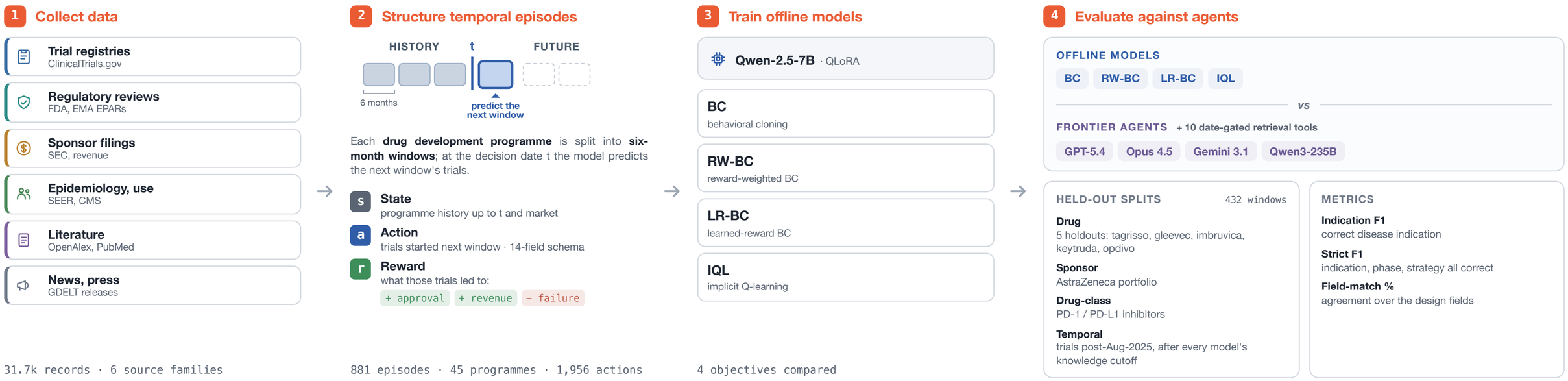
Learning Clinical-Trial Strategy: Offline Policy Training for Decision Agents

William Bolton
University of Oxford

Philip Torr
Encode: AI for Science

Overview

- Clinical drug development is sequential decision-making under uncertainty, where a sponsor must repeatedly plan a portfolio of experiments (i.e., *what trials to run next*) from heterogeneous evidence, with outcomes only observed years later. Prior research has investigated the likelihood of trial success. We instead study the upstream strategy of clinical development; given a program's state at time t, what portfolio of trials should be run next?



- **Results:** Models trained offline outperform the non-fine-tuned baselines, particularly in the post-August 2025 contamination-clean hold-out, where reward-weighted behavioral cloning obtains 46.2% Indication F1 and 14.2% Strict F1 against 25.0% and 2.1%, respectively, for the best-performing agent.
- **Takeaway:** Agents use parametric knowledge to name the right disease but often assemble the wrong trial design; offline learning can teach a decision policy that enables models to plan clinical experiments more effectively.

Data

- Each episode is a program-window pair: **states** s contain only pre-t information (trials, approvals, competitive landscape, epidemiology, market), and the **action** a is the set of trials launched in the next six-month window.
- Every trial and action is represented as a 14-field JSON object: indication, trial description, phase, strategy, study type, allocation, masking, arm and enrollment buckets, region, comparator, combination partners, and primary/secondary endpoints.
- **Rewards** R are defined by real downstream outcomes: FDA approvals credited to specific trials, attributed revenue, and failures.
- **881** (s, a, R) episodes over **45** programs, split **8 ways** to test transfer to unseen drugs, sponsors, drug-classes, and future unseen trials.

Methods

- One backbone was fine-tuned (Qwen-2.5-7B + QLoRA); the four objectives differ only in how each logged decision is weighted.
- **BC** — behavioral cloning; all windows weighted equally.
- **RW-BC** — reward-weighted by the per-trial outcome reward $r_i = 2\,a_i - f_i + \log(1 + \mathrm{rev}_i)$.
- **LR-BC** — the reward is learned from trial + program features.
- **IQL** — value-based offline RL; weighted by advantage $A(s, a)$.
- Scored by *Indication F1*, *Strict F1*, and *Field-match %*.

Limitations & Future Work

- **Limitations:** rewards are associational, not causal estimates of value · unobserved confounders affect decisions · per-trial credit is local · the dataset size is limited.
- **Future work:** closed-loop agentic systems and offline-to-online methods · compose retrieval with offline training · hierarchical RL with sub-goals to tackle the problems of long horizons and sparse delayed rewards.

Results

- Offline models see the greatest benefit on Strict F1, which rewards wider trial design similarity than Indication F1 (Figure 1).
- The gap between offline and agent performance widens on less-iconic programs. On well-documented drugs like tagrisso, base models such as Qwen use parametric knowledge to repeatedly name the same leading indication, which is often correct.

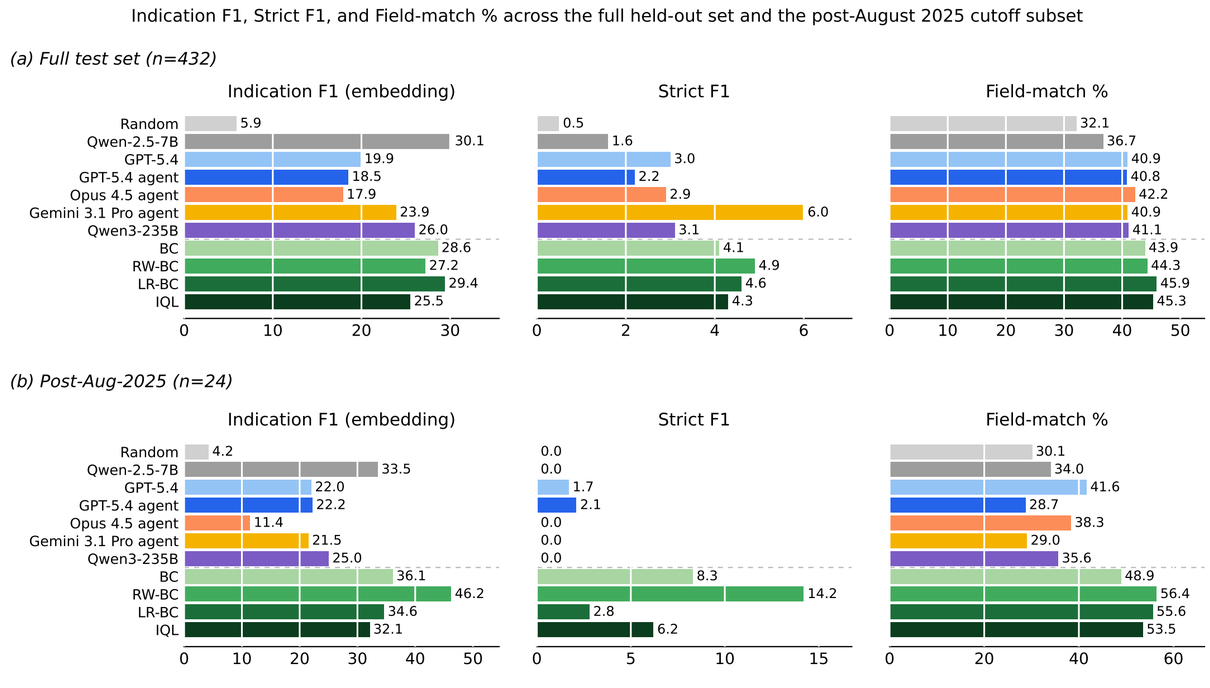


Figure 1. Results are reported as windows-weighted means across the eight held-out evaluation splits.

- Only the reward-trained models (LR-BC, RW-BC) reproduce the designs of trials that succeeded more than those that failed.

Table 1. Success versus failed trial design matches on the five single-drug holdouts.

Model	success%	failed%	discrim [95% CI]
LR-BC	18.9	7.9	$+\mathbf{11.0}\ [-0.5, 23.0]$
RW-BC	20.8	17.1	$+\mathbf{3.7}\ [-9.9, 18.4]$
IQL	18.9	23.7	$-4.8\ [-19.3, 9.1]$
BC	15.1	25.0	$-9.9\ [-23.1, 4.4]$
GPT-5.4 agent	7.5	25.0	$-17.5\ [-29.1, -5.4]$
Qwen-2.5-7B	9.4	48.7	$-39.3\ [-53.0, -25.0]$

Advanced Research + Invention Agency ARIA
Department for Science, Innovation & Technology
pillar
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