Task #80

Research Metadata Parameters for Feedback

Under Task #75 Applicability of Dimensions for LLMs

Author: Prerana Kumsi

Objective

Define a metadata schema for hybrid evaluation of response quality:

- **LLM-based** → subjective dimensions
- **Rule-based** → objective dimensions

Literature shows:

- LLM judges align ~80–90% with humans on subjective quality.
- Rule-based systems are deterministic for measurable checks.
- Hybrid evaluation is most reliable.

Dimension	Method
Tone	LLM
Helpfulness	LLM
Clarity	LLM
Relevance	LLM
Instruction Following	LLM

Length	Rule
Format	Rule
Keywords	Rule
Privacy / PII	Rule

Required Metadata

Response-Level

- `response_id`
- `prompt_id`
- `timestamp`
- `generator_model`
- `judge_model`
- `evaluation_method`

LLM Dimensions (store per dimension)

- `score`
- `confidence`
- `rationale`

Rule-Based Checks

- `token_count`
- `format_valid`
- `schema_errors`
- `pii_detected`
- `keyword_match_rate`

Aggregation

- Weighted sum of LLM scores
- Penalty for rule violations
- Hard fail for PII

Fields:

- `overall_score`
- `critical_failure`
- `requires_human_review`

JSON Schema

```
{
 "response_id": "string",
 "evaluation_method": "Hybrid",

 "rule_based": {
   "token_count": 0,
   "format_valid": true,
   "pii_detected": false,
   "keyword_match_rate": 1.0
 },

 "llm_evaluation": {
   "tone_score": 0,
   "helpfulness_score": 0,
   "clarity_score": 0,
   "relevance_score": 0,
   "instruction_following_score": 0,
   "confidence": 0.0
 },

 "overall_score": 0,
 "critical_failure": false
}
```

Conclusion:

The schema encodes the research finding that LLMs evaluate subjective quality, rule-based systems enforce structure, and hybrid evaluation ensures reliable feedback.