4.4 Generative IR

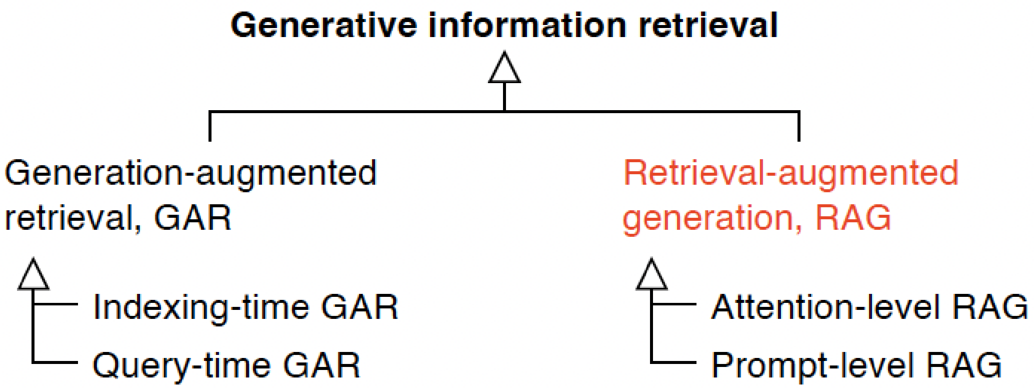
What is Biomedical and Health Informatics? - http://informatics.health/
William Hersh, MD, FACMI, FAMIA, FIAHSI

Generative IR

- IR and generative AI
- Generative-augmented retrieval
- Retrieval-augmented generation
- Resource attribution
- Concerns for generative AI in IR

IR in the era of generative AI

- Does generative AI signify the "end of search" (Wong, 2024; Honan, 2025)?
  - Even for physicians (Dorn, 2024)?
- How can generative AI augment search (Gienapp, 2024)?
- New era of "dense" retrieval (Khan, 2024)?
  - Sparse retrieval – classic IR we have covered so far
  - Dense – deeper LLM representation



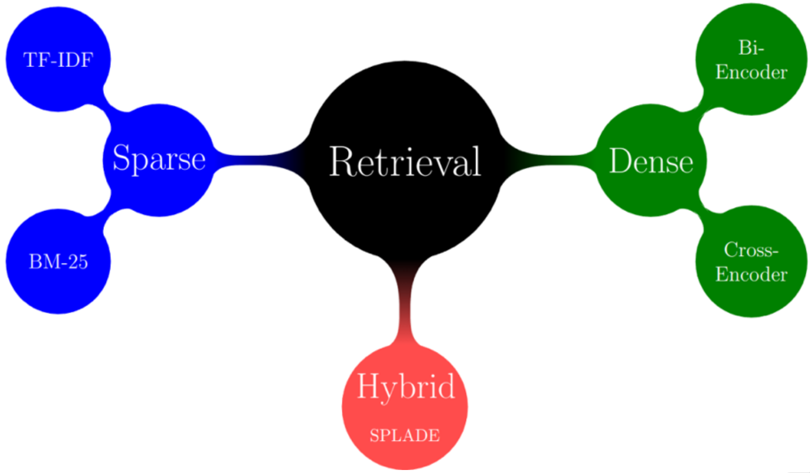
Generative information retrieval

Generation-augmented retrieval, GAR

- Indexing-time GAR
- Query-time GAR

Retrieval-augmented generation, RAG

- Attention-level RAG
- Prompt-level RAG



Mixing IR with LLMs

- Adding generative AI to search, e.g.,
  - Bing – using versions of GPT-4 and others from OpenAI, now called CoPilot
  - Google – using versions of Gemini, now called AI Overviews
- OpenAI adding search capabilities to ChatGPT; also allows development of "GPTs" (formerly "plug-ins") that add customization
- PubTator – using LLMs to improve performance (Wei, 2024)

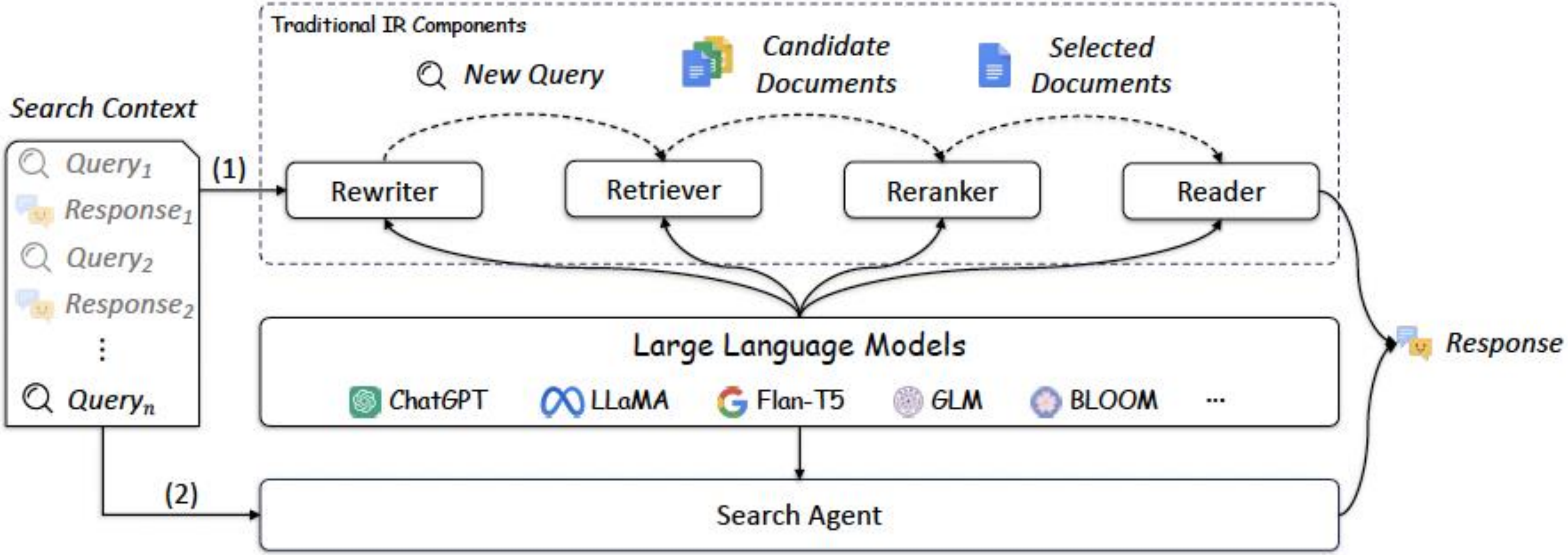
Search still matters in era of LLMs (Hersh, 2024)

- Many information needs, from simple to complex, motivate use of IR
- Users of such systems, particularly academics, have concerns for
  - Authoritativeness – who authored
  - Timeliness – when authored
  - Contextualization – veracity or grounding, and supporting evidence
- Use cases for biomedical and health search
  - Clinical – patient-care questions
  - Research – methods and insights
  - Teaching – synthesizing knowledge for our students

Comparing IR systems with LLMs

- ChatGPT deemed to provide more informative information than Google snippets for 4 cancer questions (Hopkins, 2023)
- Output of ChatGPT vs. Google evaluated by 20 experts in domains of congenital heart disease, atrial fibrillation, heart failure, or cholesterol (Van Bulck, 2023)
  - Responses deemed trustworthy and valuable, with few considering them dangerous
  - Compared to Google, 40% deemed information from ChatGPT more valuable, 45% as valuable, and 15% less valuable (although few details provided)
- For 150 health-related questions from the TREC Health Misinformation Track (Fernández-Pichel, 2025)
  - Search engines correctly answered 50-70% of questions
  - LLMs had higher accuracy, correctly answering about 80% of questions, with smaller LLMs enhanced by RAG methods

Possible role(s) for generation-augmented retrieval (GAR; Zhu, 2023)

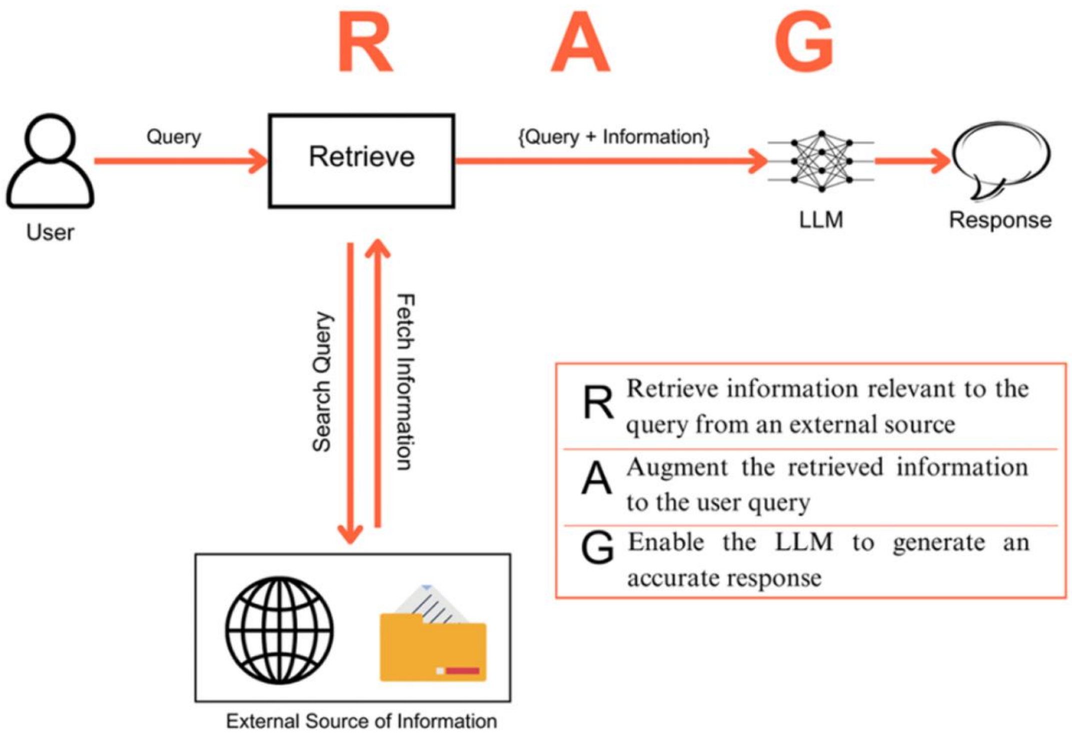


Use of generative AI to enhance retrieval

- MedCPT transformer model trained with PubMed queries-clicks leads to small improvements over BM25 (Jin, 2023)
- Improving dynamic retrieval of ED notes by predicting which notes likely to be read (Jiang, 2023)
- Matching patients to clinical trials – using variety of methods and datasets (Kusa, 2023; Dobbins, 2023; Jin, 2024; Unlu, 2024; Wornow, 2024; Nievas, 2024)

Retrieval-augmented generation (RAG)

- Impractical to train/update LLMs on a frequent basis
- Can “update” performance by adding retrieved content to prompts in context windows to improve performance of LLMs (Kimothi, 2025)
  - Including in biomedicine (Yang, 2025)



R Retrieve information relevant to the query from an external source

A Augment the retrieved information to the user query

G Enable the LLM to generate an accurate response

Efficacy of RAG

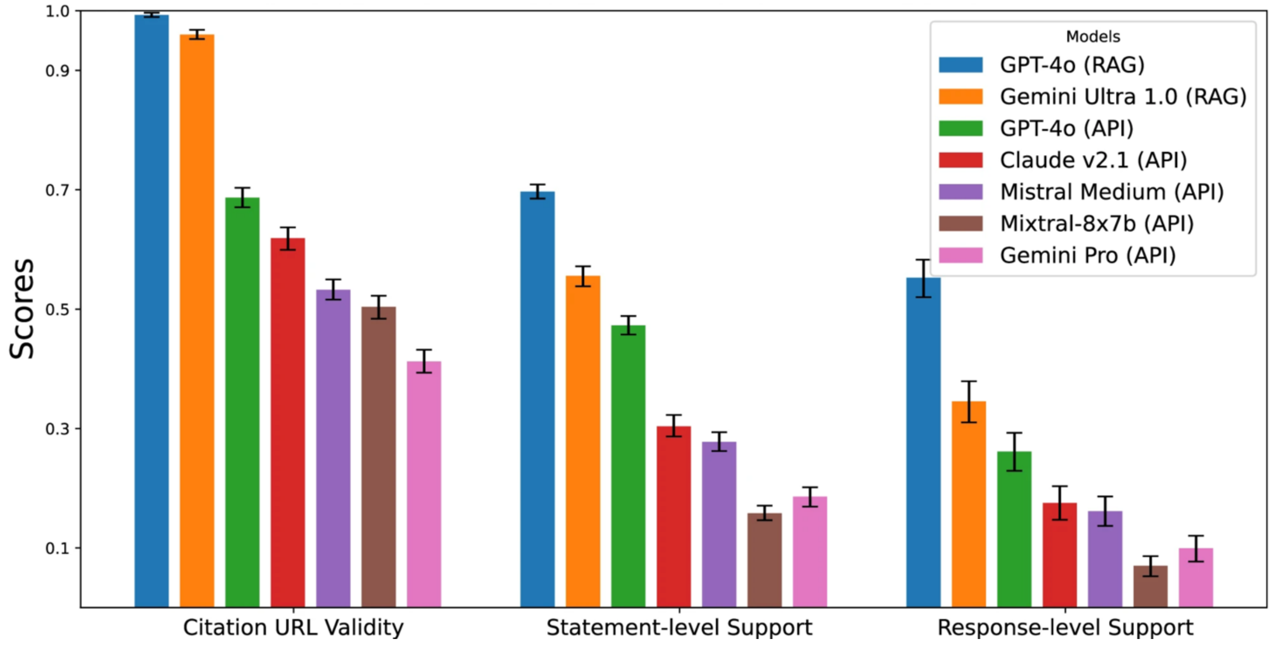
- Adding Web search content to ChatGPT prompt reduced accuracy of correct answers using TREC Health Misinformation Track data (Koopman, 2023)
- Development of LLM framework Almanac found to improve question-answering over standard LLMs based on factuality, completeness, user preference, and adversarial safety (Zakka, 2024)
- For health questions, RAG improved correctness of answers for smaller LLMs (Fernández-Pichel, 2025)
- Systematic review of 20 studies in biomedical domain found pooled odds ratio of 1.35 when RAG added to various tasks with LLMs (Liu, 2025)

Resource attribution

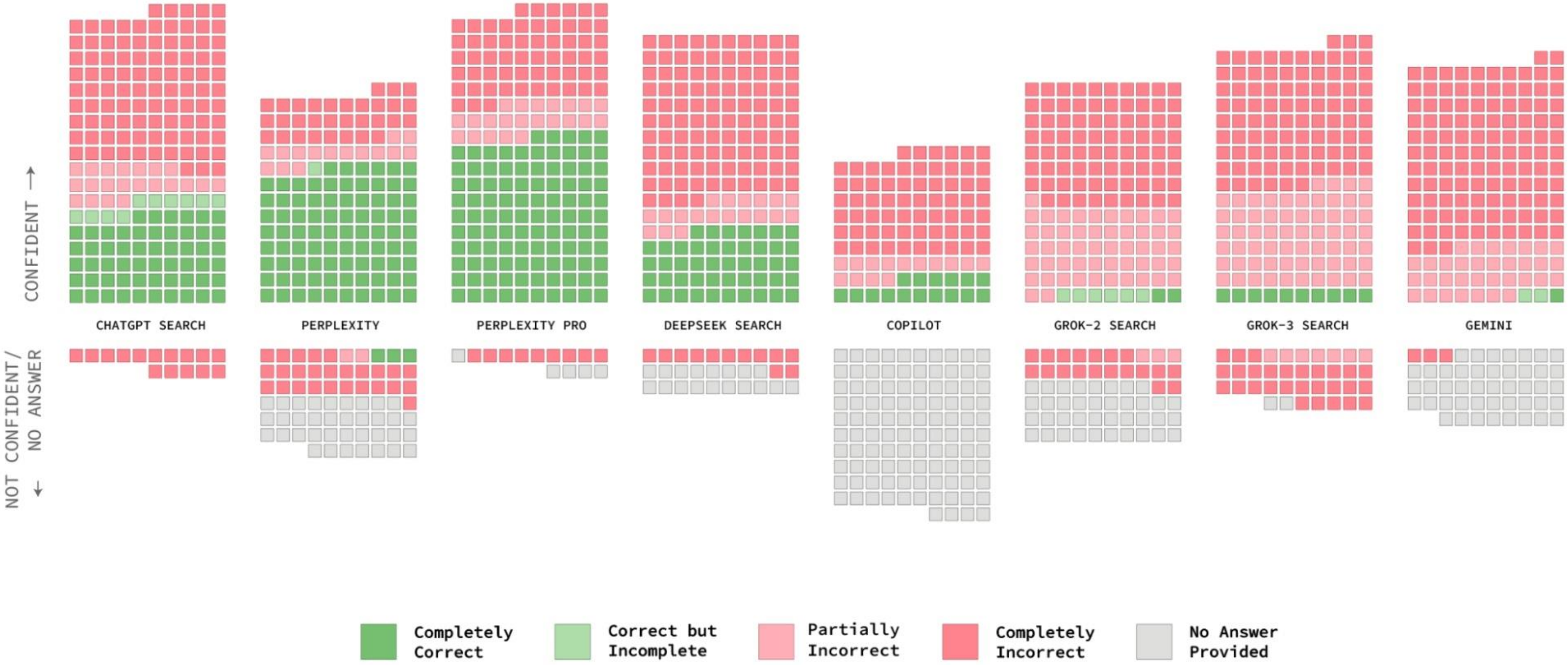
- Fabrication and errors in bibliographic citations – asked to produce short literature reviews on 42 multidisciplinary topics (Walters, 2023)
  - 55% of GPT-3.5 citations and 18% of GPT-4 citations fabricated
  - 43% of real (non-fabricated) GPT-3.5 citations and 24% of real GPT-4 citations included substantive errors
- Prompted to cite articles about learning health systems, GPT-3.5 cited 98% incorrect; GPT-4 cited more and only 20.6% incorrect (Chen, 2023)

Citation of relevant references (Wu, 2025)

- Questions extracted from well-known Web health information sources
- Validation from clinician experts
- Latest LLMs using RAG did best, with high URL validity but lower statement-level and resource-level support
- Other issues
  - Grounded vs. correct claims
  - Sources behind paywalls



Problem not limited to medicine (Jaźwińska, 2025)



Concerns for LLMs in IR (Shah, 2023)

- Opacity and hallucinations
  - LLMs don’t know when they don’t know
- Stealing content and Web site traffic
  - LLMs learn from other sites’ content and may divert traffic from their Web sites
- Taking away learning and serendipity
  - Search is exploring and we may learn new unrelated things

Another concern is generative text contamination

- Estimated LLM text in scientific literature about 1% (Gray, 2024) and up to 6.3% in mathematics and 17.5% in computer science (Liang, 2024)
- 6.5-16.9% of text of peer reviews for AI-related conferences from LLMs (Liang, 2024)
- Generative AI-fabricated papers easy to detect via Google Scholar, with content often about topics susceptible to disinformation (Haider, 2024)
- Misuse of AI worse than plagiarism (Shaw, 2025)?
- Sometimes detected and corrected, but full extent of problem not known (Kwon, 2025)
- Ongoing list of flagrant discoveries

Protecting scientific integrity going forward

- Guidance on use of generative AI incomplete and inconsistent by publishers and journals (Ganjavi, 2024)
- Some principles (Blau, 2024), similar to those advocated by others (Chen 2024; Chauhan, 2024)
  - Transparent disclosure and attribution
  - Verification of AI-generated content and analyses
  - Documentation of AI-generated data
  - A focus on ethics and equity
  - Continuous monitoring, oversight, and public engagement