🧬 SAR LANG LAB Multi-Tier Evaluation Framework

Status	In progress

> **Project Summary**
>
> The **PharmaSwarm–SAR LANG LAB Multi-Tier Evaluation Framework** was developed during an advanced **Data Science and Machine Learning Bootcamp** to bridge the gap between AI reasoning and experimental science in **drug discovery**.
>
> While current AI models can predict promising molecules, they often fail to explain *why* or *how* those predictions hold up under experimental validation. This framework introduces a structured, reproducible method to evaluate whether **large language model (LLM)** outputs are both mechanistically sound and scientifically testable.
>
> The system integrates two components: **SAR-LANG LAB (Tiers 0–2)**, which assesses reasoning accuracy and causal structure in *Structure–Activity Relationship* (SAR) analysis, and **PharmaSwarm (Tier 3)**, which validates hypotheses through *in silico* and *in vitro* experiments. Together, they create an evidence-based pipeline demonstrating how AI can move beyond prediction to deliver explainable, verifiable contributions to real-world drug discovery.

Tier 0 – SAR Report Format & Evaluation Rubric

Goal: Standardize and quantitatively evaluate model-generated SAR reports before higher-tier validation.

- **Format:** SAR Report V4 (structured template with visual and reference sections)
- **Evaluation Rubric V2:** Scores reports across **Accuracy, Comprehensiveness, Clarity, Reasoning, Evidence**

Dimension	Criteria	Metric	Score	Remarks
Accuracy	Match with reference report	Domain accuracy	10	Count × 2
	Quantitative errors (SMILES, pKi, etc.)	– 5 (Penalty)	Deduct per error	
Comprehensiveness	Template compliance (V4)	Structure adherence	5	—
Clarity & Communication	Visualization + highlights	Completeness	5	—
Reasoning & Evidence	RAGAS metrics (≥ 0.7 each)	Faithfulness / relevance	15	× 3
	Logical causal structure	Quantified reasoning	10	× 2
	Literature citation	Citation completeness	5	—
Total			**50 pts**	

Experimental Step:

Select one *Activity Cliff* (e.g., *(R)- vs (S)-thalidomide*), generate 5 reports each for Base, Tier 1, Tier 2 models, and evaluate using the rubric.

Tier 1 – Agent Performance Evaluation

Agents: `base_agent`, `biological_agent`, `futurehouse_agent`, `structural_agent`

Benchmark: Five Activity Cliffs evaluated across vendors (`openai`, `gemini`, `futurehouse`).

Parameters (Periodically Retuned):

Task	Temperature	Max Tokens	Purpose
Accuracy Evaluation	0.0 – 0.2	500 – 1000	Deterministic replication
Hypothesis Generation	0.6 – 0.8	1500 +	Creative reasoning
Automated Reporting	0.3 – 0.5	2000 +	Balanced narrative

Stage-Specific Guide

Stage	Task	Temp.	Tokens	Rationale
1	Activity Cliff Detection	0.0–0.2	100–300	Stable numeric outputs
2	LLM Hypothesis Generation	0.6–0.8	500–1000	Mechanistic reasoning
3	Report Generation	0.3–0.5	1000–2000	Structured summary
4	Brainstorming / Novel Idea Exploration	0.7–1.0	1500–2500	Creative divergence

Other Controls

- **Safety:** Dictionary-based taboo filtering → automatic penalty/rejection
- **Evidence:** Check PMID/DOI links retrieved via RAG for proper citation
- **Weighting:** Expert consensus tunes `embedding / ROUGE vs form / domain / safety`

(RAG evaluation itself handled in Tier 2.)

Tier 2 – Utility.app (RAG / RAGAS Evaluation)

Objective:

Measure how well retrieval (PubMed via RAG) supports generated reasoning.

- **Compare:** RAG On vs Off
- **Vary:** k = 1, 5, 10

```
params = {
    'db': 'pubmed',
    'term': search_term,
    'retmax': max_results,  # number of top papers fetched
    'sort': 'relevance'
}
```

Setting	Effect
`max_results = 1`	Highest precision (one paper)
`max_results = 5`	Balanced recall/precision
`max_results = 20`	Highest recall + noise risk

↑ `retmax` → ↑ Recall, ↓ Precision

Composite Metric:

Integrate RAGAS scores (`context_precision`, `recall`, `faithfulness`, `answer_relevancy`) into a weighted index for final Tier validation.

→ Ref: Judy-Choi RAGAS Tutorial

Tier 3 – PharmaSwarm Hypothesis Validation (4 Tiers)

> SAR-LANG LAB feeds into PharmaSwarm's scientific validation pipeline for drug discovery LLMs.

PharmaSwarm Tier	Stage Title	Objective	Methods / Techniques	Metrics	Key Question
1	Retrospective Benchmarking	Recreate known drug discoveries with period data	Historical case reconstruction (e.g., oncology)	Recall@K, Precision@K, Kendall's Tau, MAP	Can the system rediscover past success cases?
2	Prospective *In Silico* Assessment	Validate new targets/compounds computationally	Docking (Vina, Glide), MD (50–100 ns), ADMET (pkCSM, ADMETlab)	Docking affinity, MD stability, ADMET concordance	Are hypotheses physically and pharmacologically sound?
3	Experimental Evaluation	Wet-lab verification	SPR/ITC (Kd), IC_{50} assay, *in vivo* model	Kd < 100 nM, IC_{50} < 1 μM = "hit"	Do predicted compounds show measurable efficacy?
4	Expert User Studies	Human-in-the-loop evaluation	Comparative user studies vs standard workflow	Time savings, novelty/plausibility, Wilcoxon test	Does PharmaSwarm improve scientific productivity and credibility?

Integrated View

Pipeline Layer	System / Module	Validation Focus
Tier 0–2	**SAR-LANG LAB**	Report accuracy → Agent tuning → Retrieval reasoning
Tier 3	**PharmaSwarm**	Scientific validation (*in silico*, experimental, expert)

Together they form a **closed-loop evaluation ecosystem**—from LLM-generated SAR reasoning to computational and experimental verification.