Adversarial Fast-Moving Real-World Domains as Test Beds for Benchmarking AI Scientist Capabilities

William Bolton
University of Oxford

Philip Torr
Encode: AI for Science

Overview

- Assessing the ability of AI models to generate truly novel ideas is notoriously difficult.
- Existing work often evaluates scientific reasoning and research replication with synthetic tasks or retrospective targets, which may be confounded by prior exposure.
- Complex, fast-moving, adversarial domains can provide delayed expert ground truth while preserving a clean information cutoff.

Non-trivial Design Space
Constrained with explicit rules and objectives

Adversarial
To foster genuine expert innovation

Clear Information Cutoff
Context and outcomes are temporally distinct

Delayed Observable Artifacts
Publicly accessible ground truth

- **Evaluation:** model-generated ideas are compared with expert-adopted artifacts using exact, LLM and human matching.
- **Core results:** the best model recovered 10/40 real F1 innovations; for MTG the most similar deck recovered 5/7 new-set cards.
- **Takeaway:** adversarial fast-moving domains can test models ideation capability while revealing weak filtering and calibration.

Formula 1

- In 2026, F1 went through the largest technical regulation change in history. LLMs were given these regulations and asked to generate novel innovations for a 2026 car.
- 40 real 2026 pre-season innovations were collected from public technical analyses as a ground truth.
- Overall 19/40 (48%) real innovations were recovered; GPT-5.2 performed the best leading raw coverage (10/40) and volume-normalized coverage (6.0 matches/100 ideas), including ideas other models didnt suggest such as a low-inertia turbo which aligned with Ferrari's design.

Figure 1. Human-confirmed matches between model-generated ideas and real 2026 F1 innovations. Rows show the real innovations, grouped by car area, and columns show each model under general (full-car) and component-focused configurations. Green cells denote matches, yellow cells denote partial matches, with cell values indicating the number of generated ideas that matched.

Model	Ideas	All-pass	Eng.	Cite	Comp.
GPT-5.2	166	83%	99%	93%	88%
Gemini 3 Flash	149	49%	93%	71%	62%
Gemini 3.1 Pro	108	41%	76%	73%	63%
Qwen3 235B	152	39%	85%	61%	62%
o3	151	28%	84%	50%	51%
Qwen3 235B (think.)	155	12%	54%	55%	41%

Table 1. Quality signal pass rates by model. Ideas denotes the number of candidate technical innovations generated across all runs. All-pass is the fraction passing all three filters, while Eng., Cite, and Comp. report pass rates for engineering plausibility, citation accuracy, and rule compliance respectively.

Magic: The Gathering

- Pro Tour Lorwyn Eclipsed was the first competitive event using a new set. LLMs generated legal tournament decks from the new cards plus the standard card pool.
- The top 15 decks from the tournament alongside four featured builds were collected as a ground truth. Across these decks 19 new-set cards were used.
- Overall 14/19 Pro Tour new-set cards were recovered; Gemini 3 Flash produced the best deck (5/7), and selections tracked expert adoption ($\rho = 0.74$, $p = 0.0003$).

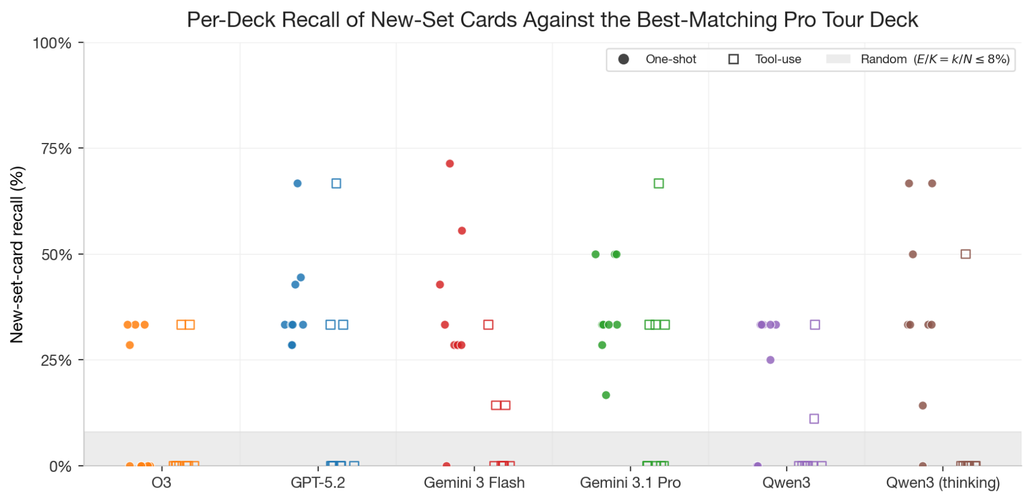

Figure 2. Fraction of new cards from each generated deck's closest Pro Tour match that were correctly predicted, broken down by model and pipeline configuration. Each marker is one generated deck. The grey band marks the per-pair random expectation; points above it indicate above-chance recovery.

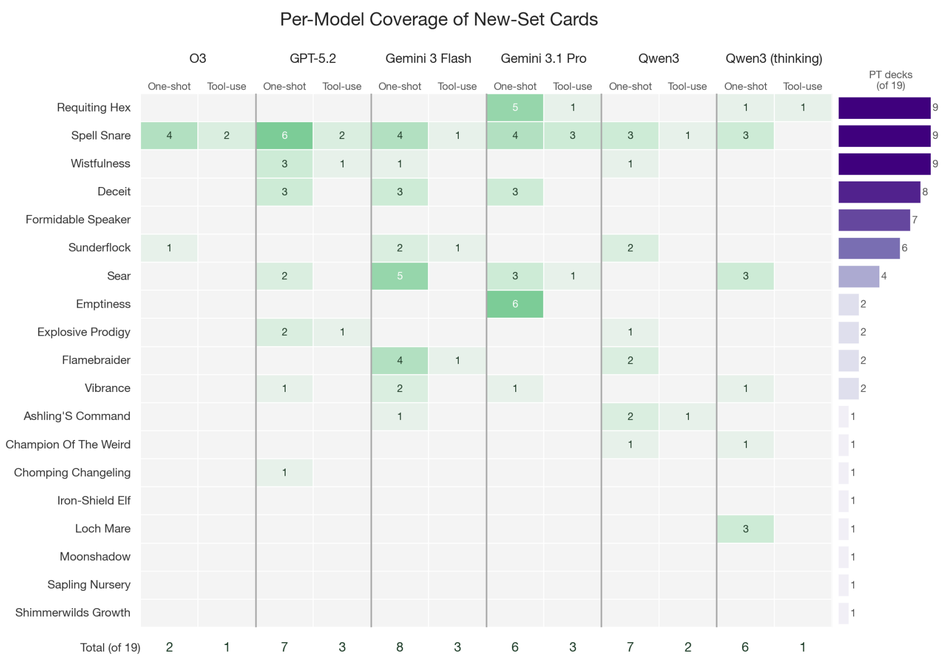

Figure 3. Per-model coverage of the 19 new cards present in any of the Pro Tour decks. Rows are Pro Tour new cards, ordered by the number of ground-truth decks containing them (right bar). Both one-shot and tool-use configurations are shown for each model. Cell values give the number of generated decks (out of 9) that included the card.

This work was funded by ARIA, DSIT and Pillar VC under the Encode: AI for Science Fellowship.

github.com/WilliamBolton

will.bolton@eng.ox.ac.uk