FACTKG Methodology: Multi-Approach Verification Pipeline

Legend:
- Data/Input
- Symbolic Methods
- Neural Encoders
- LLM Methods
- Results

FACTKG Dataset
108,675 claims from DBpedia

Train 86,367 | **Val** 13,267 | **Test** 9,041

Symbolic Baseline | **Graph Neural Networks** | **BERT Baseline** | **LLM Approaches**

Ontology-Aware Features (31) | **QA-GNN Architecture** | **Linearize KG Subgraphs** | **LLM Semantic Filtering** | **LLM Direct Reasoning**

Graph Features:
• Degree stats
• Relation diversity
• Entity coverage

Components:
• RoBERTa encoder
• Relational GNN
• Message passing

Input Format:
[CLS] Claim [SEP]
Triple₁ ... Tripleₙ

GPT-4.1-mini:
Select top-10 relevant triples

300 Test Claims:
• GPT-4o-mini
• GPT-4.1-mini

Classifiers:
• Logistic Reg
• Random Forest
• XGBoost

Variants:
• Baseline QA-GNN
• + Cross-attention

Fine-tune:
• BERT-base
• Max 512 tokens
• Binary classifier

Train BERT on:
• 9,706 filtered
• 9,706 unfiltered

Two Conditions:
① Memorization (claim only)
② KG-grounded (claim + evidence)

Symbolic Ceiling: 63.96% Acc | **GNN Result:** ~70% Acc | **BERT Result:** 92.68% Acc | **Filtered: 78.85% vs Unfiltered: 52.70%** | **KG-grounded:** 84.00-84.33% vs Mem: 71-75%

Comparative Analysis Across Reasoning Types

Per-Reasoning-Type Evaluation:
Single-hop | Multi-hop | Multi-claim | Existence | Negation | Substitution

Key Findings:
✓ Symbolic ceiling at 63.96% (ontology rules insufficient)
✓ GNNs underperform (~70%) despite graph structure
✓ BERT achieves 92.68% via neural semantics over linearized KGs
✓ LLM semantic filtering: +26% gain (78.85% vs 52.70%)
✓ KG-grounded LLMs outperform memorization (+9-13%)

Conclusion: Neural semantics + explicit KG grounding essential for robust verification