Contents

Evaluation Framework

Overview

Framework sistematico per valutare e comparare architetture agentiche attraverso metriche multi-dimensionali.

Dimensioni di Valutazione

1. Efficacia (Task Success)

Metriche: - **Success Rate**: % task completati correttamente - **Partial Success Rate**: % task parzialmente completati - **Failure Rate**: % task falliti completamente

Misurazione:

```
Success Rate = (Fully Completed Tasks) / (Total Tasks)

Stratified per:
- Complessità task
- Dominio
- Input variability
```

2. Efficienza

Metriche: - **Latenza**: p50, p95, p99 response time - **Costo**: $ per task - **Resource Usage**: Token, API calls, compute time

Misurazione:

```
Efficiency Score = Success Rate / (Cost * Latency)

Higher is better
```

3. Robustezza

Metriche: - **Error Recovery Rate**: % errori recuperati automatically - **Graceful Degradation**: Performance sotto fault conditions - **Variance**: Stabilità output tra runs

Misurazione:

```
Robustness = (Recovered Errors + Degraded Success) / Total Errors

Test under:
- Tool failures
- Network issues
- Invalid input
- Resource constraints
```

4. Estensibilità

Metriche: - **New Task Adoption**: Effort per aggiungere nuovo task type - **Tool Integration**: Effort per aggiungere tool - **Domain Transfer**: Success rate su nuovo domain

Misurazione:

```
Extensibility = Success_new / Success_original

Dopo:
- Adding N new tools
- Applying to new domain
- Handling new task types
```

5. Tracciabilità

Metriche: - **Decision Explainability**: % decisioni spiegabili - **Replay Capability**: % esecuzioni riproducibili - **Audit Completeness**: % informazioni tracciabili

Misurazione:

```
Traceability Score = (Explainable + Reproducible + Auditable) / 3

Binary per task: yes/no
```

Benchmark Suite

Task Categories

Simple (Classe A): 1. Question Answering 2. Summarization 3. Simple Classification 4. Content Generation

Complex (Classe B): 5. Multi-step Problem Solving 6. Code Generation 7. Data Analysis 8. Research Tasks

Critical (Classe C): 9. Medical Decision Support (simulated) 10. Financial Analysis 11. Legal Document Analysis

Real-Time (Classe D): 12. Rapid Classification 13. Real-time Recommendations

Collaborative (Classe E): 14. Multi-agent Planning 15. Distributed Problem Solving

Evaluation Protocol

```
For each architecture A and task set T:

1. Run A on T (N=100 runs per task)

2. Measure all metrics

3. Statistical analysis:
   - Mean, median, std dev
   - Confidence intervals
   - Significance tests vs baselines

4. Report:
   - Overall scores
   - Per-category breakdown
   - Failure analysis
```

Comparative Methodology

Baseline Architectures

Compare against: 1. **Zero-Shot LLM**: Single LLM call, no structure 2. **ReAct**: Standard ReAct implementation 3. **LangChain Agent**: Off-the-shelf LangChain 4. **Custom Baseline**: Domain-specific comparison

Statistical Significance

```
Required for claims:
- N >= 100 per condition
- p < 0.05 (t-test or Mann-Whitney)
- Effect size reported (Cohen's d)
- Confidence intervals shown
```

Reporting Standards

Required Information: - Architecture details (components, configuration) - LLM used (model, version, parameters) - Hardware (compute, memory) - Task dataset (size, source, difficulty) - Metrics (all 5 dimensions) - Statistical analysis - Failure cases analysis

Validation Checklist

- ☐ Success rate measured on benchmark
- ☐ Latency distributions reported
- ☐ Cost per task calculated
- ☐ Error recovery tested
- ☐ Extensibility validated (new domain)
- ☐ Trace ability verified
- ☐ Compared against baselines
- ☐ Statistical significance confirmed
- ☐ Failure cases analyzed
- ☐ Limitations documented

Example Report Template

```
# Architecture Evaluation Report

## Architecture
- Name: [Pattern name]
- Components: [List]
- LLM: [Model]
- Configuration: [Key params]

## Results

### Efficacia
- Success Rate: X% (CI: [low, high])
- vs Baseline: +Y% (p=0.0Z)

### Efficienza
- Latency p95: X seconds
- Cost: $X per task

### Robustezza
- Error Recovery: X%
- Variance: =X

### Estensibilità
- New domain success: X%
- Tool integration effort: X hours
```

```
### Tracciabilità
- Explainability: X%
- Reproducibility: X%

## Failure Analysis
- Top failure modes: [List]
- Root causes: [Analysis]

## Conclusions
- Strengths: [List]
- Weaknesses: [List]
- Recommended use cases: [List]
```

Continuous Evaluation

Production Monitoring: - Real-time success rate tracking - Latency monitoring - Cost tracking - Error rate alerts - Drift detection (performance degradation)

Next: 06-open-questions.md -> Research limitations and future directions