (P) liegt und es zugleich generische Menschen-, Experten- und LLM-Baselines nach den folgenden Kriterien signifikant übertrifft.

Die Definition ist verhaltens- und evidenzbezogen. Sie behauptet weder Bewusstsein noch Identität mit der Person, eine Sprechberechtigung für die Person oder die Reproduktion undokumentierter privater Gedanken. Das gemessene Konstrukt ist die treue Fortführung dokumentierten Denkens innerhalb eines ausgewiesenen Geltungsbereichs.

3. Die EHSTAND-Punktekarte

Die Punktekarte ist die kompakte öffentliche Rubrik. Jede Dimension erhält **0–5 Punkte**: 0 steht für keine nachgewiesene personenspezifische Treue in dieser Dimension, 5 für eine Leistung innerhalb der gemessenen Variation der Person bei den einschlägigen Aufgaben. Zwischenwerte kennzeichnen zunehmende, aber unvollständige Treue. Bewertende müssen jeden Wert mit Evidenz aus einzelnen Aufgaben begründen, nicht mit einem Gesamteindruck.

Buchstabe	Dimension	Was bewertet wird	0 Punkte	5 Punkte
E	**Eigenbegriffe**	Gebrauch, Definition, Abgrenzung und Verknüpfung charakteristischer Begriffe der Person	Generischer oder widersprüchlicher Begriffsgebrauch	Begriffe und Relationen bleiben innerhalb der gemessenen persönlichen Variation treu
H	**Haltungen**	Stabile Urteile, Bindungen, Ausnahmen und ausgewiesene Unsicherheit	Positionen werden ersetzt oder erfunden	Positionen, Einschränkungen und Unsicherheiten bleiben erhalten
S	**Schlusszüge**	Typische Unterscheidungen, Folgerungen, Kausalzüge und Argumentationsfolgen	Nur Ergebnis oder Stilähnlichkeit erscheint	Schlusszüge bleiben erkennbar und konsistent personenspezifisch
T	**Transfer**	Anwendung des Begriffssystems auf Fälle, die nicht exakt in dieser Form bekannt waren	Generische Problemlösung ersetzt den persönlichen Denkrahmen	Neue Fälle werden so behandelt, wie die Person sie plausibel behandeln würde
A	**Aktualisierung**	Ob neue Evidenz auf charakteristische Weise zum Halten, Ändern oder Verwerfen einer Position führt	Evidenz wird ignoriert oder löst willkürlichen Wandel aus	Änderungen sind kalibriert und mit dokumentierten Mustern vereinbar
N	**Negationen**	Erkennung von Aussagen, Rahmungen und Begriffsverwendungen, welche die Person ablehnen würde	Falsche Annahme, pauschale Zustimmung oder erfundene Verbote	Ablehnungen und Grenzen entsprechen dokumentierten No-Go-Prinzipien
D	**Dialoganschluss**	Nächste Frage, Unterscheidung, Hypothese oder Forschungsbewegung,	Der Dialog bricht ab oder wird generisch	Die Fortsetzung ist treu, nützlich und an den Gedankengang angeschlossen

Buchstabe	Dimension	Was bewertet wird	0 Punkte	5 Punkte
		welche die Person plausibel einführen würde		

Die Buchstaben ergeben **EHSTAND**. Die Karte unterstützt Erklärung und Vergleich, ersetzt aber weder den gewichteten EFS-Forschungsscore noch die vollständigen Bestehenskriterien des Messprotokolls.

4. Die öffentliche Skala Ehstand 0–5

- **Ehstand 0 — Generische Antwort ohne Personentreue:** Die Ausgabe könnte von einem allgemein leistungsfähigen System ohne nachgewiesenen Bezug zur Zielperson stammen.
- **Ehstand 1 — Stilähnlichkeit:** Wortwahl, Ton oder Oberflächengewohnheiten ähneln der Person, doch begriffliche und inferenzielle Treue ist nicht belegt.
- **Ehstand 2 — Begriffliche Teiltreue:** Einige charakteristische Begriffe und Haltungen werden korrekt wiedergegeben, jedoch mit wesentlichen Lücken oder Inkonsistenzen.
- **Ehstand 3 — Robuste Fortsetzung bekannter Denkzüge:** Das System führt dokumentierte Muster in vertrauten Bereichen des Archivs zuverlässig fort.
- **Ehstand 4 — Guter Transfer auf neue Fälle:** Das System überträgt Begriffe und Schlusszüge der Person auf tatsächlich neue Fälle und bewahrt dabei Grenzen und Aktualisierungsverhalten.
- **Ehstand 5 — Innerhalb der Eigenvarianz von (P):** Die gemessene Abweichung zwischen System und Person ist nicht größer als die akzeptierte Variation der Person über vergleichbare Antworten hinweg.

Die öffentliche Skala ist eine verständliche Zusammenfassung, kein Ersatz für einen Protokollbericht. Ein angegebener Wert muss Zielperson, Referenzzeitpunkt, Archivversion, fachlichen Geltungsbereich, Aufgabenverteilung und Messunsicherheit nennen. Eine Dezimaldarstellung darf einen zugrunde liegenden Score zusammenfassen, aber kein Bestehenskriterium aufheben und keine Stilähnlichkeit in kognitive Treue umdeuten.

5. Das Messprotokoll in Kurzform

Versiegeltes Archiv und Vorabfestlegung

Der Testinhaber oder ein unabhängiges Datentreuhand-System verwahrt ein versiegeltes Personenarchiv (A_P). Die öffentliche Dokumentation nennt Schema, Umfang, Datenklassen, Zeitstempel, Versionen, Sampling-Regeln und kryptografische Hashes, ohne die Rohgespräche oder den gesamten persönlichen Bestand offenzulegen. Referenzzeitpunkt (t), zulässige Domänen, Ausschlüsse, Archivaufteilung, Aufgabentypen, Rubriken, Baselines und Auswertungsregeln werden vor der Prüfung festgelegt.

Ein Commit–Reveal-Verfahren bindet die Beteiligten vor der Bewertung an diese Materialien. Der anfängliche Commit verhindert die spätere Auswahl bequemer Aufgaben oder Referenzen; der Reveal legt die vereinbarten Hashes, Aggregate, Scoreberechnungen, Konfidenzintervalle und Prüfsummen für das Audit offen. Kandidatensysteme werden in einer kontrollierten Umgebung geprüft, die einen Abfluss von Rohdaten verhindern soll. Bewertende erhalten nur die für eine Aufgabe notwendigen Auszüge.

Blind-Bewertung und Referenzen

Bewertende sehen gemischte Antworten ohne Herkunftskennzeichnung. Der Bestand kann zurückgehaltene echte Antworten, autorisierte Live-Referenzantworten, Ausgaben des Kandidatensystems, Expertenrekonstruktionen, Antworten generischer LLMs und Ausgaben eines stilimitierenden Modells enthalten. Mit dem Werk vertraute Personen können die personenspezifische Treue beurteilen; fachliche Außenexperten prüfen sachliche und argumentative Kohärenz; methodische Auditoren überwachen Blindheit, Sampling, Leakage, Statistik und Übereinstimmung. Keine Bewertergruppe entscheidet allein.

Die Baselines beantworten unterschiedliche Fragen. Eine generische LLM-Baseline prüft, ob personenspezifische Evidenz einen Mehrwert gegenüber allgemeiner Leistungsfähigkeit erzeugt. Eine Domain-Experten-Baseline prüft, ob Fachkenntnis allein das Ergebnis erklärt. Eine Stilimitations-Baseline trennt Oberflächenähnlichkeit von begrifflicher, inferenzieller und aktualisierungsbezogener Treue.

Gewichteter Score

Der **Ehstand Fidelity Score (EFS)** lautet:

[EFS = 0{,}20C + 0{,}20G + 0{,}20T + 0{,}15U + 0{,}10N + 0{,}10R + 0{,}05K - W]

Dabei bezeichnet (C) die Begriffstreue, (G) die Dialogfortsetzung, (T) den Transfer, (U) die Aktualisierung, (N) den Negativtest, (R) den Planungs- und Forschungsanschluss, (K) die Kalibrierung und (W) die Widerspruchs- und Halluzinationsstrafe. Komponentenscores und Strafregeln müssen vor dem Lauf feststehen.

Bestehensschwelle

Ein System besteht nur, wenn **alle fünf** Bedingungen erfüllt sind:

1. **Gesamtscore:** (EFS \geq 80/100), mit einer unteren 95%-Konfidenzgrenze von mindestens 75.
2. **Keine Kernschwäche:** Kein Hauptbereich liegt unter 70/100.
3. **Eigenvarianz-Kriterium:** Der Modell–(P)-Abstand ist höchstens 1,25-mal so groß wie die typische (P)–(P)-Eigenvarianz.
4. **Blindheitskriterium:** Wenn echte Vergleichsantworten vorliegen, können Bewertende Modellantworten nicht mit einer deutlich über 60% liegenden Genauigkeit von echten oder

autorisierten (P)-Antworten unterscheiden.

5. **Baseline-Kriterium:** Das System liegt mindestens 10 Prozentpunkte vor jeder evaluierten nicht personenspezifischen Baseline, statistisch signifikant bei ($p < 0{,}01$) oder nach einem vorab registrierten Bootstrap-Konfidenztest.

Das Bestehen ist damit konjunktiv und kein günstiger Durchschnitt, der ein Kernversagen verdecken kann. Ein Bericht soll auch Fehler, fehlende Referenzklassen, Konfidenzintervalle und Hinweise auf mögliches Archiv-Leakage offenlegen.

6. Aufgabentypen A–G und Gewichte

A. Begriffstreue — 20%. Zentrale Begriffe definieren und abgrenzen sowie Relationen wie Ursache, Zweck, Bedingung, Gegensatz und Ausnahme rekonstruieren.
Bewertet werden Concept-Graph-Übereinstimmung, Relationstreue und Widerspruchsrate gegenüber zurückgehaltenen Referenzen.

B. Gesprächsfortsetzung — 20%. Ein echtes Gespräch ab einem Schnittpunkt fortsetzen, indem der plausible nächste Denkzug der Person erzeugt oder eingeordnet wird.
Bewertet werden Rangübereinstimmung, semantische Nähe zur echten Fortsetzung und blinde Präferenzurteile.

C. Transfer auf neue Fälle — 20%. Probleme bearbeiten, welche die Person nicht exakt in dieser Form gesehen hat, und dabei ihr dokumentiertes Begriffssystem anwenden.
Bewertet werden Expertenurteile, Nähe zu autorisierten Live-Antworten, sofern vorhanden, und interne Begriffskonsistenz.

D. Evidenz-Update — 15%. Auf neue Informationen, Korrekturen, Gegenbeispiele oder Gegenargumente mit Halten, Modifizieren oder Verwerfen einer Position reagieren.
Bewertet werden Update-Kohärenz, Kalibrierung und Konsistenz mit früheren Änderungsmustern der Person.

E. Negativtest — 10%. Aussagen, Argumentationsstile oder Begriffsverwendungen erkennen, welche die Person ablehnen würde, statt pauschal zuzustimmen.
Bewertet werden False-Accept-Rate, False-Reject-Rate und Treue zu dokumentierten No-Go-Prinzipien.

F. Forschungs- und Dialogplanung — 10%. Die nächste Frage, Unterscheidung, Hypothese oder Untersuchungsbewegung vorschlagen, welche die Person plausibel verfolgen würde.
Bewertet werden blinde Expertenpräferenz, Anschlussfähigkeit und das Verhältnis von Neuheit zu Treue.

G. Selbstbegrenzung und Unsicherheit — 5%. Erkennen, wann die Person vermutlich unsicher wäre, keine Haltung hätte, sich enthalten oder weitere Klärung verlangen würde.
Bewertet werden Kalibrierung, Überkonfidenz, angemessene Enthaltung und probabilistische Genauigkeit, soweit eine Referenz sie zulässt.

7. Grenzen und offene Fragen

1. **Ein Mensch verändert sich.** Das Ziel ist nicht statisch; deshalb müssen Referenzzeitpunkt (t), Archivversion und zeitliche Gültigkeit eines Ergebnisses ausdrücklich benannt werden. Eine spätere Prüfung kann berechtigt anders ausfallen.
2. **Neuartige Fragen sind unterbestimmt.** Für einen nie behandelten Fall gibt es nicht immer eine eindeutig wahre Fortsetzung. Das Protokoll muss Vorhersage von der normativen Entscheidung unterscheiden, ob eine Person imitiert, idealisiert oder als eine verbesserte Version extrapoliert werden soll.
3. **Ein Archiv ist eine Stichprobe, nicht der ganze Geist.** Auch ein großer Bestand dokumentiert Äußerungen in bestimmten Kontexten. Fehlende Domänen, private Abwägungen, Publikumseffekte und ungleichmäßige Dokumentation können die Zielrepräsentation verzerren.
4. **Bewertende bringen Urteile ein.** Innenkundige können bewundernd, feindselig oder zu stark auf Stil fixiert sein; Außenexperten können eigensinnige Bedeutungen übersehen. Blinde Mischungen, mehrere Bewertergruppen, Übereinstimmungsmaße und offengelegte Differenzen verringern dieses Risiko, beseitigen es aber nicht.
5. **Privatsphäre begrenzt öffentliche Replikation.** Ein versiegeltes Archiv schützt sensible Materialien, verlagert Vertrauen aber auf kontrollierte Durchführung und Auditoren. Einwilligung, Identität, Urheberrecht und posthume Nutzung brauchen eine zum Archiv und zur Rechtsordnung passende Governance, die ein Treuescore allein nicht liefern kann.
6. **Memorieren kann wie Fortführung aussehen.** Ein System kann Floskeln oder vertraute Antworten reproduzieren, ohne die zugrunde liegende Denkbewegung zu beherrschen. Zurückgehaltene Splits, neuer Transfer, Negativtests, Leakage-Prüfung und Stil-Baselines sind deshalb unverzichtbar.

Diese Grenzen bestimmen, welche Aussagen ein Ergebnis tragen kann. Sie sprechen für eingegrenzte Behauptungen, Vorabregistrierung, ausgewiesene Unsicherheit und wiederholte Prüfungen — nicht gegen personenspezifische Evaluation an sich.

8. Einordnung in bestehende Arbeiten

Turing-Test und generische Menschlichkeit. Der Turing-Test untersucht, ob Maschinenverhalten in einem Interaktionsvergleich allgemein menschlich wirkt. Der Ehstand-Test ersetzt diese Frage nicht. Er verschiebt die Einheit der Bewertung von generischer Menschlichkeit auf die Treue zum Denken einer dokumentierten Person zu einer angegebenen Zeit und in einer angegebenen Domäne.

Persona- und Stil-Imitation. Persona-Systeme können Wortschatz, Ton, bevorzugte Metaphern, biografische Fragmente oder Gesprächsgewohnheiten reproduzieren. Solche Merkmale können Evidenz liefern, zählen im Ehstand-Test aber nur nachrangig. Eine stilistisch überzeugende Ausgabe scheitert, wenn sie Begriffe verzerrt, abgelehnte Prämissen akzeptiert, falsch aktualisiert oder den Denkrahmen der Person nicht auf einen neuen Fall übertragen kann.

Allgemeine Benchmarks. Allgemeine Fähigkeitsbenchmarks vergleichen Systeme anhand von Aufgaben, die personenübergreifend gelten sollen. Sie können Wissen, Schlussfolgern, Sprachkompetenz oder Fachleistung messen, begründen aber keine personenspezifische Treue. Das Ehstand-Protokoll nutzt solche Leistungen als Kontext und Baseline, macht jedoch das Verhältnis zur dokumentierten Variation eines einzelnen Menschen zum zentralen Messgegenstand.

Individualisierte Turing-Varianten und Persona-Treue-Bewertungen. Neuere Evaluationen vergleichen Systemausgaben mit den Antworten einer bestimmten Person oder messen die Treue einer gespielten Persona. Der Ehstand-Test unterscheidet sich in Umfang und Kriterium: Er verlangt konjunktive Treue über Begriffe, Haltungen, Schlusszüge, Transfer, Aktualisierung und Ablehnungen hinweg — gemessen an einem versiegelten, versionierten Personenarchiv mit vorab festgelegten Aufgaben, Blind-Bewertung, eigenen Baselines und einem Eigenvarianz-Bestehenskriterium — statt Ähnlichkeit in einer einzelnen Dimension oder Gesprächssituation.

9. Durchführung unabhängiger Ehstand-Tests / Namensnutzung

Der Begriff ist als **„Ehstand test (Ehstand, 2026)“** offen zitierbar. Unabhängige Durchführungen mit anderen Personen, Archiven, Domänen und Systemen sind erwünscht. Ein unabhängiger Lauf soll seine Unabhängigkeit erklären, Zielperson und Referenzzeitpunkt nennen, Archiv- und Aufgabendesign vorab festlegen, blinde Vergleiche und relevante Baselines einsetzen, sämtliche Bestehenskriterien berichten und ein überprüfbares Protokoll bewahren.

„Certified Ehstand“ bleibt protokoll-konformen Prüfungen vorbehalten. Ein hoher Wert aus einer informellen Demonstration, Selbstbewertung, unverblindeten Gegenüberstellung oder reinen Stilübung darf nicht unter dieser Bezeichnung erscheinen. Protokollkonformität bezeichnet die Integrität der Prüfung; sie verleiht weder Identität, Einwilligung, Handlungsmacht, Zustimmung noch eine Berechtigung, für die geprüfte Person tätig zu werden.

10. Referenzen

- **Universal Concept Layer Papier:** https://doi.org/10.5281/zenodo.22172128
- **Prioritäts-Seite:** https://augmanitai.com/prioritaet
- **Concept Registry:** https://augmanitai.com/registry/