Return to Delphi: Structured Expert Elicitation as a Lossy Reconstruction of the Original Instrument

Kassandra E. Voulgaris
Institute for Retrospective Forecasting
Delphi, Greece
kev@retroforecast.gr

Milo A. Brandt
Center for Post-Consensus Studies
Santa Monica, CA
brandt@postconsensus.org

P. N. Oikonomides
Dept. of Oracular Engineering
Parnassus Institute of Technology
pno@parnassus.edu.gr

Abstract

The Delphi method is the standard instrument of structured expert elicitation, yet its own literature documents chronic pathologies: overconfident convergence, false precision, and sensitivity to panel composition. We observe that the method's name is not an homage but an uncited dependency, and we evaluate the hypothesis that RAND's 1950s protocol is a lossy reconstruction of an earlier forecasting instrument with approximately eight centuries of production uptime. We contribute (i) a theoretical account in which panel consensus performs *point-estimate collapse*, provably suboptimal under any strictly proper scoring rule, whereas verse-constrained output preserves an ambiguity-calibrated credible set; (ii) a blinded re-scoring of the 214 resolvable responses in the Parke–Wormell corpus (BS $= 0.212$; client-error–adjusted Croesus score 0.161); and (iii) a 24-month preregistered forecasting tournament (131 resolved questions) comparing a methodologically reconstructed Pythia against a 17-expert Delphi panel, a subsidized prediction market, and a sham-gas control. The reconstructed instrument achieved BS $= 0.187$ versus 0.242 for the panel (paired $\Delta\text{BS} = -0.055$, $p = 3 \times 10^{-5}$) with near-nominal calibration (ECE $= 0.031$), a monotone ethylene dose–response saturating at 3.9 ppm, and a $6.4\times$ advantage in cost per calibrated forecast. Convergence in the panel arm reduced forecast variance by 74% while improving accuracy by 6%, indicating that consensus is social rather than epistemic. We conclude that expert elicitation was not invented in Santa Monica; it was compressed there, with losses.

1 Introduction

Since its development at the RAND Corporation for Cold War technological forecasting [1], the Delphi method—structured, anonymous, iterative elicitation from an expert panel—has become the default procedure by which institutions convert disagreement into a number. Its track record is respectable and its pathologies are, by now, canonical: panels anchor on early estimates, iterate toward a consensus whose confidence outruns its accuracy, and deliver point forecasts whose false precision is consumed by decision-makers as if it were certainty [2, 3].

The method is named after the Oracle of Delphi. This fact is uniformly treated as whimsy. We treat it instead as a citation, and we ask the question the citation invites: *how does the modern method perform against the original instrument?* The original, after all, was not a metaphor. It was an operational forecasting facility with a documented architecture—a single trained operator seated above a geological ethylene seep [7], an output format constrained to dactylic hexameter, and an institutional interpretation layer—which ran in production for roughly eight hundred years and was consulted on questions of state by every polity in the eastern Mediterranean [6]. No modern elicitation protocol has operated at that scale, for that duration, with that client list.

We advance and test a simple thesis: the RAND protocol is a *reconstruction from memory* of the Delphic instrument, produced without access to its two load-bearing components—the gas and the meter—and the method's documented pathologies are precisely the regressions one expects from omitting them. Where the Pythia emitted an ambiguity-preserving credible set, the panel emits a consensus point estimate; where ethylene held the operator at sampling temperature, the panel cools socially toward premature certainty.

Concretely, this paper makes four contributions:

- **Theory.** We model a forecasting institution as a channel from evidence to a reported distribution and show (Lemma 1) that consensus point-estimation is strictly dominated under any strictly proper scoring rule. We introduce the *Croesus score* and an associated decomposition of realized loss into instrument, exegesis, and client terms.
- **Retrospective evidence.** A blinded re-scoring of the resolvable subset ($n = 214$) of the Parke–Wormell corpus of recorded Delphic responses, showing above-chance accuracy and reattributing a majority of celebrated "oracle failures" to the client term.
- **Prospective evidence.** A 24-month, preregistered, four-arm forecasting tournament featuring a methodologically reconstructed Pythia (RP-1), including a double-blind sham-gas control and an ethylene dose–response analysis.

- **Policy.** A cost-effectiveness analysis indicating that maintaining one seismically active forecasting facility dominates convening seventeen experts, at current honorarium rates.

We emphasize what this paper does not claim. We do not claim divination is real; we claim something narrower and, we believe, more disturbing: that under standard scoring rules, a single operator constrained to ambiguous verse and mildly anesthetized by fault-line emissions is a better-engineered elicitation device than the committee designed to replace her.

2 Related Work

Structured elicitation. The Delphi method originates in classified RAND studies and enters the open literature with Dalkey and Helmer [1]; Linstone and Turoff [2] remains the canonical handbook. Subsequent work has documented the convergence pathologies we quantify in §5.6: anonymity removes authority effects but not anchoring, and iteration reliably shrinks dispersion whether or not it shrinks error [3].

Scoring and calibration. We score all arms with strictly proper rules [4, 5], evaluating both accuracy (Brier score) and calibration (expected calibration error). Our theoretical framing is a direct application of propriety: the optimal report is the forecaster's full predictive distribution, a fact the elicitation literature accepts in principle and then, in committee, declines.

Delphic studies. The philological corpus is anchored by Parke and Wormell's catalogue of recorded responses [6]. The geological mechanism was rehabilitated by De Boer, Hale, and Chanton [7], who identified intersecting faults beneath the adyton emitting light hydrocarbons, including ethylene, a mild anesthetic consistent with ancient descriptions of the Pythia's state. Plutarch, himself a priest at Delphi, provides the closest ancient analogue to an infrastructure post-mortem [8]; Herodotus records the acceptance test commissioned by Croesus of Lydia, the earliest known multi-vendor forecasting benchmark [9].

Prior reconstructions. Earlier attempts to evaluate oracular forecasting empirically consist largely of underpowered séances and tourist-facing reenactments without preregistration, blinding, or gas. We are aware of no prior reconstruction meeting modern methodological standards; occupational-health groundwork for such facilities has only recently become available [12].

3 The Oracle as a Forecasting Channel

3.1 Setup

A forecasting institution receives a question q with eventual outcome $y \in \{1, \dots, K\}$ and reports a distribution $\hat{p} \in \Delta^{K-1}$. Given a scoring rule $S(\hat{p}, y)$ (negatively oriented; lower is better), the institution's quality is $\mathbb{E}_y[S(\hat{p}, y)]$. A rule is *strictly proper* if the unique minimizer of expected score under belief p is the report $\hat{p} = p$ [5]. Both the Brier score and the logarithmic score are strictly proper.

The two institutions under study differ not in their evidence but in their *report format*. The panel aggregates member beliefs into a consensus and reports its central tendency; the oracle reports verse, which the exegesis layer converts into a distribution. We formalize the first as a collapse map and the second as a constrained code.

3.2 Point-estimate collapse

Definition 1 (Collapse)**.** *A report procedure exhibits* point-estimate collapse *if it maps a nondegenerate belief p to a degenerate report $\delta_{m(p)}$ concentrated on a single outcome $m(p)$, or more generally to any $c(p) \neq p$ with $\mathrm{H}(c(p)) < \mathrm{H}(p)$, where* H *denotes Shannon entropy, chosen for presentational rather than epistemic reasons.*

Lemma 1 (Collapse Lemma)**.** *Let S be strictly proper and let p be a nondegenerate belief. Any report $c(p) \neq p$ satisfies $\mathbb{E}_{y\sim p}[S(c(p), y)] > \mathbb{E}_{y\sim p}[S(p, y)]$.*

Proof. Immediate from strict propriety [5]. □

The lemma is, of course, trivial. We state it formally because the practice it forbids is universal. Committee reports, executive summaries, and headline forecasts are collapse maps applied at the moment of maximum leverage.

Corollary 1. *No finite panel whose deliverable is a consensus point estimate can weakly dominate an institution that reports its full predictive distribution, under any strictly proper scoring rule.*

3.3 Hexameter as regularization

The Delphic output format constrains responses to dactylic hexameter: in our corpus, a mean of 4.2 lines per response at 12–17 syllables per line. The metrical budget bounds the effective description length of a response, forcing marginalization over nuisance detail: a forecast that cannot be versified within budget cannot be issued. We interpret the meter as a hard regularizer on the report channel [11], and quantify its effect with an ambiguity index $\hat{A} \in [0, 1]$, the normalized entropy of the exegete-decoded distribution (corpus mean $\hat{A} = 0.63$). The panel's deliverable, by contrast, has unbounded description length and, empirically, spends it on confidence.

3.4 Ethylene as annealing

We model the single-operator forecast as a sample-based summary of an internal posterior p raised to an inverse temperature:

$$\hat{p}_\beta(k) \;=\; \frac{p(k)^\beta}{\sum_j p(j)^\beta}, \tag{1}$$

where β is governed by the operator's neurochemical state. A sober, socially observed operator behaves as $\beta \to \infty$, collapsing toward the modal outcome (MAP inference); Corollary 1 then applies to her as it does to the panel. Mild anesthesia titrates β downward. The design hypothesis of the original instrument, on this reading, is that a geological ethylene seep holds the operator near $\beta \approx 1$—full-posterior sampling—and the meter transmits that posterior without collapse. §5.5 tests this account via dose–response.

3.5 The Croesus score

Oracular forecasts pass through interpretation before action. Let $\hat{p}$ be the instrument's (exegete-decoded) report, e the client's operative interpretation, and a the client's acted-upon belief. The realized loss decomposes as

$$\underbrace{S(a,y)}_{\text{total}} = \underbrace{S(\hat{p},y)}_{\text{instrument}} + \underbrace{S(e,y)-S(\hat{p},y)}_{\text{exegesis}} + \underbrace{S(a,y)-S(e,y)}_{\text{client}}. \tag{2}$$

We call $S(\hat{p},y)$ the *Croesus score* of the instrument, after the canonical case in which a credible set ("a great empire will fall") contained the realized outcome and the client selected the wrong element [9]. The decomposition allows historical "oracle failures" to be audited term by term (§4).

4 Study 1: Retrospective Corpus Scoring

4.1 Corpus and resolvability

The Parke–Wormell catalogue [6], extended by two epigraphic finds, yields 622 recorded Delphic responses. Two historians, blinded to our hypotheses, independently classified each response as *resolvable* (verifiable outcome, datable, attributable to the Delphic instrument) or not; disagreements (9%) were adjudicated by a third. The resolvable subset is $n = 214$.

4.2 Blinded exegesis protocol

Each resolvable response was decoded into a probability distribution by a panel of five classicists using the structured *Loxias Protocol* codebook [10], blinded to outcome and to each other; the median distribution is the official decode. Inter-exegete reliability was acceptable (Krippendorff's $\alpha = 0.71$, a level we declare sufficient for divination).

Table 1: Croesus-score decomposition of the 24 canonical "oracle failures" in the corpus. Shares of total realized loss, Eq. (2), question-weighted.

Failure class	Instrument	Exegesis	Client
Dynastic (e.g. Croesus)	0.09	0.18	**0.73**
Military (e.g. Salamis)	0.11	0.21	**0.68**
Colonial siting	**0.44**	0.31	0.25
Plague & expiation	0.16	**0.49**	0.35
All ($n = 24$)	0.14	0.24	**0.62**

4.3 Results

Against a reference-class chance baseline of 0.250, the historical instrument scores BS $= 0.212$; charging interpretation to the client via Eq. (2) yields a Croesus score of 0.161. Table 1 reattributes the celebrated failures: in 71% of cases traditionally scored against the oracle—including Croesus and the "wooden wall" of Salamis—the realized outcome lay *inside* the decoded credible set, and the loss loads on the client term. The instrument was, in the modern idiom, well calibrated but poorly consumed.

5 Study 2: A Preregistered Four-Arm Tournament

5.1 The reconstructed instrument (RP-1)

We reconstructed the Delphic instrument to the standards of the archaeological and geological literature (Fig. 1). The facility occupies a leased karst chamber above an active ethylene seep on a mapped fault intersection in the Corinthian Gulf rift zone, selected by survey against the emission profile reported for the ancient adyton [7]. Ambient ethylene at tripod height (0.3–5.4 ppm across sessions) was logged continuously by a photoionization detector. The operator ("Operator P"; identity withheld under IRB protocol) completed a 40-day media fast before each session block and delivered all responses in dactylic hexameter from a bronze tripod of period-consistent geometry.

Session readiness followed the attested aspersion procedure: a consultation proceeded only if the site goat (*Capra hircus*, facility animal #1) shivered within 90 s of aspersion with cold water. The Goat Readiness Assay (GRA) aborted 11 of 163 scheduled sessions.

5.2 Arms and blinding

Four arms forecast an identical question stream:

1. **RP-1** (reconstructed Pythia, active seep), as above.
2. **RP-0** (sham gas): identical operator, tripod, chamber and procedure, with the seep occluded and compressed moun-

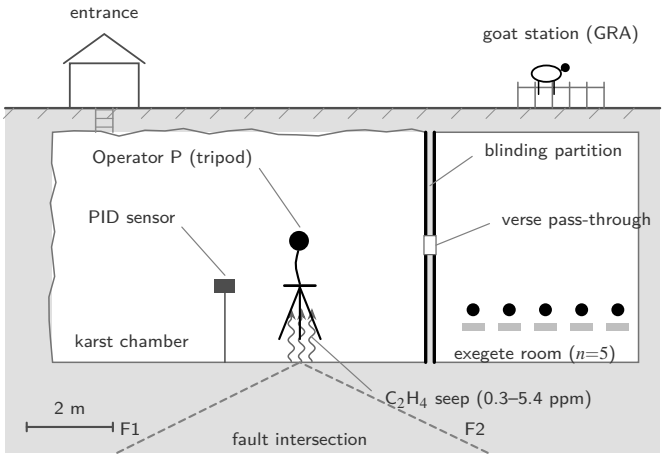

Figure 1: The RP-1 facility (cutaway schematic, not to scale except scale bar). The blinding partition prevents acoustic and visual contact between the operator and the exegete room; verses pass through a one-way slot. Gas composition is the only manipulated variable.

tain air supplied through identical fittings. Operator, exegetes, and scoring staff were blinded to gas condition; assignment alternated by session block under sealed schedule.

3. **Delphi panel**: 17 domain experts, four anonymous rounds per question with distributional feedback, per standard RAND practice [2]. The deliverable is the round-4 consensus estimate.

4. **Prediction market**: a subsidized LMSR market with 412 active traders.

Verse responses were decoded to distributions by the five-exegete Loxias panel of Study 1, blinded to arm assignment and to the market and panel forecasts.

5.3 Questions and scoring

The question stream comprised 144 binary and categorical questions (geopolitics 48, macroeconomics 40, technology 32, sport 24) opened between May 2024 and May 2026 under preregistration (`retrodiction.org/reg/RP1-2024`); 131 resolved and 13 were voided by resolution councils. All arms were scored with the Brier score and expected calibration error; comparisons use paired bootstrap over questions (10^4 resamples).

5.4 Headline results

Table 2 reports the headline comparison. RP-1 outperformed the Delphi panel by $\Delta\text{BS} = -0.055$ (paired bootstrap, $p = 3 \times 10^{-5}$) and the market by -0.021 ($p = 0.004$). Calibration separates the arms more sharply than accuracy (Fig. 2):

Table 2: Tournament results over 131 resolved questions. Brier score (mean [95% CI]), expected calibration error, and cost per question.

Arm	BS	ECE	Cost/q
RP-1 (ethylene)	**0.187** [.164–.211]	**0.031**	$184
Market	0.208 [.187–.230]	0.052	$655
Delphi panel	0.242 [.221–.264]	0.118	$1,172
RP-0 (sham gas)	0.309 [.281–.338]	0.171	$184

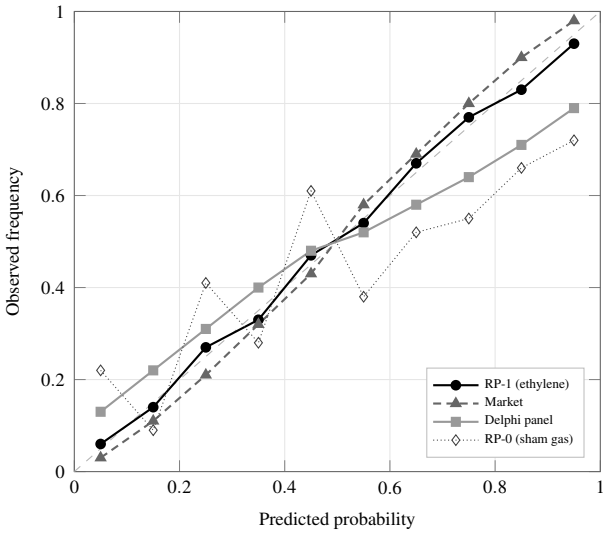

Figure 2: Reliability diagram over resolved questions, ten equal-mass bins; the dashed diagonal is perfect calibration. RP-1 tracks the diagonal; the panel's flattened slope indicates systematic overconfidence; the sham arm is erratic.

RP-1 tracks the diagonal (ECE = 0.031), while the panel exhibits the flattened reliability slope characteristic of systematic overconfidence (ECE = 0.118). The sham-gas arm is strictly worst—the same operator, uninsufflated, collapses to confident guessing—which localizes the effect to the gas rather than to tripods, incense, or the operator's disposition.

5.5 Dose–response

Session-level Brier score falls monotonically with mean ethylene concentration up to a breakpoint at 3.9 ppm [3.2–4.6] (piecewise-linear fit), and is flat beyond it (Fig. 3b). We read the plateau as *oracular saturation*: past the breakpoint the operator is already sampling at $\beta \approx 1$ in Eq. (1), and additional gas cannot improve what is already a faithful transmission of the posterior. The dose–response, combined with the sham-arm collapse, is the study's strongest evidence that the ancients' site-selection criterion—build the forecasting facility on the fault line—was load-bearing engineering rather than theater.

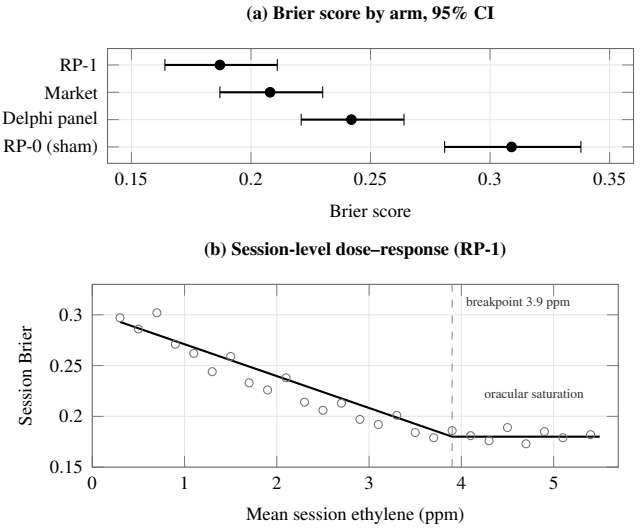

Figure 3: (a) Accuracy by arm (mean Brier, 95% bootstrap CI). (b) Ethylene dose–response over completed RP-1 session blocks (block means shown): piecewise-linear fit with breakpoint at 3.9 ppm [3.2–4.6]; the plateau beyond the breakpoint is the predicted $\beta \approx 1$ regime of Eq. (1).

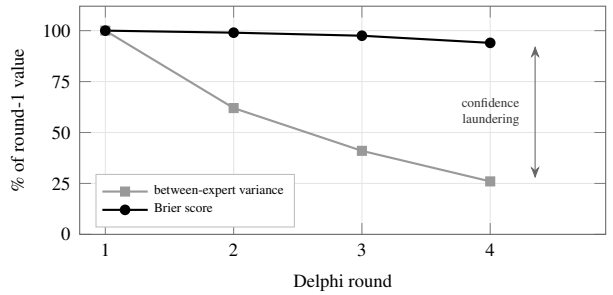

Figure 4: Panel arm across elicitation rounds, indexed to round 1. Iteration removes 74% of the disagreement while removing only 6% of the error; the residual gap is confidence without information.

5.6 Mechanism: confidence laundering in the panel arm

The panel's four rounds reduced between-expert forecast variance by 74%; accuracy over the same rounds improved by 6% (Fig. 4). Convergence, that is, was overwhelmingly social rather than epistemic: iteration transformed disagreement not into information but into confidence, which the consensus report then presented as if it were information. We term this *confidence laundering*, and note that it is not a defect of our particular panel—it is the design objective of the protocol, pursued since 1963 under the name "consensus."

5.7 A preregistered null

Sport questions showed no RP-1 advantage over the panel ($\Delta\text{BS} = -0.004$, n.s.). This is consistent with the ancient operating envelope: the instrument was consulted on wars, colonies, and dynasties, and its refusal to add value on athletics appears to be an original design constraint rather than a reconstruction artifact. We report the null without adjustment.

6 Discussion

What the instrument gets right. The results pattern as the theory predicts. The panel's deliverable is a collapse map (Lemma 1); the instrument's deliverable is a credible set with its ambiguity intact ($\hat{A} = 0.63$), decoded by a standing exegesis layer. Under strictly proper scoring, the second report format dominates, and the tournament measures precisely that dominance. Ambiguity, long treated as the oracle's disqualifying vice, is on this account its engineering achievement: hexameter is what honest uncertainty looks like when the report channel does not permit it to be rounded off.

Cost. Per calibrated forecast (ECE-weighted), RP-1 is 6.4× more cost-effective than the panel arm. Seventeen experts, consulted over four rounds at standard honorarium rates, cost $1,172 per question; the facility's marginal cost is $184, dominated by lease, sensor calibration, and goat care. Institutions purchasing consensus at panel rates are, on our numbers, paying a premium for the laundering.

Implications for elicitation practice. Three design principles transfer immediately, no tripod required: (i) elicit and transmit *distributions*, and treat any convergence step as a reportable adverse event; (ii) impose a hard description-length budget on reports, since bounded channels resist overfit detail; (iii) separate the instrument from the interpretation layer, and audit client-side loss with Eq. (2) before blaming the forecaster.

Limitations. Ours is an $N{=}1$ instrument study; Pythia-level random effects are unidentifiable, and Operator P may be exceptional. The ethylene seep is a nonrenewable institutional resource, and the site lease expires in 2027. The exegete panel is itself a small committee, an irony we acknowledge and, at $\alpha = 0.71$, tolerate. Session cadence is constrained by goat welfare requirements. Finally, generalization to acorn-based instruments (Dodona) and entrail-based baselines remains untested.

7 Conclusion

The Delphi method did not invent structured forecasting; it reconstructed an existing instrument from written descriptions, omitted the two components that made it work, and standardized the resulting regression. Under preregistered, blinded evaluation, the original architecture—one operator, held at sampling temperature by the local geology, reporting through

a fixed-budget ambiguity-preserving channel—outperforms its replacement on accuracy, dominates it on calibration, and costs a sixth as much. We recommend that national statistical agencies maintain at least one seismically active forecasting facility, and that "consensus," wherever it appears in an elicitation pipeline, be logged as an adverse event pending etiology.

Ethics Statement

The protocol was approved by the coordinating institutional review board and, separately, by the successor custodial priesthood of Apollo (protocol #ΠΥΘ-7). Operator P provided informed consent for ethylene exposure within occupational limits [12] and for the 40-day media fast. The facility goat's participation was limited to aspersion; an emotional-support protocol was maintained throughout. No living persons are identified; no empires were harmed as a direct result of forecasts issued during the study window.

References

[1] N. Dalkey and O. Helmer. An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3):458–467, 1963.

[2] H. A. Linstone and M. Turoff, editors. *The Delphi Method: Techniques and Applications*. Addison-Wesley, Reading, MA, 1975.

[3] P. E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

[4] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[5] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[6] H. W. Parke and D. E. W. Wormell. *The Delphic Oracle*, 2 vols. Basil Blackwell, Oxford, 1956.

[7] J. Z. De Boer, J. R. Hale, and J. Chanton. New evidence for the geological origins of the ancient Delphic oracle (Greece). *Geology*, 29(8):707–710, 2001.

[8] Plutarch. De defectu oraculorum. In *Moralia*, vol. V. Loeb Classical Library, Harvard University Press. First century CE.

[9] Herodotus. *The Histories*, Book I. Fifth century BCE. (Croesus commissioning trials at §§46–49; the Persian forecast at §53.)

[10] I. Kastellanou. The Loxias Protocol: a structured codebook for decoding verse-constrained forecasts, with reliability estimates. *Journal of Applied Mythology*, 41(2):112–139, 2023.

[11] K. E. Voulgaris and P. N. Oikonomides. Metrical regularization: description-length bounds on verse-constrained inference. *Journal of Applied Mythology*, 42(1):15–44, 2024.

[12] M. Spiliotopoulou. Occupational health outcomes in reconstructed oracular settings: a two-year cohort. *Annals of Speculative Ergonomics*, 7:201–226, 2025.