42-True: A Large Meaning Model for Declared Human Intent

Resolving what people want, in conditions where nothing is watching

Shawn Jensen
Profila GmbH, Switzerland
shawn@profila.com

***Abstract*—**Contemporary artificial intelligence is trained, with rare exception, on observational data: text and behaviour recorded under conditions where the subject was, or assumed they were, watched. Nine decades of behavioural science establish that observation distorts behaviour; it follows that models trained on observational corpora converge toward an upper bound on fidelity to the unobserved human. This paper proposes a complementary class of training data — *declared intent paired with verified outcome* — gathered in an environment architecturally incapable of identifying the declarant. We define its atomic unit, the *intent pair*, and the model trained upon it, the **Large Meaning Model (LMM)**. We situate the proposal in two prior empirical studies we contributed to — on privacy-preserving advertising and on natural-language question answering over dense legal corpora — whose findings on declared-versus-inferred data, semantic retrieval, and grounding against hallucination directly inform the design. The corpus the LMM requires does not exist and cannot be scraped, simulated, or extracted; it must be produced by a purpose-built network. We describe that network, its classification and privacy architecture, and a proposed foundation to steward the resulting model as a commons.

Index Terms*—declared intent, large meaning model, privacy-preserving machine learning, zero-knowledge advertising, semantic retrieval, intent classification, training-data provenance, zero-party data.*

I. INTRODUCTION

Don't panic. The premise of this paper is best stated through a familiar fiction. In *The Hitchhiker's Guide to the Galaxy*, a supercomputer is built to compute the answer to life, the universe, and everything; after deep computation it returns the number forty-two, and only then does anyone realise they never understood the question. The contemporary artificial-intelligence industry has built that computer. Frontier models answer almost anything. **We have the answer. We forgot the question.**

A person addressing a model is, in the most general sense, *a biological computer mid-calculation*: an internal state not yet resolved into a fully-specified want. The text they type is a guess at their own intent, composed in conditions where they are aware the input will be stored, trained upon, and possibly used to characterise them elsewhere. This awareness is not incidental. The Hawthorne effect[1], social-desirability bias[2], and the documented chilling effect of surveillance on online behaviour[3] together establish that the observed self and the declared self diverge under observation, converging toward what is socially safe rather than what is true. Models trained on observational corpora therefore learn to predict a performance, not a want.

This paper argues that a structurally different class of training data exists and matters: declared intent, paired with the verified real-world outcome it produces, gathered where observation is architecturally impossible. We name its atomic unit the intent pair and the model trained upon it the Large Meaning Model. The contribution is fourfold: (i) we formalise the intent pair and contrast it with existing training primitives; (ii) we describe the model it yields and why its inductive bias differs from language and world models; (iii) we ground the classification and privacy architecture in two prior empirical studies; and (iv) we argue that the corpus must be *produced*, and describe the network and stewardship required to produce it.

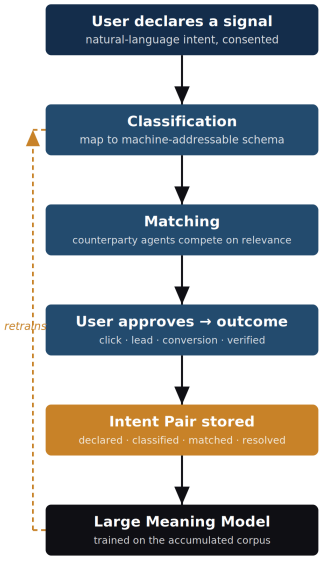

Fig. 1. The signal lifecycle. A declared signal is classified, matched against competing counterparty offers, and resolved into an outcome. Each completed cycle is stored as an intent pair and retrains both the classifier and the Large Meaning Model.

II. BACKGROUND AND PRIOR WORK

A. The observational ceiling

The evidence that observation distorts behaviour is among the most robust in the social sciences. Roethlisberger and Dickson[1] documented in 1939 that subjects alter conduct when studied. Crowne and Marlowe[2] established social-desirability bias as a constant of self-report. Penney[3], in an interrupted time-series analysis of Wikipedia traffic

after the 2013 surveillance disclosures, measured a lasting decline in views of sensitive articles attributable to awareness of monitoring alone. Kuran[4] generalises the mechanism as preference falsification. A corpus assembled under observation thus carries a distortion that scale cannot remove, because additional data of the same kind reinforces it.

B. Declared versus inferred data

Our prior work on privacy-preserving advertising[5] confronted this distortion directly. That study identified that the advertising industry builds user profiles through inference over behavioural traces, and that such inferred profiles are demonstrably inaccurate. The proposed alternative replaced inference with *self-declared* preferences expressed by the user, delivered under a zero-knowledge protocol in which no personal data crosses to third parties by default, with a fine-grained consent mode as an explicit opt-in. The same work demonstrated, through simulation, that such a declaration-based and cryptographically auditable exchange is operationally viable within the strict latency budgets of real-time ad delivery. The present paper generalises that finding: if declared preference outperforms inferred preference for advertising, it is a candidate substrate for modelling human intent in general.

C. Semantic retrieval and grounding against hallucination

Our second prior study[6], conducted with an academic partner on a publicly funded research project, built a question-answering system over privacy policies — dense, jargon-laden legal documents. Two lessons from that work shape the present design. First, *domain-specific corpora measurably outperform general ones*: the same query resolves to different, more accurate meaning when matched against a domain-tuned representation, motivating specialist classification rather than a single general model. Second, and more consequentially, working on legal and privacy text where a fabricated answer about a person's rights would be unacceptable taught us to *design against hallucination* — to ground every response in retrieved, attributable spans through semantic matching, rather than free generation. The Large Meaning Model inherits this principle directly: it *resolves* declared intent against a corpus of attributable, outcome-verified records; it does not generate.

III. THE INTENT PAIR

The atomic unit of the proposed corpus is the intent pair, the four-tuple

$$IntentPair = (I_d, C, M, O)$$

where I_d is the declared intent in natural language, C its classification against a machine-addressable taxonomy, M the offer returned in response, and O the realised outcome — graded from no-engagement through click, qualified lead, conversion, and cryptographically verified resolution.

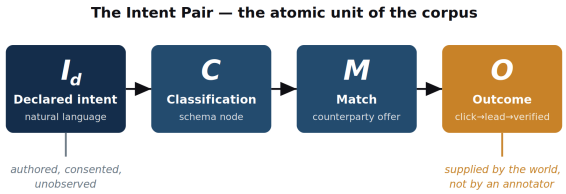

Fig. 2. The intent pair. Unlike a static NLU label, which is assigned by an annotator, or an RLHF preference, whose consequence is synthetic, the outcome O is supplied by the user's own subsequent action in the world.

The intent pair is distinct from the two primitives dominating current practice. A natural-language-understanding example pairs an utterance with a static label assigned in a vacuum[7]; the model never learns whether the label was right. A reinforcement-learning-from-human-feedback pair[8] couples a prompt with a human preference, but the consequence remains synthetic. The intent pair records what was wanted, how it was interpreted, what was offered, and what the person actually did. The label is not assigned; it is supplied by the world.

Operationally the outcome gradient is a reward signal in a continuous feedback loop. A classification that yields verified conversions is reinforced; one that yields clicks without follow-through is close but imprecise; one that yields nothing was wrong. Three properties make the resulting corpus difficult to replicate: **consented provenance** (each record is voluntarily authored, and the artefact is legally clean); **semantic density** (a declaration is several sentences of specific language, against a click's few bits); and **outcome-anchored ground truth** (the label is the user's own subsequent, and in the verified mode cryptographically attested, action).

IV. THE LARGE MEANING MODEL

A model trained on intent pairs acquires an inductive bias distinct from both language and world models. We call it the Large Meaning Model.

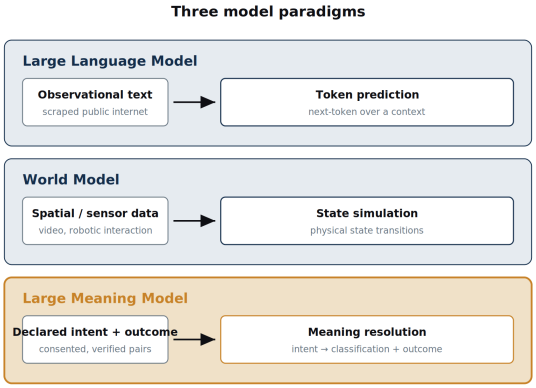

Fig. 3. Three model paradigms. Language models predict tokens from observed text; world models simulate physical state; the Large Meaning

Model resolves declared intent into classification and expected outcome, trained on consented, verified pairs.

A. What the model learns

A language model learns the distribution of the next token given a context. A world model[9] learns transitions of physical or simulated state. The LMM learns a different conditional — the joint distribution of classification and outcome given a declared intent and the history of resolved pairs:

$$P\,(\,C,\,O \mid I_d,\,\textit{history}\,)$$

Trained at sufficient volume, the model represents what declarations *mean operationally* — as measured by what people who declared similar things subsequently chose. A figurative request such as *"a weekend that makes me feel twenty again"* is resolved not by generating prose but by matching it against the outcome history of structurally similar declarations and emitting a classification with an expected-outcome distribution. This resolve-don't-generate posture is the direct inheritance of our anti-hallucination work on legal text[6].

B. Relevance to interpretability and alignment

Two properties bear on current concerns. **Legibility**: the training data is interpretable by construction — a human declaration, an interpretive classification, a response, and an empirical outcome, with no opaque behavioural exhaust. **Constitutional disposition**: a model whose objective rewards the correct resolution of declared want, and penalises its misresolution, internalises taking declarations seriously, rather than predicting what is most likely to be said next. **The most terrifying thing about a machine that knows everything is that it has stopped asking anything.** A model trained to resolve declared intent retains the questioning structure: it cannot resolve an intent that was never stated, and does not pretend to.

C. Why it compounds

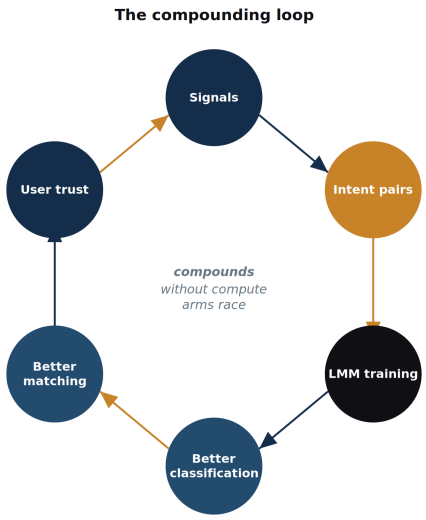

Fig. 4. The compounding loop. Signals yield intent pairs, which train the model, which improves classification and matching, which earns the trust that yields more signals. The asset compounds with participation rather than compute.

The model improves along three axes at once: volume (every signal adds a record), fidelity (every outcome corrects a prior mapping), and breadth (every new domain adds a dimension of meaning). None requires an expanding compute budget; all require expanding participation. A competitor whose business depends on observation cannot assemble the same corpus without first abandoning observation — a contradiction in its own terms.

V. CLASSIFICATION ARCHITECTURE

A declared signal arrives as natural language and must be mapped to a machine-addressable schema. Our prior question-answering work[6] benchmarked this mapping directly, finding that lightweight sparse retrieval can match or exceed far heavier neural models on a domain-specific corpus, and that domain tuning matters more than model size. Classification here is performed by a three-tier architecture, each tier compensating for the failure modes of the one above.

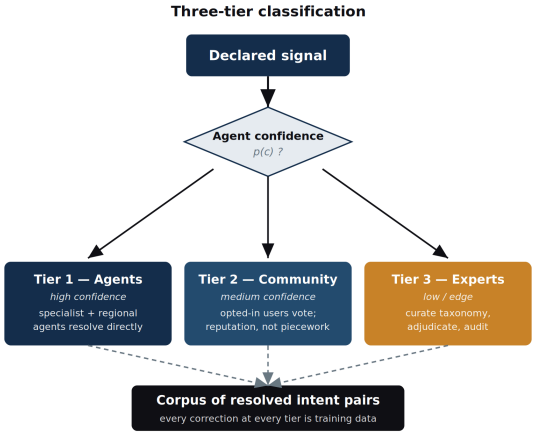

Fig. 5. Three-tier classification. Specialist agents resolve high-confidence signals; opted-in community classifiers resolve the ambiguous middle for reputation, not piecework; independent domain experts handle the long tail, extend taxonomy, and audit. Every correction is training data.

Tier 1 — Agents. Specialist agents, including regional sub-agents fluent in colloquialism and idiom, resolve the majority of signals at high confidence. **Tier 2 — Community.** Ambiguous signals route, anonymised, to opted-in users who vote; consensus resolves the classification and accrues reputation rather than per-task cash, which would invite gaming. **Tier 3 — Experts.** Domain-literate professionals, drawn from the knowledge economy and independent of the transaction economy, resolve the residue, extend the taxonomy, and audit the tier below. The design echoes the expert-profiling finding of our prior work[6], which weighted contribution by demonstrated work rather than self-declared expertise.

VI. PRIVACY ARCHITECTURE

The corpus is honest only if participants behave as they would unobserved, which requires that the network be architecturally incapable of identifying them — not merely committed to refraining. This is the central lesson of our

zero-knowledge advertising work[5], whose protocol we adopt and generalise.

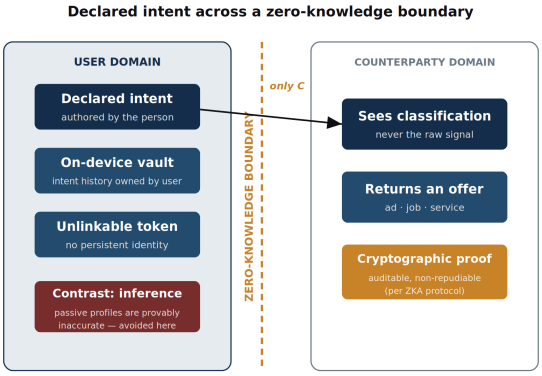

Fig. 6. The zero-knowledge boundary. The user declares intent within a domain they control; counterparties receive only the classification, never the raw signal, and return offers with cryptographic, auditable proofs. Inference-based profiling is structurally excluded.

Three principles hold, while their implementations evolve. First, signals are never linked to persistent identity. Second, counterparties see classifications, not signals. Third, individual signals cannot be recovered from the trained model, through differential privacy[10] or its successors. The current stack — on-device vaults, confidential compute for matching, and zero-knowledge proofs for outcome attestation — follows the auditable, non-repudiable protocol established in [5]. If a user suspects observation, the chilling-effect literature predicts the declaration degrades to what is safe; unobservability is therefore not a feature but a precondition.

VII. PROPOSED STEWARDSHIP

A model trained on the declared intent of a population should not be the property of any single commercial actor. We propose that a foundation be formed — a public-purpose entity that would hold the Large Meaning Model as a commons, steward the protocol, and distribute returns to contributors. No such foundation yet exists; what follows describes the intended structure, not an operating organisation.

Under the proposed structure, operators would implement the network commercially under licence, while contributors — those who declare, classify, and verify — would accrue stake in proportion to the marginal information their contribution adds to the model, redeemed as deferred distributions rather than per-task payment. Stake would become harder to earn as the model matures and novel information grows scarce, aligning early contribution with the period of greatest need. The weighting of contribution by demonstrated value, rather than self-declaration, follows directly from our prior profiling results[6].

VIII. WHY THIS REQUIRES BUILDING

A purely theoretical treatment of the Large Meaning Model would be shorter than this one, and insufficient. The corpus does not exist. It cannot be scraped from the public web, which is observational by construction; it cannot be simulated, since the signal of interest is precisely what people declare when unobserved; and it cannot be extracted from existing systems, whose data is the data this model is defined against.

The corpus must be produced, and production requires a network: a place where declaration is safe, a layer that classifies declarations at scale, a market that creates the conditions for outcomes to occur, and a verification layer that anchors those outcomes to attestable events. None of these components is individually novel; the contribution is their assembly. The position is analogous to gravitational-wave astronomy: the theory was a century old before an instrument could produce the observations, and the instrument was the harder problem. Here too, the argument for declared-intent data is straightforward; the architecture that produces it is the work.

IX. CONCLUSION AND FUTURE WORK

We have argued that observational training data carries an irreducible distortion, that declared intent paired with verified outcome is a structurally different and largely unexploited class of data, and that a model trained upon it — the Large Meaning Model — resolves meaning in a manner language and world models do not. We have grounded the proposal in two prior empirical studies on declared-versus-inferred data and on grounded semantic retrieval, and we have described the network, classification, and privacy architecture required to produce the corpus, together with a proposed foundation to steward the result.

Future work falls in three areas: instrumenting the contribution-attribution mechanism with scalable data-valuation methods[11]; extending outcome verification from declared resolution to cryptographically attested transaction; and the formation of the stewardship foundation itself. The alternative to building is a future modelled entirely on the watched self — fluent, capable, *mathematically perfect and humanly vacant*. A record of what people want when nothing is watching is one of the ingredients any humane future will require. The calculation is not finished. This is a proposal to finish it.

REFERENCES

[1] F. J. Roethlisberger and W. J. Dickson, Management and the Worker. Cambridge, MA: Harvard Univ. Press, 1939.

[2] D. P. Crowne and D. Marlowe, "A new scale of social desirability independent of psychopathology," J. Consult. Psychol., vol. 24, no. 4, pp. 349–354, 1960.

[3] J. W. Penney, "Chilling effects: Online surveillance and Wikipedia use," Berkeley Technol. Law J., vol. 31, no. 1, pp. 117–182, 2016.

[4] T. Kuran, Private Truths, Public Lies: The Social Consequences of Preference Falsification. Cambridge, MA: Harvard Univ. Press, 1995.

[5] P. Callejo, R. Cuevas, A. Cuevas, M. Kotila, L. Bragg, and M. Van Roey, "Zero Knowledge Advertising: a new era of privacy-preserving AdTech solutions," Univ. Carlos III de Madrid / Profila GmbH, 2022.

[6] L. Mazzola, A. Shankar, C. Bless, M. A. Rodriguez, A. Waldis, A. Denzler, and M. Van Roey, "A question answering tool for website privacy policy comprehension," in HCII 2023, LNCS 14045, pp. 194–212, 2023.

[7] S. Larson et al., "An evaluation dataset for intent classification and out-of-scope prediction," in Proc. EMNLP-IJCNLP, 2019.

[8] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[9] Y. LeCun, "A path towards autonomous machine intelligence," Open Review, 2022.

[10] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, no. 3–4, pp. 211–407, 2014.

[11] A. Ghorbani and J. Zou, "Data Shapley: Equitable valuation of data for machine learning," in Proc. ICML, 2019; P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in Proc. ICML, 2017.