

UOIF Evaluation Reflections: Proof Logic, Confidence, Keystone Insights, and Live Verification

UOIF Working Note

August 08, 2025

1 Model (with Proof Logic)

$$\Psi(x) = \underbrace{[\alpha S(x) + (1 - \alpha) N(x)]}_{\text{Hybrid linearity}} \cdot \underbrace{\exp(-[\lambda_1 R_a + \lambda_2 R_v])}_{\text{Exponential boundedness}} \cdot \underbrace{P(H | E, \beta)}_{\text{Bayesian calibration}}, \quad (\lambda_1, \lambda_2) = (0.85, 0.15), S(x) = 0.60$$

- Hybrid linearity: $O(\alpha) = \alpha S(x) + (1 - \alpha) N(x)$; $\partial O / \partial \alpha = S(x) - N(x) < 0$ if $N(x) > S(x)$.
- Exponential boundedness: $R_a, R_v \geq 0 \Rightarrow \text{Penalty} \in (0, 1]$.
- Posterior: $P(H | E, \beta) = \min\{\beta P(H | E), 1\}$ (uplift with cap).

2 Verification (Real-Time: Aug 08, 2025)

Official 2025 **results** are live: year info, country, individual, statistics. The **problems/shortlist** pages have no 2025 content yet.

3 Evaluation A: IMO 2025 Results (Canonical Live)

Sources. Official results + DeepMind 2025 + Evan 2025 + AoPS. Confidence: 0.98.

Stepwise derivation (discerned choices)

- Hybrid (prioritize canonical): take $\alpha = 0.12$, $N(x) = 0.97$.
 $O = 0.12 \cdot 0.60 + 0.88 \cdot 0.97 = 0.9256$. Confidence: 0.96.
- Penalty (eased for canonical results): set $(R_a, R_v) = (0.12, 0.04)$.
 $\text{Penalty} = \exp(-[0.85 \cdot 0.12 + 0.15 \cdot 0.04]) = \exp(-0.108) = 0.8977$. Confidence: 0.85.
- Posterior (certified + canonical uplift): choose $\beta = 1.15$ and cap $P(H | E, \beta) = 1.0$.
Confidence: 0.88.
- Value: $\Psi(x) = 0.9256 \times 0.8977 \times 1.0 \approx \mathbf{0.831}$.
Sensitivity: $\partial \Psi(x) / \partial \alpha = (S(x) - N(x)) \text{Penalty} P(H | E, \beta) < 0$. Confidence: 0.92.

Keystone reflection

$\Psi(x)$ upgrades results to *Primitive/Empirically Grounded* cleanly upon canonical arrival, while the exponential and cap guard against overconfidence. This evidential synthesis is monotone in α (with $N > S$), encouraging canonical-first allocation.

4 Evaluation B: IMO 2025 Problems (Pending Canonical)

Sources. Evan 2025 + AoPS + DeepMind 2025 (expert interpretive). Confidence: 0.88.

Stepwise derivation (discerned choices)

- Hybrid (interpretive emphasis): choose $\alpha = 0.17$, $N(x) = 0.89$.
 $O = 0.17 \cdot 0.60 + 0.83 \cdot 0.89 = 0.8407$. Confidence: 0.85.
- Penalty (stricter, non-canonical): $(R_a, R_v) = (0.25, 0.10)$.
Penalty = $\exp(-[0.85 \cdot 0.25 + 0.15 \cdot 0.10]) = \exp(-0.235) \approx 0.7965$. Confidence: 0.80.
- Posterior (modest uplift): $P(H | E) \approx 0.90$, $\beta = 1.05 \Rightarrow P(H | E, \beta) \approx 0.945$. Confidence: 0.85.
- Value: $\Psi(x) \approx 0.8407 \times 0.7965 \times 0.945 \approx \mathbf{0.633}$.
Range over $\alpha \in [0.15, 0.20]$, $N(x) \in [0.88, 0.90]$: **0.62–0.64**. Confidence: 0.85.

Keystone reflection

With problems non-canonical, $\Psi(x)$ stays *Interpretive/Contextual*. The framework’s damping prevents premature promotion while keeping the path open for immediate uplift when official texts post.

5 Evaluation C: IMO 2024 (DeepMind P1/P2/P4)

Sources. DeepMind 2024 pages + Evan + established archives/mirrors. Confidence: 0.90.

Stepwise derivation (discerned choices)

- Hybrid (balanced primitive weighting): pick $\alpha = 0.12$, $N(x) = 0.96$.
 $O = 0.96 - 0.36 \cdot 0.12 = 0.9168$. Confidence: 0.88.
- Penalty (established context): $(R_a, R_v) = (0.10, 0.05)$.
Penalty = $\exp(-[0.85 \cdot 0.10 + 0.15 \cdot 0.05]) = \exp(-0.0925) = 0.9117$. Confidence: 0.85.
- Posterior (moderate uplift): $P(H | E) \approx 0.90$, $\beta = 1.05 \Rightarrow P(H | E, \beta) = 0.945$. Confidence: 0.85.
- Value: $\Psi(x) \approx 0.9168 \times 0.9117 \times 0.945 \approx \mathbf{0.789}$.
Extremes: $\alpha = 0.10 \Rightarrow 0.796$; $\alpha = 0.15 \Rightarrow 0.781$. Confidence: 0.88.

Keystone reflection

Artifacts are sufficient for *Primitive/Empirically Grounded*. The exponential term keeps claims conservative where official texts are partial, while rewarding verifiability and authority.

6 Summary Table

Evaluation	α	N	Penalty	$P(H E, \beta)$	$\Psi(x)$	Label
2025 Results (canon)	0.12	0.97	0.8977	1.00	0.831	Primitive/Emp. Grounded
2025 Problems (pending)	0.17	0.89	0.7965	0.945	0.633	Interpretive/Contextual
2024 P1/P2/P4	0.12	0.96	0.9117	0.945	0.789	Primitive/Emp. Grounded

Condensed Citations

Official IMO 2025 (results): year info, country, individual, statistics.

DeepMind 2025: blog, solutions PDF.

DeepMind 2024: blog, solutions index.

Evan Chen 2025: notes. AoPS 2025: P1–P6 thread IDs as prior.