Contents

ADR-003: Tool Grounding with 20%+ External Verification

Status

Accepted

Context

LLM vs. LLM comparison (having one model verify another) measures **consistency**, not **truth**. This creates a dangerous failure mode:

1. Model A generates a plausible-sounding hallucination
2. Model B (or A again) confirms it sounds reasonable
3. The hallucination gets reinforced in memory
4. Future queries retrieve and build upon the hallucination
5. **Hallucination cementing**: False beliefs become entrenched

Without external grounding, Cato's curiosity becomes a **hallucination amplifier** rather than a learning mechanism.

Evidence

- GPT-4 self-consistency: ~85% (agrees with itself on hallucinations)
- Claude self-consistency: ~82%
- Cross-model agreement on hallucinations: ~70%

These numbers mean **most hallucinations pass LLM-only verification**.

Decision

Mandate that **at least 20% of curiosity loops must verify against external reality** through tool use:

Grounding Tools

Tool	Purpose	Use Case
Web Search	Factual verification	“Is X true?” queries
Code Execution	Computational verification	Math, algorithms, data analysis
API Calls	Real-time data	Weather, stocks, current events
Database Queries	Structured data	Historical records, statistics
Document Retrieval	Source verification	Citations, quotes, references

Grounding Policy

```
class GroundingPolicy:
    """Determines when to use external grounding."""

    ALWAYS_GROUND = [
        "factual_claim",      # "The population of X is Y"
        "numerical_claim",    # "X costs $Y"
        "temporal_claim",     # "X happened in Y"
        "attribution",        # "X said Y"
        "scientific_claim",   # "Studies show X"
    ]

    SAMPLE_GROUND = [
        "general_knowledge",  # 20% sampling
        "reasoning_chain",    # 10% sampling
        "creative_content",   # 5% sampling
    ]

    NEVER_GROUND = [
        "opinion",            # "I think X"
        "hypothetical",       # "If X then Y"
        "meta_statement",     # "I'm uncertain about X"
    ]
```

Architecture

```
                  Curiosity Question
  "What is the GDP of France in 2024?"




                  Claim Classifier
  Type: factual_claim, numerical_claim
  Decision: MUST_GROUND
```

```
                  LLM Prediction
"France's GDP in 2024 is approximately $3.1 trillion"
```

```
                  Tool Grounding
Tool: Web Search (IMF, World Bank, Statista)
Result: "$2.78 trillion (IMF 2024 estimate)"
```

```
                  NLI Comparison
Prediction vs. Ground Truth
Result: PARTIAL_MATCH (order of magnitude correct)
Surprise Score: 0.4
```

```
                  Memory Update
Store corrected fact with source attribution
Mark original prediction as "needs_update"
```

Implementation

Tool Executor Service

```typescript
interface GroundingResult {
  tool: string;
  query: string;
  result: string;
  sources: string[];
  confidence: number;
  timestamp: Date;
}

class ToolGroundingService {
  private readonly webSearch: WebSearchClient;
  private readonly codeExecutor: CodeExecutionClient;
  private readonly apiClient: ExternalAPIClient;
  async ground(
    claim: string,
    claimType: string
  ): Promise<GroundingResult> {
    // Select appropriate tool
    const tool = this.selectTool(claimType);

    // Execute grounding
    switch (tool) {
      case 'web_search':
        return this.groundWithWebSearch(claim);
      case 'code_execution':
        return this.groundWithCode(claim);
      case 'api_call':
        return this.groundWithAPI(claim);
      default:
        throw new Error(`Unknown tool: ${tool}`);
    }
  }

  private async groundWithWebSearch(
    claim: string
  ): Promise<GroundingResult> {
    // Generate search query from claim
    const query = await this.generateSearchQuery(claim);

    // Execute search
    const results = await this.webSearch.search(query, { limit: 5 });

    // Extract relevant facts
    const facts = await this.extractFacts(results, claim);

    return {
      tool: 'web_search',
      query,
      result: facts.summary,
      sources: facts.sources,
      confidence: facts.confidence,
      timestamp: new Date()
    };
  }
}
```

Grounding Budget

To prevent excessive API costs, grounding has its own budget:

Tool	Cost per Call	Daily Limit	Monthly Budget
Web Search	$0.01	1,000	~$300
Code Execution	$0.001	5,000	~$150
API Calls	Varies	500	~$100
Total			**~$550/month**

Consequences

Positive

- **Hallucination prevention**: External reality check breaks confirmation loops
- **Source attribution**: All facts traceable to external sources
- **Confidence calibration**: Grounding provides ground truth for calibration
- **User trust**: Can cite sources when asked

Negative

- **Higher latency**: Tool calls add 500ms-2s per grounding
- **Additional cost**: ~$550/month for grounding tools
- **Complexity**: Tool integration and error handling
- **Rate limits**: External APIs have usage limits

Metrics

Track grounding effectiveness:

Metric	Target	Description
Grounding Ratio	20%	% of curiosity loops with tool grounding
Correction Rate	30%	% of LLM predictions corrected by grounding
Source Coverage	80%	% of facts with external source attribution
Hallucination Rate	10%	% of responses containing unverified claims

References

- TruthfulQA: Measuring How Models Mimic Human Falsehoods
- Tool-Augmented Language Models
- Retrieval-Augmented Generation