Dynamic Construction of Causal Knowledge Graphs for Scientific Reasoning in Search Agents

Research Proposal

F. Sun

December 15, 2025

Outline

Recent Advances in Search Agents

Recent LLM-based search agents have achieved impressive results:

Multi-step Search & Reasoning:

- Tongyi DeepResearch (2025)
- WebDancer (2025)
- MaskSearch (2025)

✓ Complex multi-step planning
✓ Long trajectory training
✓ Iterative refinement

KG-Enhanced Search:

- PaSa (2025): Citation network
- DynaSearcher (2025): Wikidata
- CausalKG (2021): Causal relations

✓ Structured knowledge
✓ Improved retrieval
✓ Rich representations

Question

These systems work well for general QA. **What about scientific reasoning?**

The Unique Challenges of Scientific Reasoning

Scientific questions require more than information retrieval:

Example: Vitamin D and COVID-19

User asks: "Does vitamin D prevent severe COVID-19?"

Current systems might say:

> *"Studies show vitamin D levels are associated with COVID-19 severity..."*

Problem: Association $\neq$ Causation

Scientific reasoning requires:

- Distinguish **correlation** from **causation**
- Check for **reverse causation**: Severe patients $\rightarrow$ hospitalised $\rightarrow$ less sunlight $\rightarrow$ low vitamin D
- Evaluate **evidence type**: RCTs show no effect, observational studies show correlation
- Assess **evidence quality**: GRADE framework
- Quantify **effect size** with uncertainty: RR = 0.95, 95%CI: [0.82, 1.10]

What Current Systems Do Well

Tongyi DeepResearch / WebDancer

Strengths:

- ✓ Multi-step search
- ✓ Agent orchestration
- ✓ Long trajectory handling
- ✓ Iterative refinement

Limitations for science:

- × No causality distinction
- × No evidence grading
- × LLM black-box reasoning
- × No structured representation

PaSa / DynaSearcher

Strengths:

- ✓ KG-enhanced retrieval
- ✓ Structured knowledge
- ✓ Domain-specific search

Limitations for science:

- × Static, pre-built KGs
- × General relations, not causal
- × No evidence assessment
- × Focus on retrieval, not reasoning

Standing on the Shoulders of Giants

Key Insight

These systems provide 80% of what we need (search, agents, KG).
We need the remaining 20%: **causal reasoning with evidence assessment**.

The Gap: Scientific Causal Reasoning

Capability	Tongyi	PaSa	CausalKG	Needed
Multi-step search	✓	✓	—	✓
KG-enhanced	×	✓	✓	✓
Causal relations	×	×	✓	✓
Causation vs correlation	×	×	×	✓
Evidence quality (GRADE)	×	×	×	✓
Literature-driven KG	×	×	×	✓
Dynamic KG construction	×	×	×	✓

Why existing KG approaches don't work:

- **PaSa/DynaSearcher:** Use static, general KGs (Wikidata) — no causal relations, no evidence grading
- **CausalKG:** Data-driven (learns from observational data) — not literature-driven, no evidence assessment

Core Problem

Core Problem

No existing system dynamically constructs causal KGs from scientific literature with evidence assessment.

Research Question

How to build an agent that dynamically constructs causal knowledge graphs from scientific literature to enable evidence-based reasoning?

Key distinction from prior work:

NOT our approach:

- Pre-build a large causal KG
- Store all scientific knowledge
- Query static database

→ This is database engineering

Our approach:

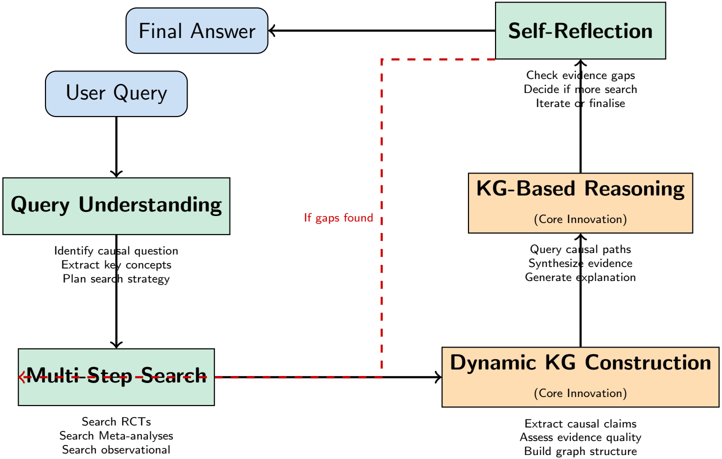
- User asks a question
- Agent searches literature
- **Dynamically builds KG** from results
- Reasons over temporary KG

→ This is agent research

Three Research Challenges

1. **On-the-Fly Causal KG Construction**
   - How to extract causal relations from retrieved papers in real-time?
   - How to represent effect sizes, conditions, and evidence sources?
   - How to handle contradictory studies?
2. **Evidence-Guided Structured Reasoning**
   - How to automatically assess evidence quality (GRADE framework)?
   - How to reason over the KG to synthesize conclusions?
   - How to distinguish causation from correlation?
3. **Causal-Aware Search Strategy**
   - How to identify when a query requires causal reasoning?
   - How to plan searches for different evidence types (RCT vs observational)?
   - How to iteratively refine the KG based on gaps?

System Overview: Dynamic Causal KG Agent



Key difference from existing agents: Intermediate structured representation (KG) enables verifiable, evidence-grounded reasoning

Example Walkthrough: "Does Vitamin D Prevent COVID-19?"

Phase 1: Query Understanding

- Recognise: Causal question
- Extract: Intervention = Vitamin D, Outcome = COVID-19 severity
- Plan: Search for RCTs, Meta-analyses, Observational studies

Phase 2: Multi-Step Search

- Search 1: "vitamin D COVID-19 RCT" $\rightarrow$ 3 papers
- Search 2: "vitamin D COVID-19 meta-analysis" $\rightarrow$ 2 papers
- Search 3: "vitamin D COVID-19 observational" $\rightarrow$ 5 papers

Phase 3: Dynamic KG Construction (Core Innovation)

Extract from each paper:

- RCT 1: RR=0.98, CI:[0.85,1.12], n=500
- Meta: RR=0.95, CI:[0.82,1.10], 7 RCTs
- Obs: Correlation r=-0.45, n=5000

Build temporary KG:

- Node: VitaminD, COVID_Severity
- Edge: causal_effect_unclear
- Evidence: RCT (High), Obs (Low)
- Conclusion: Insufficient evidence

Example Continued

Phase 4: KG-Based Reasoning (Core Innovation)

1. **Query KG:** Does VitaminD causally reduce COVID_Severity?
2. **Check evidence types:**
   - High-quality evidence (RCT, Meta): No significant effect
   - Low-quality evidence (Observational): Shows correlation
3. **GRADE Assessment:**
   - Initial grade: High (RCTs available)
   - Downgrade: Effect not significant (CI crosses 1.0)
   - Final grade: Low evidence of causal effect
4. **Check alternative explanations:**
   - Reverse causation: Severe patients → hospitalized → low vitamin D
   - Confounder: Healthy lifestyle → vitamin D + immunity

Phase 5: Self-Reflection

- Evidence sufficient? Yes (RCTs + Meta available)
- Need more search? No
- Generate final answer

KG Representation (Using RDF*) I

Inspired by CausalKG, but extended for literature-driven construction:

```
<<:VitaminD :causal_effect :COVID_Severity>>
    # Effect quantification
    :effectSize ''RR=0.95''^^xsd:float ;
    :confidenceInterval ''[0.82, 1.10]'' ;
    :statisticalSignificance ''p=0.48'' ;

    # Evidence sources (key extension)
    :supportedBy [
        :Study_RCT_2023 [
            :studyType :RandomizedControlledTrial ;
            :sampleSize 500 ;
            :effectSize 0.98 ;
            :qualityScore :High
        ],
        :Study_Meta_2024 [
            :studyType :MetaAnalysis ;
            :includedStudies 7 ;
            :effectSize 0.95 ;
            :heterogeneity ''I2=15%'' ;
            :qualityScore :High
        ]
    ] ;

    # Evidence assessment (key extension)
    :evidenceGrade :Low ;
    :gradeJustification ''Effect not statistically significant'' ;
```

KG Representation (Using RDF*) II

```
# Alternative explanations (key extension)
:possibleReverseCausation true ;
:confounders [:Hospitalization, :HealthyLifestyle] ;

# Temporal metadata
:constructedAt ''2024-12-13''^^xsd:date ;
:querySpecific ''Does vitamin D prevent COVID-19?'' .
```

Core Innovation 1: On-the-Fly KG Construction

Why not pre-build a KG?

- Scientific literature: 30M+ papers in PubMed alone
- User queries: Long-tail distribution, impossible to anticipate
- Knowledge updates: New papers published daily

Our approach: Build KG dynamically for each query

Aspect	**Pre-built KG**	**Dynamic KG (Ours)**
Coverage	Limited, static	Query-specific, focused
Freshness	Outdated	Includes latest papers
Scalability	Need to process all papers	Only process relevant papers
Feasibility	Infeasible for 30M+ papers	Feasible (10-20 papers/query)

Technical challenges:

- Fast extraction: Quickly build KG from 10-20 papers
- High accuracy: Causal claims, effect sizes, study types
- Conflict resolution: Handle contradictory studies

Core Innovation 2: Evidence-Guided Reasoning

Why not just use LLM to synthesize?

LLM Black-box Reasoning:

- × Cannot verify reasoning steps
- × May "hallucinate" evidence
- × Citation accuracy issues
- × Unclear evidence weighting

KG-Based Reasoning:

- ✓ Explicit reasoning paths
- ✓ Traceable to sources
- ✓ Verifiable evidence chain
- ✓ Systematic GRADE scoring

GRADE Framework Integration:

1. **Initial grading:** RCT/Meta = High, Observational = Low
2. **Downgrade factors:**
    - Risk of bias (study quality)
    - Inconsistency (heterogeneity across studies)
    - Indirectness (different populations/outcomes)
    - Imprecision (wide confidence intervals)
3. **Upgrade factors:** Large effect, dose-response gradient
4. **Final grade:** High / Moderate / Low / Very Low

Core Innovation 3: Causal-Aware Search

Different from general search agents:

Aspect	General Agent	Causal-Aware (Ours)
Query analysis	Keywords	**Causality detection**
Search strategy	Broad retrieval	**Evidence type-specific** (RCT, Meta, Observational)
Stopping criteria	Enough info	**Evidence sufficiency** (per GRADE)
Iteration	General refinement	**Gap-driven search** (missing evidence types)

Example search plan for "Does aspirin reduce heart attacks?"

1. Identify: Causal question → Need causal evidence
2. Search 1: "aspirin myocardial infarction RCT" → Find experimental evidence
3. Search 2: "aspirin heart attack meta-analysis" → Find synthesised evidence
4. Check KG: Do we have high-quality evidence? Yes → Stop
5. If no: Search 3 "aspirin MI cohort study" → Lower quality but more coverage

Standing on the Shoulders of Giants

Our work builds on and extends existing research:

Prior Work	What We Borrow	What We Add
Tongyi DeepResearch	Multi-step search framework Session-level RL	+ Causal reasoning + Evidence grading
WebDancer	Tool use (search + browse) Iterative refinement	+ Specialised tools + Evidence extraction
PaSa	Academic search domain KG-enhanced retrieval	+ Content reasoning + Causal KG
DynaSearcher	KG + Doc hybrid retrieval Multi-reward RL	+ Dynamic KG + Literature-driven
CausalKG	Rich causal representation RDF* for complex relations Causal reasoning patterns	+ From literature + Evidence assessment + Dynamic construction

Positioning

80% foundation from prior work + 20% critical extension = Novel contribution

The 20% (causal reasoning, evidence grading, dynamic KG) is essential for scientific reasoning but missing from all existing systems.

Why Simple Extensions Don't Work

Could we just prompt existing systems differently?

Attempt: Enhanced Prompt for Tongyi

"Please distinguish causation from correlation, evaluate evidence quality using GRADE, check for confounders, and quantify effect sizes."

Why this fails:

1. **LLM black-box:** Cannot verify if GRADE was actually applied
   - LLM might output "GRADE: High" without actual assessment
   - No way to check reasoning steps
2. **Lack of structure:** No enforcement of systematic process
   - Prompt is suggestion, not requirement
   - LLM may skip steps or hallucinate
3. **Citation accuracy:** Hard to trace claims to sources
   - LLM may misattribute findings
   - Cannot verify "RR=0.80" came from Paper X

KG solves these: Structured representation forces systematic extraction and enables verification

Evaluation Dataset: CausalReasoningQA

Inspired by LegalSearchQA (L-MARS), build scientific causal reasoning benchmark

Dataset Specs:

- **Size:** 200-300 questions
- **Domain:** Biomedical (Stage 1)
- **Source:** Cochrane reviews
- **Annotation:** Medical experts

Question Types:

- Causal judgment (40%)
- Evidence assessment (30%)
- Conditional queries (20%)
- Conflict detection (10%)

Example Questions:

Type 1: Causal judgment

> *"Does aspirin reduce myocardial infarction risk?"*
>
> *Gold: Established (RR=0.80, GRADE: High)*

Type 2: Evidence assessment

> *"How strong is the evidence that vitamin D prevents COVID-19?"*
>
> *Gold: Low (RCTs show no effect)*

Type 3: Conditional

> *"For age less than 40 without risk factors, does aspirin reduce MI risk?"*
>
> *Gold: No evidence / Unlikely*

Evaluation Metrics

Metric	Definition
Causal Accuracy	Correct classification: Established / Probable / Unlikely / Disproven
GRADE Accuracy	Correct evidence grading: High / Moderate / Low / Very Low
Evidence Completeness	% of high-quality studies cited (Recall of RCTs/Meta-analyses)
Effect Size Accuracy	Correct extraction of RR, OR, CI
Confounder Detection	% of relevant confounders identified
Explanation Quality	Human evaluation: Clarity, correctness, evidence support

Baselines

Baselines:

- GPT-4 (no tools)
- GPT-4 + Web Search (standard agent)
- GPT-4 + Web Search + Static KG (Wikidata)
- **Our System:** GPT-4 + Web Search + Dynamic Causal KG

Expected improvements:

- Causal accuracy: improvements vs GPT-4 baseline
- Evidence completeness: improvements vs single-search baseline

Stage 1: Biomedical Deep Dive

Infrastructure

- Design KG schema (RDF*), Core extraction prompts
- Implement BiomedicalAdapter (UMLS/MeSH integration)

Prototype System

- Implement extraction pipeline (causal claims, effect sizes, study types)
- Implement GRADE assessment module, Build Mini KG (50 papers)

Full System & Data

- Implement KG reasoning module, Conflict detection
- Build CausalReasoningQA (100 questions), Iterate on system

Evaluation & Writing

- Run experiments, Compare baselines
- Analyse results, Draft paper

Stage 2* & 3*: Generalisation

Stage 2*: Validate Transferability

- Select second domain (Materials Science or Social Science)
- Implement domain adapter
- Identify cross-domain patterns vs domain-specific needs
- Refactor core architecture based on learnings

Stage 3*: Abstract Framework

- Extract common causal reasoning patterns
- Design adapter development guide
- Open-source framework + documentation
- Write methodology paper

Algorithm 1: Dynamic KG Construction & Reasoning

Algorithm 1: Dynamic Causal KG Construction and Reasoning

```
Input: user_query (e.g., "Does aspirin reduce heart attack risk?")
Output: answer (conclusion, evidence_grade, explanation, sources)
query_info ← parse_query(user_query);
// query_info = {type: "causal", intervention: X, outcome: Y}
KG ← initialize_empty_graph();
search_plan ← generate_search_plan(query_info);
// search_plan = ["X Y RCT", "X Y meta-analysis", ...]
for each search_query in search_plan do
    papers ← web_search(search_query);
    for each paper in papers do
        study_info ← llm_extract(paper.abstract);
        // Extract: study_type, effect_size, CI, sample_size
        if validate_extraction(study_info) then
            study_info.grade ← assess_grade(study_info);
            // GRADE: High/Moderate/Low/Very Low
            KG.add_relation(query_info.intervention,;
                            query_info.outcome,;
                            study_info);
        end
    end
    if has_sufficient_evidence(KG, query_info) then
        break;
    end
end
relation ← KG.query(query_info.intervention, query_info.outcome);
if relation = null then
    return {conclusion: "No evidence found"};
end
```

Algorithm 2 (Continued): Reasoning Rules

Algorithm 2: Reasoning Rules (continued from Algorithm 1)

```
// Apply reasoning rules
if relation has ≥ 1 RCT with High/Moderate grade then
    if aggregate_effect is significant then
        conclusion ← "Established causal";
    else
        conclusion ← "No causal effect";
    end
else
    if relation has only Low/Very Low grade then
        conclusion ← "Insufficient evidence";
    else
        conclusion ← "Unclear";
    end
end
explanation ← generate_explanation(conclusion, relation);
return {conclusion, relation.overall_grade, explanation, relation.sources};
```

Algorithm 3: GRADE Evidence Assessment

Algorithm 3: GRADE Evidence Quality Assessment

```
Input: study_info (study_type, effect_size, sample_size, ...)
Output: grade ∈ {High, Moderate, Low, Very Low}
if study_info.type ∈ {RCT, Meta-analysis} then
    initial_grade ← 4 ;                                        // High
else
    initial_grade ← 2 ;                                        // Low
end
downgrades ← 0;
// Imprecision (rule-based)
if study_info.sample_size < 100 then
    downgrades ← downgrades + 1;
end
if study_info.CI is wide then
    downgrades ← downgrades + 1;
end
if study_info.effect not significant then
    downgrades ← downgrades + 1;
end
// Risk of bias (LLM-assisted)
bias_assessment ← llm_assess_bias(study_info);
downgrades ← downgrades + bias_assessment.downgrade;
final_grade ← max(1, min(4, initial_grade − downgrades));
grade_map ← {4 : High, 3 : Moderate, 2 : Low, 1 : Very Low};
return grade_map[final_grade];
```

Summary of Contributions

Technical Contributions (System & Methods)

1. **Dynamic KG Construction:** Multi-stage extraction, evidence-aware schema, incremental building
2. **Evidence-Graded Reasoning:** GRADE integration, hybrid rule-LLM, verifiable inference
3. **Causal-Aware Search:** Query classification, evidence type-specific planning, gap-driven iteration

Empirical Contributions (Data & Evaluation)

4. **CausalReasoningQA Benchmark:** 200-300 questions, multi-dimensional annotations
5. **Evaluation Framework:** Beyond accuracy, ablation studies, design validation

Summary of Contributions

Potential Impact (Applications & Extensions)

6. **Clinical & Research Tools:** Decision support, literature review assistance
7. **Extensible Framework:** Domain adapters, open-source, community-driven

Key message: Our contributions lie in **how to effectively combine** existing components (LLMs, search, KG) for scientific reasoning, not merely in training new models.

Potential Risks & Mitigation

Risk	Challenge	Mitigation Strategy
Extraction Accuracy	LLM may hallucinate causal claims or effect sizes	Multi-stage verification: (1) Few-shot extraction (2) Rule-based validation (3) Self-consistency checks (4) Confidence scoring for manual review
GRADE Automation	GRADE requires expert judgment (e.g., indirectness assessment)	(1) Automate objective components (study type, sample size) (2) LLM-assisted subjective components (3) Human-in-loop for ambiguous cases (4) Compare with Cochrane assessments
Speed Requirements	Building KG from 10-20 papers could be time-consuming	(1) Parallel processing of papers (2) Caching of common extractions (3) Progressive disclosure (show partial KG while building)
Benchmark Quality	Need expert annotations (expensive)	(1) Start with Cochrane reviews (pre-annotated) (2) Begin with 100 questions, expand after validation
Generalisation	Focuses on 1-2 domains first; generalisation unproven	(1) Clear positioning: depth-first strategy (2) Design modular architecture from start (3) Document patterns for community extensions (4) Provide adapter development guide

Questions for Discussion

I would greatly appreciate your guidance on these feasibility concerns:

1. **Technical Feasibility & Validation**
   - What accuracy threshold makes the system useful?
     - If LLM extraction achieves 70% accuracy, is KG-based reasoning still valuable?
     - How do we validate that KG actually helps vs pure LLM? (if ablation shows minimal difference?)
2. **Success Criteria**
   - What improvement over baseline constitutes success?
     - +10% causal accuracy? +20%? What's realistic and publishable?
3. **Scope & Backup Plans**
   - If full system proves too ambitious, acceptable simplifications:
     - Focus on evidence grading only (drop full causal reasoning)?
     - Single evidence type (RCTs only)?
   - At what point (month/milestone) should I consider pivoting?

Summary

Problem

Current search agents lack systematic causal reasoning capabilities needed for scientific questions

Proposed Solution

Dynamic Causal KG Agent: Constructs causal knowledge graphs on-the-fly from retrieved literature

- Query-specific, not pre-built
- Literature-driven, not data-driven
- Evidence-graded with GRADE framework
- Structured reasoning, not LLM black-box

Summary

Core Innovations

1. On-the-fly causal KG construction from literature
2. Evidence-guided structured reasoning (GRADE integration)
3. Causal-aware search strategy (evidence type-specific)

Timeline & Milestones

- Stage 1: Biomedical system + benchmark + paper draft
- Stage 2*: Second domain + transferability analysis
- Stage 3*: General framework + methodology paper

Thank you.

Backup: Detailed Example Output

User: "Does aspirin reduce heart attack risk?"

System Output:

Conclusion: Established Causal Relationship
Aspirin reduces myocardial infarction (MI) risk by approximately 20%.
Evidence Summary:

- **Effect Size:** *RR = 0.80 (95% CI: 0.73-0.88)*
- **Evidence Grade:** *High (GRADE assessment)*
- **Based on:** *Meta-analysis of 287 RCTs (n=135,000)*

Applicable Conditions:

- *Age $\geq$ 50 years*
- *Presence of cardiovascular risk factors*
- *Dosage: 75-100 mg/day*
- *Duration: Minimum 6 months*

Important Considerations:

- *Also increases bleeding risk (RR=1.5 for GI bleeding)*
- *Benefit-risk assessment needed for each patient*

Primary Source: *Antithrombotic Trialists' Collaboration (2009)*
DOI: 10.1111/j.1538-7836.2009.03307.x

Backup: Comparison with Existing Systems

Same query to different systems:

GPT-4 (no tools):

"Yes, aspirin is commonly used to reduce heart attack risk. It works by preventing blood clots..."

Issues:

- No effect quantification
- No evidence citation
- No conditions specified
- No risk-benefit discussion

Tongyi-style agent:

"Research shows aspirin reduces heart attack risk. A meta-analysis found significant reduction. However, there are bleeding risks..."

Issues:

- Vague "significant reduction"
- No GRADE assessment
- No systematic evidence review

Our system:

"Established causal: RR=0.80 (95% CI: 0.73-0.88), GRADE: High, based on 287 RCTs. Applicable for age $\geq$50 with risk factors, 75-100mg/day, $\geq$6 months. Note: increases bleeding risk (RR=1.5)."

Advantages:

- Precise effect size + CI
- Evidence grade (GRADE)
- Specific conditions
- Risk-benefit quantified
- Source traceable via KG

Backup: Why This is LLM/Agent Research

Core technical challenges are all LLM/Agent-related:

1. **Few-shot Information Extraction**
   - Extract structured causal information from unstructured text
   - Challenge: Achieve high accuracy with minimal examples
   - Techniques: CoT prompting, self-consistency, verification
2. **LLM-Assisted Evidence Assessment**
   - Automate GRADE scoring components
   - Challenge: Match expert judgment
   - Techniques: Reasoning chains, multi-step verification
3. **Agent Orchestration**
   - Multi-step planning, execution, reflection
   - Challenge: When to search more vs conclude
   - Techniques: ReAct, self-critique, iterative refinement
4. **Structured Reasoning**
   - Reasoning over KG structure
   - Challenge: Combine symbolic (KG) and neural (LLM)
   - Techniques: Neuro-symbolic integration
5. **Explanation Generation**
   - Generate human-readable explanations from KG
   - Challenge: Clarity + evidence grounding
   - Techniques: Template-based + LLM generation