GW TAI
TRUSTWORTHY AI INITIATIVE

Information Synthesis Workflows with Large Language Models:
A Review of Current Practices, Opportunities, and Challenges in Literature Review Tasks

Bryce Huffman (presenter),
Dr. Caitlin Grady & Dr. Zoe Szajnfarber

AI4SE & SE4AI Workshop 2025
September 17, 2025
Engineering Management and Systems Engineering
The George Washington University

Agenda

Introduction and motivations

Methods

Results and key findings

Wrap-up

Advancements in AI Large Language Models have led to rapid increases in commercial, purpose-built AI tools for unstructured text workflows

1 GENERAL MODELS

- Public availability catalyzed in late-2022
- Ongoing advancements

ANTHROP\C
OpenAI
Gemini

2 OPEN QUESTIONS

- How is their potential harnessed?
- Where could they be used?
- What are they good for?
- What are they not good for?

3 PROPOSAL WRITING?

- Common across industries
- Core challenge is unstructured text processing and synthesis
- Direct similarities to SE artifact generation and other related tasks

Exploding market of commercial LLM tools claiming to support proposal writing workflows

AutogenAI
AutoRFP.ai
govdash.com
GOVEAGLE
GovSignals
grant assistant by FREEWILL
Grantable
grantboost
inventive
Loopio
RSci.AI
responsive
sweetspot

/ IDENTIFIED SET

GW TAI TRUSTWORTHY AI INITIATIVE

Note: Identified commercial platforms restricted to those involving RFI/RFP/RFx and/or grant/proposal writing purposes; set is non-exhaustive
Sources: [1]–[17]

Agenda

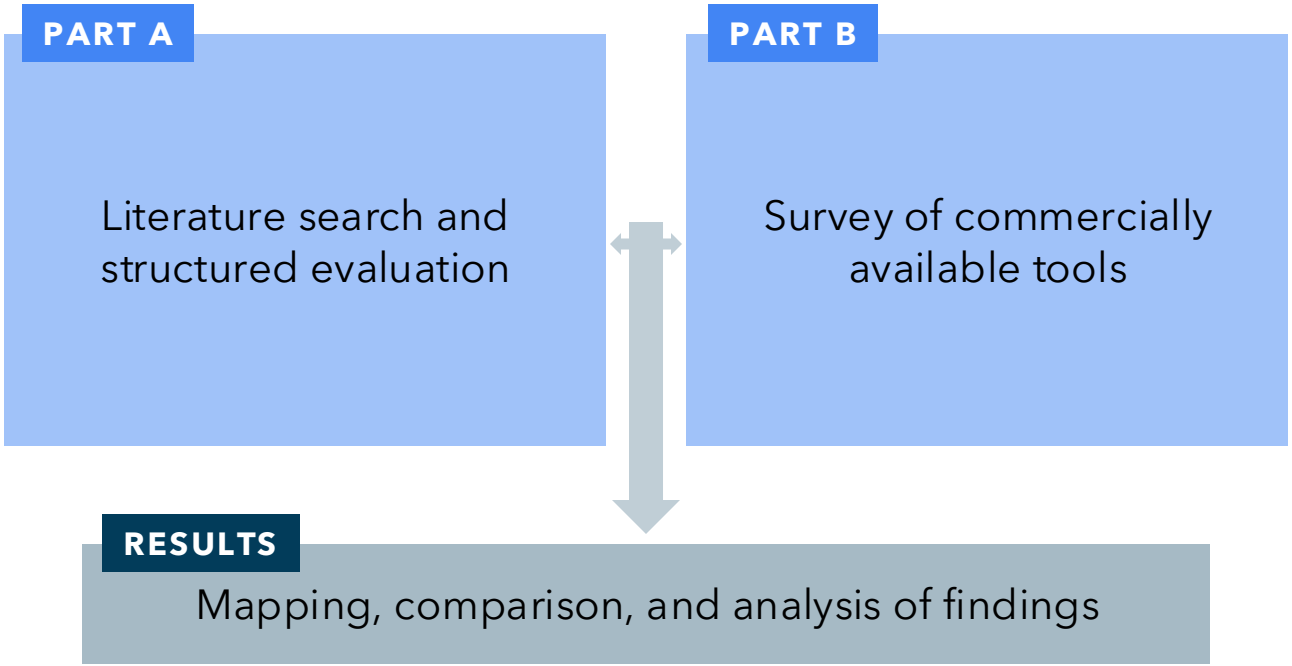
Introduction and motivations

Methods

Results and key findings

Future research

This study employed a 2-part research design



Agenda

Introduction and motivations

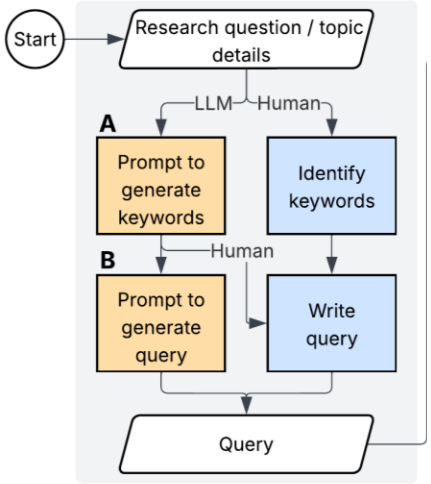
Methods

Results and key findings

Future research

Five core tasks emerge in the LLM-supported literature review workflow for unstructured text synthesis, based on a review of 22 studies

1: Keyword generation and query formulation



KEY

LLM | Human

2: Query search execution

3: Article screening

4: Data extraction

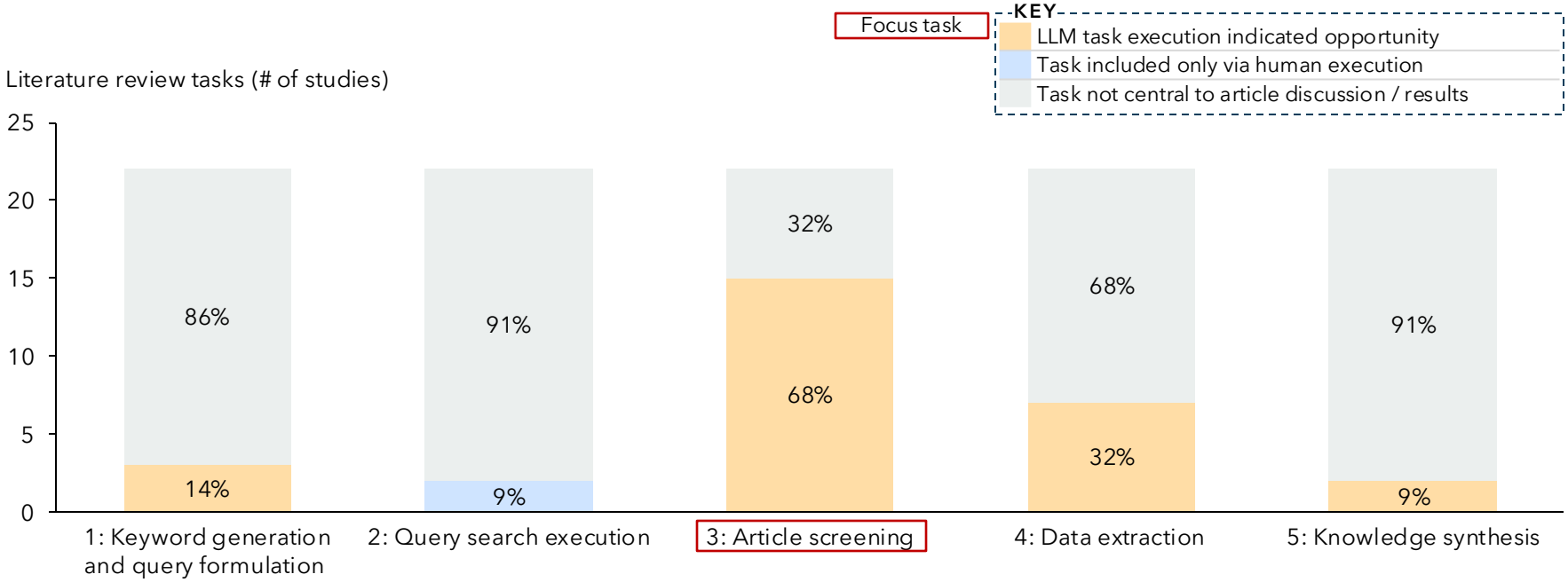
5: Knowledge synthesis

GW TAI TRUSTWORTHY AI INITIATIVE

Sources: Extracted and synthesized from [18-39]

Across these five tasks, LLMs gain the most traction in the literature within article screening, where studies find potential

Opportunity identified in ~70% of studies for article screening; only ~9% for synthesis



GW TAI TRUSTWORTHY AI INITIATIVE

Sources: Extracted and synthesized from [18-39]

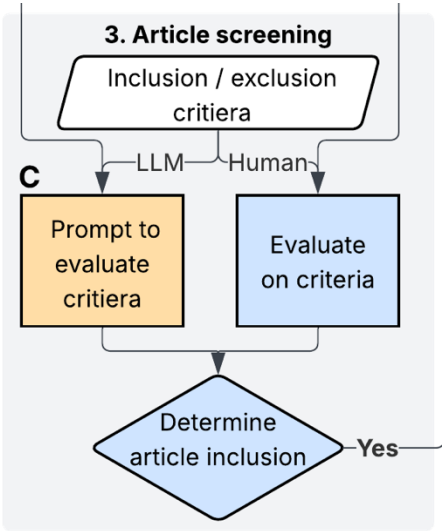
As literature review workflows are analogous to those in proposal writing, we can ground claims about purpose-built tools in evidence

Literature Review Tasks		Proposal Writing Tasks
1	Keyword generation and query formulation	Define search keywords
2	Query search execution	Search for documents in knowledge base
3	Article screening	Determine if documents are relevant to use
4	Data extraction	Sift through documents to parse out relevant information
5	Knowledge synthesis	Draft information into a narrative

GW TAI TRUSTWORTHY AI INITIATIVE

Mapping purpose-built tool features to core workflow tasks reveals the breadth of stated and implied functionality

Article screening is the primary focus within literature



All mapped tools claim end-to-end capabilities while only explicitly stating ~40% of core tasks

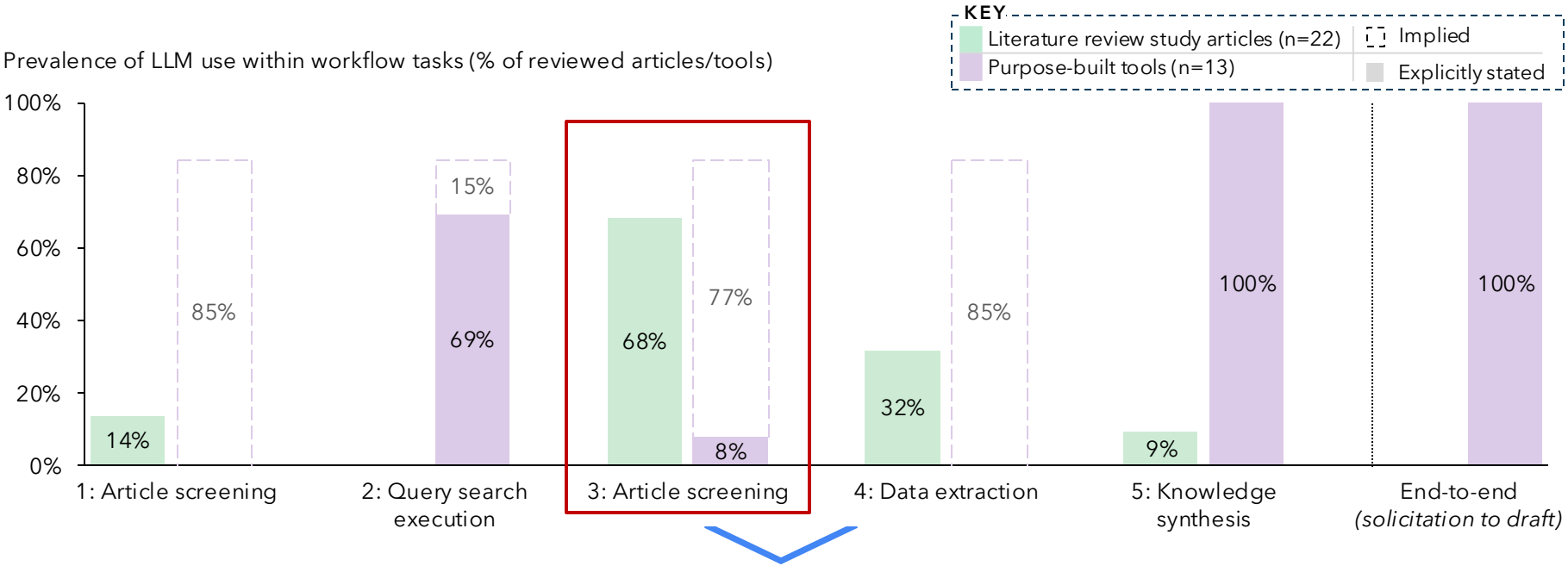
Tool	1: Keyword generation and query formulation	2: Query search execution	3: Article screening	4: Data extraction	5: Knowledge synthesis	End-to-end *(solicitation to draft)*
AutogenAI	•	✓	•	•	✓	✓
AutoRFP.ai	•	✓	✓	•	✓	✓
govdash.com	•	✓	•	•	✓	✓
GOVEAGLE	•	✓	•	•	✓	✓
GovSignals	•	✓	•	•	✓	✓
grant assistant	x	x	x	x	✓	✓
Grantable	•	•	•	•	✓	✓
grantboost	x	x	x	x	✓	✓
inventive	•	✓	•	•	✓	✓
loopio	•	✓	•	•	✓	✓
RFP GEN AI	•	✓	•	•	✓	✓
responsive	•	✓	•	•	✓	✓
sweetspot	•	•	•	•	✓	✓

KEY ✓ Explicitly stated • Implied x Not applicable

GW TAI TRUSTWORTHY AI INITIATIVE

Note: Identified commercial platforms restricted to those involving RFI/RFP/RFx and/or grant/proposal writing purposes; mapping focuses on core proposal writing tasks (e.g., data extraction for proposal writing rather than bid/no-bid requirements); set of tools is non-exhaustive // Sources: LHS [18-39], RHS [5-17]

Stated commercial tool functionality diverges markedly from opportunities in the current literature conversation



The gap between marketed commercial tool capabilities and the opportunities identified in literature necessitates robust frameworks for performance evaluation

GW TAI TRUSTWORTHY AI INITIATIVE

Sources: Extracted and synthesized from [5-39]

Summary of key findings

Five core tasks emerge from the unstructured text processing workflow in the literature, with article screening being the most indicated opportunity

Stated commercial tool functionality diverges markedly from opportunities in the current literature conversation

The gap between marketed commercial tool capabilities and the opportunities identified in literature necessitates robust frameworks for performance evaluation

Agenda

Introduction and motivations

Methods

Results and key findings

Future research

We can look towards evaluation techniques in the literature, although varied and ad-hoc, to inform an evaluation framework

Evaluation metric *(non-exhaustive)*	**1: Keyword generation and query formulation**	**3: Article screening**	**4: Data extraction**	**5: Knowledge synthesis**
Retrieval	✓			✓
Coverage (% of articles)	✓			✓
Completeness	✓			
Classification	✓	✓	✓	
Accuracy		✓	✓	
Sensitivity / Recall		✓		
Specificity		✓		
F-scores (F1/F2/F3)	✓	✓		
Agreement		✓	✓	
Cohen's kappa		✓		
Prevalence-Adjusted Bias-Adjusted kappa (PABAK)		✓	✓	
Text analysis		✓		✓
Flesch-Kincaid Grade Level		✓		
Simple Measure of Gobbledygook (SMOG)				
Recall-Oriented Understudy for Gisting Evaluation (ROUGE)				✓
Qualitative evaluation			✓	✓
Categorical error typology			✓	
Human / expert assessment				✓

✓ Commonly referenced in literature

Sources: Extracted and synthesized from [18-39]

Initial results from the literature indicate opportunity but do not show performance matching the current statements from purpose-built tools

1 Keyword generation and query formulation

- Keyword generation yielded **<10% of articles** *(coverage)*[38]
- Query search terms were **incorrect ~50% of the time** *(completeness)*[31]

3 Article screening

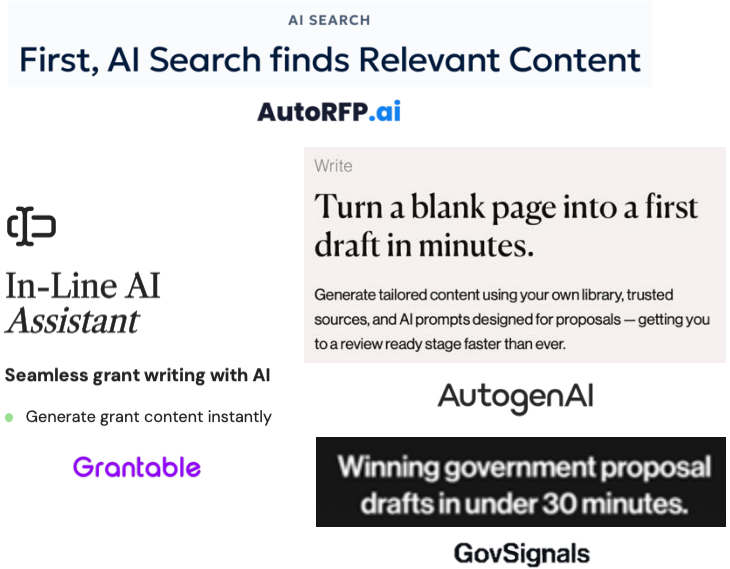
- Higher performing study found **~75% *sensitivity*; findings vary**[28-30]

4 Data extraction

- *F1-scores* between **mid-70s and 90s**[33-35]
- Performance was lower for questions requiring inference or synthesis[34]

5 Knowledge synthesis

- Performance evaluation **ranked** different approaches; no conclusive quality score[38]



Sources: RHS [5, 6, 9, 11]

All together, the gap between marketed tool capabilities, the literature, and initial performance necessitates robust evaluation frameworks

Key considerations

- How is a **threshold** for relevant defined?
- Is there a **ground truth**?
- What is the **reference point**?
- What **other techniques** may be useful for evaluation?
- How do you manage performance **trade-offs** (e.g., sensitivity vs. specificity)?
- At what point does an error become **truly problematic**?

Thank you!

Five core tasks emerge from the unstructured text processing workflow in the literature, with article screening being the most indicated opportunity

Stated commercial tool functionality diverges markedly from opportunities in the current literature conversation

The gap between marketed commercial tool capabilities and the opportunities identified in literature necessitates robust frameworks for performance evaluation

Q / A