



International AI Innovation
OLYMPIAD

2026 IAI²O AI INNOVATORS CHALLENGE

FINAL - PROBLEM DESIGN AND RATIONALE

A systematic explanation of the design principles, learning objectives, assessment targets, and difficulty calibration behind the 2026 AIIC Final.

Prepared by the 2026 IAI²O Scientific Committee

DOCUMENT PURPOSE

This document explains why each problem was included in the 2026 AI Innovators Challenge Final, what knowledge and reasoning skills it was designed to measure, how the subparts progress in depth, and how Part 1 and Part 2 work together to evaluate broad AI competence.

AI competence = conceptual understanding + mathematical reasoning + implementation + experimental judgment

International AI Innovation Olympiad (IAI²O)

2026 Global Final



2	Part 1 - AI Theoretical Foundations	4
2.1	Question 1 - Multiple-Select Conceptual Foundations	4
2.2	Question 2 - Classical Machine Learning	6
2.3	Question 3 - CNN, Softmax, and Backpropagation	7
2.4	Question 4 - Advanced Deep Learning: Self-Attention	8
3	Part 2 - Applied Machine Learning	9
3.1	Problem 1 - Diamond Price Prediction	9
3.2	Problem 2 - Galaxy Morphology Classification	10
4	Relationship Between Part 1 and Part 2	11
5	Assessment Philosophy and Contestant Differentiation	12



mathematical reasoning, and the ability to connect mathematical principles to modern machine-learning systems.

Rather than rewarding memorization of definitions or familiarity with particular software libraries, the problems are designed to determine whether contestants can reason from first principles, recognize structural relationships between different models, and explain why an algorithm behaves in a particular way.

The examination deliberately spans classical machine learning, neural networks, sequential decision making, search, recommender systems, convolutional networks, optimization, and Transformers. This breadth is not intended simply to test how many topics a contestant has studied. Instead, it tests whether the student can move comfortably between different AI paradigms while identifying the mathematical principles shared among them.

A second important design principle is progressive depth. Question 1 evaluates broad conceptual fluency. Question 2 requires quantitative reasoning across several classical machine-learning methods. Question 3 connects neural-network architecture, forward propagation, probability, loss functions, and gradient computation. Question 4 then requires contestants to derive and prove a structural property of self-attention.

recognize → calculate → derive → prove and interpret

This structure is intended to distinguish students who have encountered AI terminology from students who have developed a deeper mathematical understanding of how AI systems actually work.

Part 2 complements this theoretical assessment with applied tasks. The design intentionally shifts from mathematical reasoning to data judgment, model construction, validation strategy, and experimental iteration.

Question 1 contains five multiple-select subproblems spanning recurrent neural networks, reinforcement learning, A* search, softmax, and recommender systems. The goal is broad conceptual coverage without reducing the assessment to definition recall.

Each option is designed to be individually plausible. Contestants must distinguish statements that are always true, conditionally true, or false. This format is particularly effective in AI, where many misconceptions arise from oversimplified explanations.

Question 1.1 - Recurrent Neural Networks and LSTMs

Design objective. Test whether students understand gradient propagation through time at the level of repeated Jacobian multiplication, rather than simply knowing that RNNs may suffer from vanishing gradients and that LSTMs help.

Key concepts assessed. Backpropagation through time; products of Jacobians; spectral norms; vanishing and exploding gradients; activation-function limitations; LSTM cell-state dynamics; forget gates.

Why the distractors matter. The ReLU option targets the misconception that a different nonlinearity automatically removes gradient pathologies. The forget-gate option tests whether students understand that values near one preserve cell-state information while values near zero erase it.

Rationale. This subproblem separates memorized vocabulary from mechanism-level understanding. Gradient propagation is a foundational idea that later reappears in residual networks, deep optimization, and Transformer training.

Question 1.2 - Reinforcement Learning and Q-Learning

Design objective. Combine a simple numerical update with conceptual distinctions between the TD target, learning rate, behavior policy, target policy, terminal states, and exploration.

Key concepts assessed. Bellman-style bootstrapping; temporal-difference targets; off-policy learning; ϵ -greedy behavior; terminal-state handling.

Deeper design intent. A recurring difficulty in reinforcement learning is distinguishing how data is collected from how value estimates are updated. A contestant may compute the update correctly yet still misunderstand why Q-learning is off-policy.

Rationale. The problem rewards students who understand reinforcement learning as a learning process involving targets, policies, and bootstrapping rather than as a single formula.

Why this is more than a definition question. Contestants must reason from inequalities such as $h(n) \leq h^*(n)$, and from the consistency inequality, rather than simply recall terminology.

Rationale. Heuristic search is an early but important example of combining domain knowledge with formal guarantees. The problem tests whether students can manipulate an AI concept mathematically.

Question 1.4 - Softmax, Temperature, and Cross-Entropy

Design objective. Test whether contestants understand softmax as a mathematical transformation, including invariance, ranking, limiting behavior, and gradients.

Key concepts assessed. Additive shift invariance; effect of positive rescaling; temperature; probability ranking; high-temperature uniform limit; cross-entropy derivatives.

Why temperature is included. Temperature scaling appears across modern AI, including calibration, knowledge distillation, contrastive learning, and sampling. It is mathematically accessible but broadly relevant.

Rationale. The subproblem connects probability, optimization, and modern neural-network practice and rewards students who can reason qualitatively without relying on heavy computation.

Question 1.5 - Recommender Systems and Matrix Factorization

Design objective. Introduce a practical AI system while still requiring mathematical reasoning about latent factors, missing data, cold start, and model identifiability.

Key concepts assessed. Matrix factorization; dot-product prediction; missing-not-at-random data; latent representations; ID-only cold start; scaling non-identifiability; capacity versus generalization.

Without regularization, $p' = c p$ and $q' = q / c$ leave the dot-product prediction unchanged.

Rationale. Recommender systems demonstrate that successful machine learning depends both on model structure and on understanding what the observed and missing data actually mean.

applying one type of calculation.

Question 2(a) - Naive Bayes

Key concepts assessed. Bayes rule; Laplace smoothing; class-conditional likelihoods; posterior normalization; conditional independence; model misspecification.

Deeper rationale. Many introductory questions stop after computing a posterior. This problem deliberately asks what happens when the conditional-independence assumption is later found to be false. The critical insight is that a model may be misspecified while still producing the correct class decision.

What strong students should recognize. Model correctness, probability calibration, and classification correctness are distinct concepts. Correlated features may distort estimated probabilities without necessarily changing the predicted label.

Question 2(b) - Support Vector Machines

Key concepts assessed. Functional margins; canonical support hyperplanes; decision-boundary invariance; geometric normalization.

Design intent. The scaling transformation $(w,b) \rightarrow (3w,3b)$ was chosen because it exposes why raw functional margins are representation-dependent while geometric margins are not.

The hyperplane $w^T x + b = 0$ is unchanged by positive common scaling, while the normalized geometric margin remains invariant.

Rationale. Rather than merely calculating a margin, contestants must connect algebraic representation to geometric meaning.

Question 2(c) - XGBoost

Key concepts assessed. Gradient and Hessian statistics; regularized optimal leaf weights; split gain; structural penalties; model-selection tradeoffs.

Design intent. Many students have used boosting through libraries. This problem reverses the black-box experience and exposes the optimization used to decide whether a split should exist and what each leaf should predict.

Why γ is included. The final subpart shows that regularization can change model structure itself, not merely fitted parameter values.

Rationale. The problem tests whether students understand a widely used algorithm internally rather than recognizing only its name or software interface.

Question 3(a) - CNN Architecture Analysis

Key concepts assessed. Convolution output dimensions; channels; pooling; trainable parameter counts; receptive fields.

Design intent. These are not merely implementation details. They test whether students understand how local computations propagate spatially and how model size is determined.

Why receptive field is included. Students often know how to compute tensor shapes yet cannot explain which input pixels influence a later activation. Receptive-field reasoning therefore distinguishes framework familiarity from architectural understanding.

Question 3(b) - Forward Propagation

Key concepts assessed. Dense-layer logits; softmax normalization; exact probability computation; cross-entropy.

Number design. Sparse activations and $W_{1,1} = \ln 2$ are chosen so that $e^{(\ln 2)} = 2$. This keeps the arithmetic exact and lets the problem measure understanding instead of calculator speed.

Rationale. Contestants must connect a linear layer to probabilistic predictions and then to the training objective.

Question 3(c) - Backpropagation

Key concepts assessed. Softmax-cross-entropy gradient; chain rule; matrix calculus; outer-product gradients; sign interpretation under gradient descent.

$$\partial L / \partial W = a (\nabla_z L)^T$$

Design intent. Students begin from a known local derivative and must derive the full matrix gradient rather than quote a memorized formula.

Why the sign interpretation matters. The final part checks whether students understand what the derivative means: the true-class logit should be driven upward while competing logits are pushed downward.

Rationale. This question integrates calculus, linear algebra, and optimization into one coherent explanation of neural-network learning.

information and causal masking.

Question 4(a) - Forward Pass

Key concepts assessed. Construction of Q , K , V ; dot-product score matrix; row-wise softmax; attention output.

Design intent. Before asking for an abstract proof, the problem makes students compute a complete miniature attention layer exactly. This provides a concrete anchor for the structural reasoning that follows.

Question 4(b) - Permutation Equivariance

Key concepts assessed. Permutation matrices; matrix transformations; row and column permutations; row-wise softmax; equivariance proofs.

$$Q' = P Q, K' = P K, V' = P V, S' = P S P^T, \text{ and ultimately } O' = P O.$$

Why this is the hardest part. Contestants must track dimensions, reason about how P acts on both rows and columns, justify how row-wise softmax transforms, and compose these facts into a proof.

Rationale. This moves beyond using a Transformer formula. It asks students to derive a structural property of an architecture from first principles, which is much closer to the reasoning required in AI research.

Question 4(c) - Positional Information and Causal Masking

Key concepts assessed. Interpretation of permutation equivariance; positional encodings as symmetry breaking; fixed causal masks; conditions under which equivariance fails.

Deeper rationale. The key lesson is that model capability can arise by deliberately breaking a mathematical symmetry. Pure self-attention has no inherent concept of token order; positional encodings and causal masks inject structure that the attention operation itself does not contain.

diamond prices. The task requires exploratory analysis, preprocessing, feature handling, modeling, validation, and submission generation rather than direct fitting of a perfectly prepared matrix.

Official metric. RMSLE (Root Mean Squared Logarithmic Error), lower is better.

Key Skills Assessed

- Data inspection and exploratory data analysis
- Handling imperfect or inconsistent data
- Categorical and numerical feature processing
- Feature engineering and interaction discovery
- Validation design and leakage avoidance
- Regression modeling and model comparison
- Hyperparameter selection and competition submission discipline

Why Diamond Pricing

Diamond prices depend on nonlinear interactions among carat, dimensions, cut, color, clarity, and related features. The dataset therefore rewards contestants who explore interactions and representation choices instead of relying on a single linear assumption.

Why RMSLE

Prices span a wide range. RMSLE emphasizes relative error more than raw absolute error and reduces the extent to which a small number of very expensive diamonds dominate the score. This makes the metric better aligned with proportional prediction quality across the dataset.

Difficulty Calibration

A conventional tree-based model with sensible preprocessing can produce a respectable baseline, while stronger contestants can gain performance through more careful feature engineering, validation, model choice, ensembling, and tuning. The problem is intentionally accessible at entry level while retaining substantial optimization headroom.

Key Skills Assessed

- CNN architecture design
- PyTorch implementation
- Image preprocessing and augmentation
- Training stability and optimization
- Regularization and learning-rate control
- Class imbalance and class-aware objectives
- Validation strategy under limited time and compute
- Experimental iteration and reproducibility

Why Galaxy Morphology

Galaxy classes depend on meaningful spatial structure such as smoothness, spiral arms, bars, merger morphology, and edge-on appearance. The task therefore requires the network to learn useful visual representations instead of exploiting trivial color cues.

Why Orientation Augmentation Matters

Galaxy morphology is largely invariant to image rotation. A spiral galaxy does not change its physical class because the image is rotated. This introduces an important machine-learning principle: known invariances of the real problem should be reflected in model design and data augmentation.

Why Macro F1

The dataset is class-imbalanced. Plain accuracy could reward a model that performs very well on common classes while neglecting rare morphologies. Macro F1 computes class-level F1 scores first and then averages them, giving each morphology category equal influence on the final result.

Difficulty Calibration

The starter CNN provides a functional baseline so that contestants can focus on modeling rather than infrastructure. Stronger contestants can improve performance through architecture design, normalization, residual connections, regularization, class weighting, augmentation, scheduling, and more disciplined validation.



Part 1 asks whether contestants understand why machine-learning algorithms work.

Part 2 asks whether they can make those algorithms work on real data.

Part 1 examines model assumptions, geometry, optimization, CNN architecture, probability, gradients, attention, and symmetry at a conceptual and mathematical level. Part 2 then places contestants in settings where those same ideas affect practical performance.

The Diamond task emphasizes data understanding and classical machine-learning judgment. The Galaxy task emphasizes representation learning, optimization, class imbalance, and experimental deep learning.

AI competence = conceptual understanding + mathematical reasoning + implementation + experimental judgment

The competition is therefore designed not to identify the contestant who has memorized the largest number of algorithms, nor simply the contestant who can use the most sophisticated library. It aims to identify students who can understand, derive, implement, evaluate, and improve intelligent systems.



	Problem Solving
Foundational fluency	Recognizes core concepts and standard formulas; can complete straightforward numerical steps.
Operational understanding	Can apply concepts correctly in unfamiliar numerical settings and explain standard behaviors.
Structural understanding	Can distinguish closely related ideas, identify invariances and assumptions, and connect algebra to geometry or probability.
First-principles reasoning	Can derive gradients, prove architectural properties, and explain why a method behaves as it does.
Research-style maturity	Can interpret consequences of a proof, identify when assumptions fail, and reason about how model structure encodes or omits information.

The full contest architecture is intentionally broad but not shallow. Breadth is used to ensure that finalists can transfer reasoning across paradigms; depth is introduced through derivation, proof, hidden assumptions, regularization, validation, and experimental judgment.

The strongest contestants should therefore demonstrate not only correctness, but also the ability to explain mechanisms, recognize when standard statements are overgeneralized, and adapt principles to unfamiliar settings.

END OF DESIGN AND RATIONALE

2026 IAI²O AI Innovators Challenge Final

International AI Innovation Olympiad