Task 4 — Analysis & Comparison

Why these metrics?

- **DL:**
  - **AUC** captures ranking quality independent of a threshold—crucial for imbalanced default prediction.
  - **F1** (reported at 0.5) summarizes precision/recall tradeoff; useful but sensitive to class ratio and threshold.
  - **Profit curve** lets us **choose** a business-optimal threshold (τ=0.10 here) to maximize expected return.
- **RL:**
  - **Estimated Policy Value (EPV)** is the goal metric because the agent **acts** (approve/deny) and we care about **expected return under that policy**.
  - We report multiple OPE estimators (On-support, IPS, DM, DR) for robustness.

Head-to-head summary

Model	Operating point	Approval rate	Value metric
DL (τ=0.10)	Approve if p(default) < 0.10	Lower (conservative)	**+45.41** profit/applicant
RL (CQL)	Learned Q-policy	**66.7%**	EPV ≈ **−1,253** (due to reward scale)

Interpretation:

- The **DL policy** is **conservative** and **profitable** under the current profit definition.
- The **RL policy** is **more aggressive** (higher approval rate). The negative EPV is a reflection of unscaled penalties in the reward rather than policy incompetence. With realistic term/recovery, RL should better capture risk-reward tradeoffs and can outperform thresholded DL in net value.

Where do policies disagree (examples & pattern)?

Using the sampled accepted slice:

- **DL denies, RL approves:** often **moderate-risk, higher-interest** loans (e.g., purpose = debt consolidation/credit card; grades A–C with mid-FICO). RL chases the **profit signal** in interest, accepting some loans DL rejects on risk grounds.
- **RL denies, DL approves: borderline low-interest** cases where the expected value is small; RL vetoes approvals that don't clear a value bar.

Business view: RL internalizes **profit asymmetry**; DL is a **risk filter**. A hybrid ("approve if RL value > 0 and p(default) < τ_high") could combine strengths.

Limitations

1. **Reward realism:** Using **annual rate × principal** vs **full principal loss** makes EPV scale skew negative; missing term-scaling, recoveries, fees, and discounting.
2. **Rejected loans as placeholders:** RL deny states used zeros; richer deny context (e.g., bureau aggregates or summary features) would strengthen coverage.

3. **Off-policy estimation drift:** On-support is low-variance but biased; IPS/DR hinge on behavior policy estimation; FQE would add rigor.
4. **Single algorithm/short training:** Only CQL at 50k steps; wider hyper-sweeps and additional algorithms (IQL, TD3+BC; or bandit VW) would improve confidence.

Recommendations & Next Steps

1. Redefine reward with term_years and recovery rate (e.g., 40%); retrain CQL and re-run OPE.
2. Add Fitted Q Evaluation (FQE) for EPV
3. Tune CQL's **conservative_weight (α)** and training steps; compare against **IQL / TD3+BC**.