ArXiv [Preprint]. 2024 Jul 4:arXiv:2305.13338v3. [Version 3]

Other versions

Gene Set Summarization Using Large Language Models

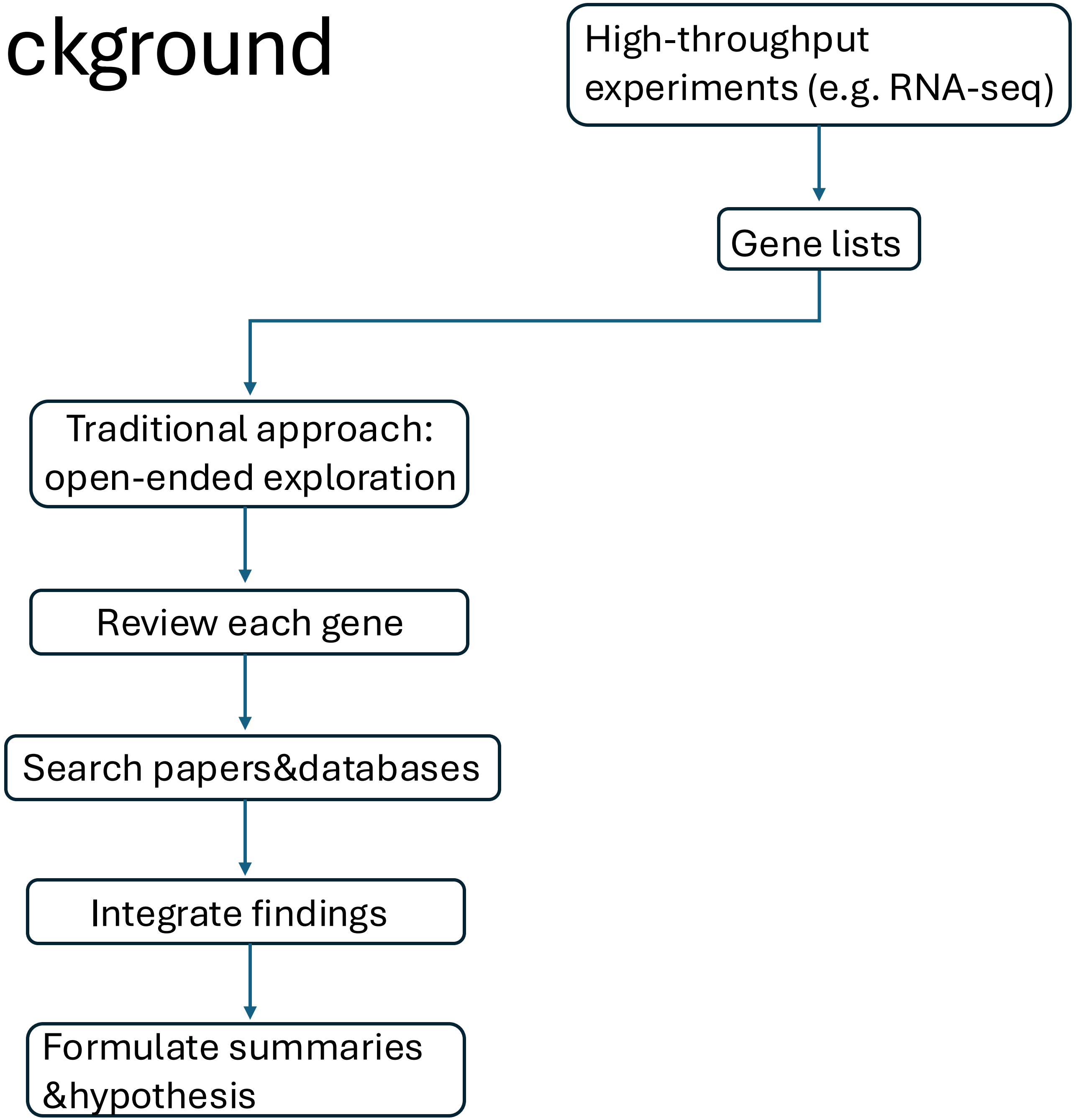
Marcin P Joachimiak [1], J Harry Caufield [1], Nomi L Harris [1], Hyeongsik Kim [2], Christopher J Mungall [1]

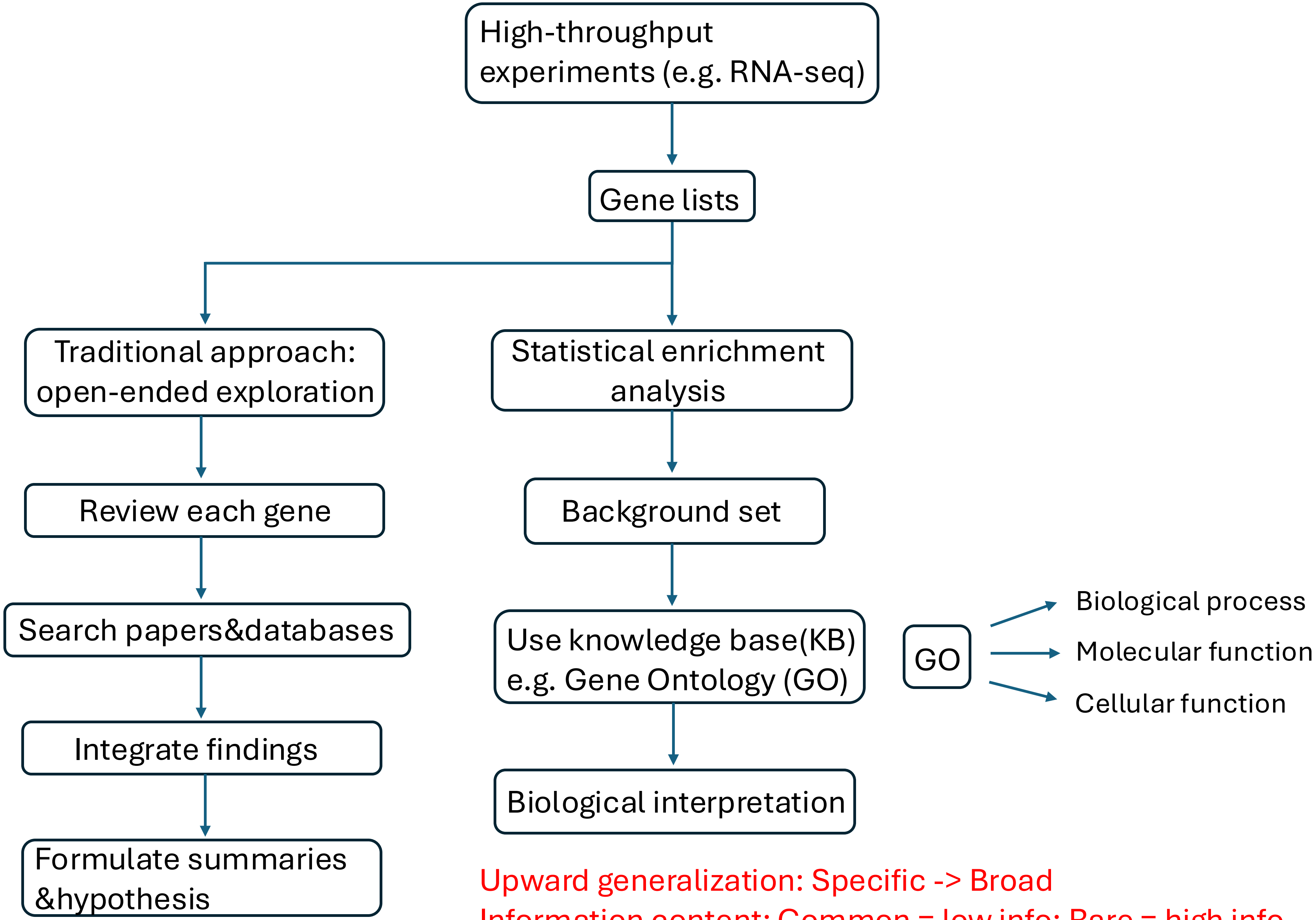
Author information   Copyright and License information

PMCID: PMC10246080 PMID: 37292480

Jenny Ge
Data Science Journal Club
9.18.2025

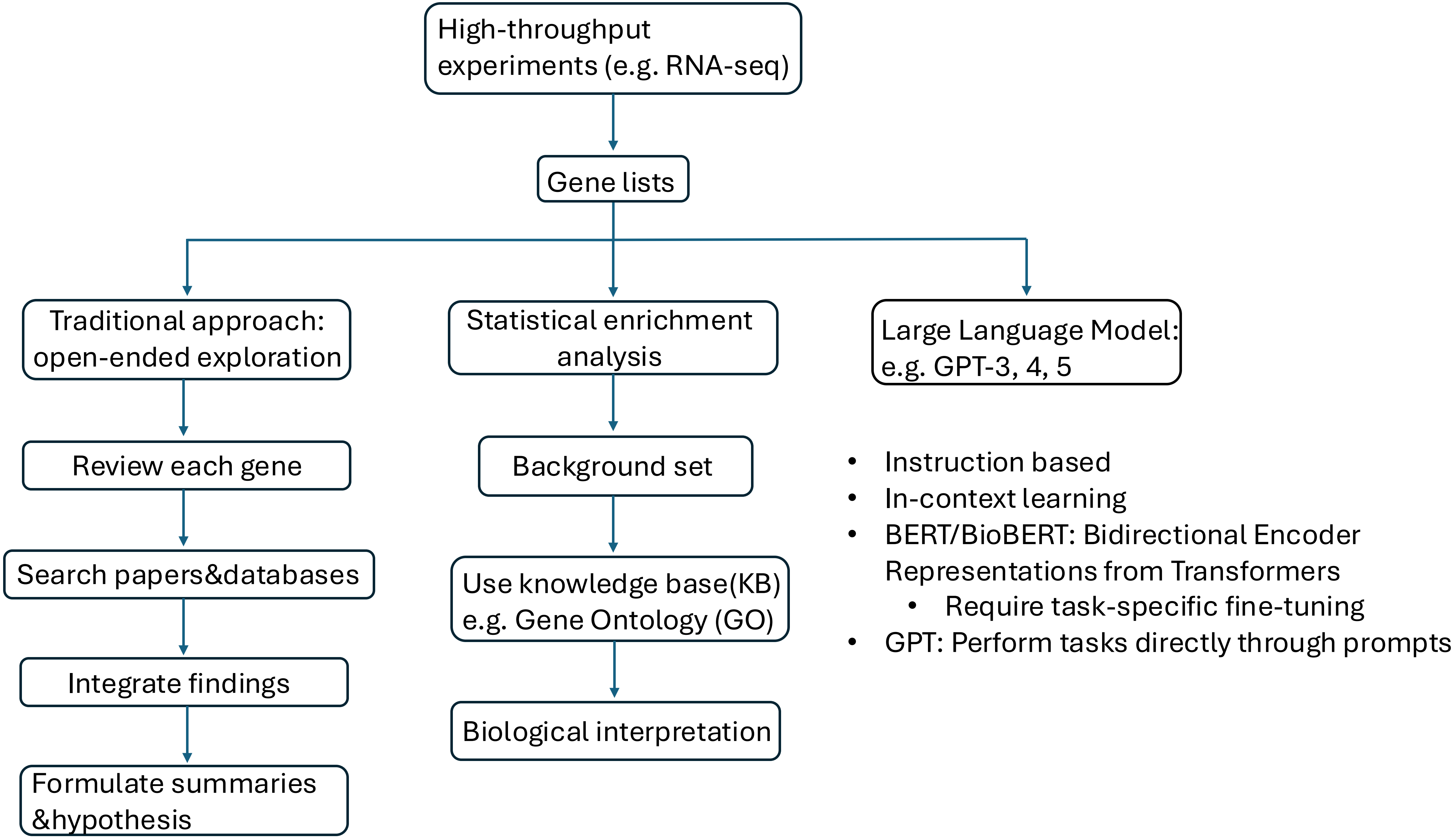
Background

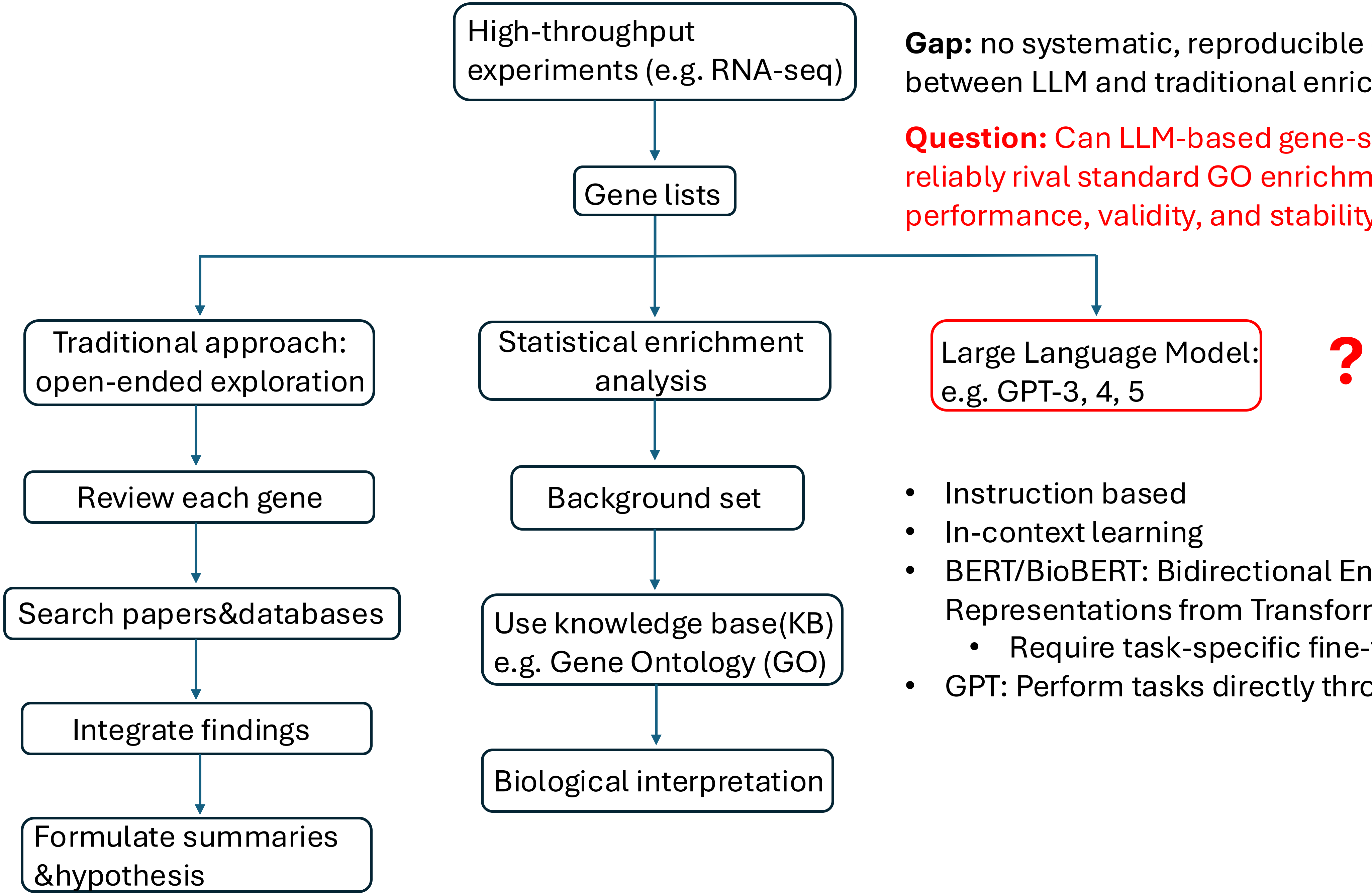

High-throughput experiments (e.g. RNA-seq)
Gene lists
Traditional approach: open-ended exploration
Review each gene
Search papers&databases
Integrate findings
Formulate summaries &hypothesis



Upward generalization: Specific -> Broad

Information content: Common = low info; Rare = high info


High-throughput experiments (e.g. RNA-seq)
Gene lists
Traditional approach: open-ended exploration
Review each gene
Search papers&databases
Integrate findings
Formulate summaries &hypothesis
Statistical enrichment analysis
Background set
Use knowledge base(KB) e.g. Gene Ontology (GO)
Biological interpretation
Large Language Model: e.g. GPT-3, 4, 5
• Instruction based
• In-context learning
• BERT/BioBERT: Bidirectional Encoder Representations from Transformers
• Require task-specific fine-tuning
• GPT: Perform tasks directly through prompts



Gap: no systematic, reproducible comparison between LLM and traditional enrichment methods

Question: Can LLM-based gene-set summaries reliably rival standard GO enrichment on performance, validity, and stability?

- Instruction based
- In-context learning
- BERT/BioBERT: Bidirectional Encoder Representations from Transformers
  - Require task-specific fine-tuning
- GPT: Perform tasks directly through prompts

Method

What is TALISMAN?
Terminological ArtificiaL Intelligence SuMmarization of Annotation and Narratives

How does it work?

Input

A gene list → Gene info from database

- Gene Symbol->gene ID
- Narrative gene description
- Automated gene description from GO

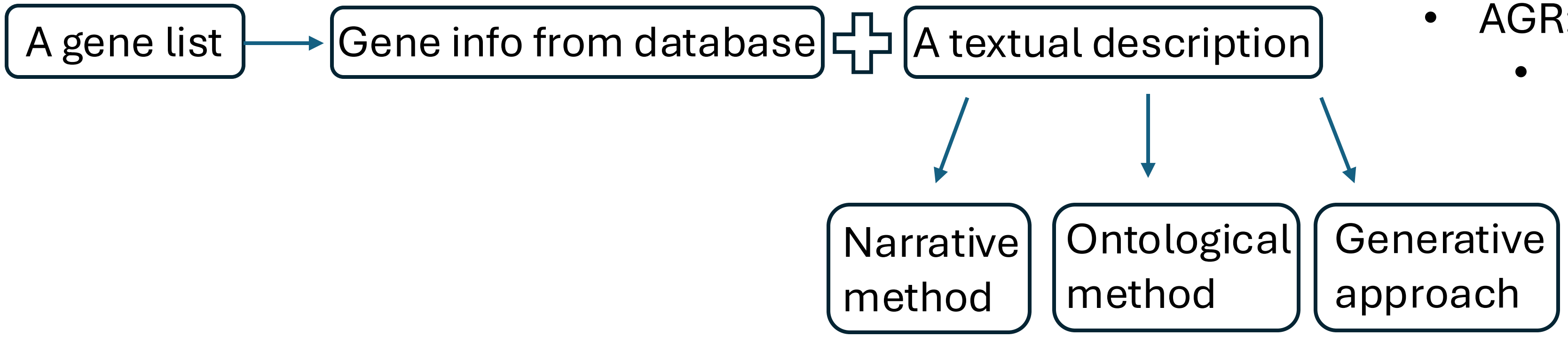
Alliance of Genome Resource API: One platform, unified gene knowledge

Method

What is TALISMAN?
Terminological ArtificiaL Intelligence SuMmarization of Annotation and Narratives

How does it work?

Input



- RefSeq database: curated, non-redundant, reference sequences for gene
- AGR: Alliance of Genome Resources
  - Integrates multiple databases using a unified annotation system.

- Narrative method: gene symbol + narrative description (RefSeq)
- Ontological method: gene symbol + ontology term summaries (GO/AGR controlled natural language)
- Generative approach: only gene symbols

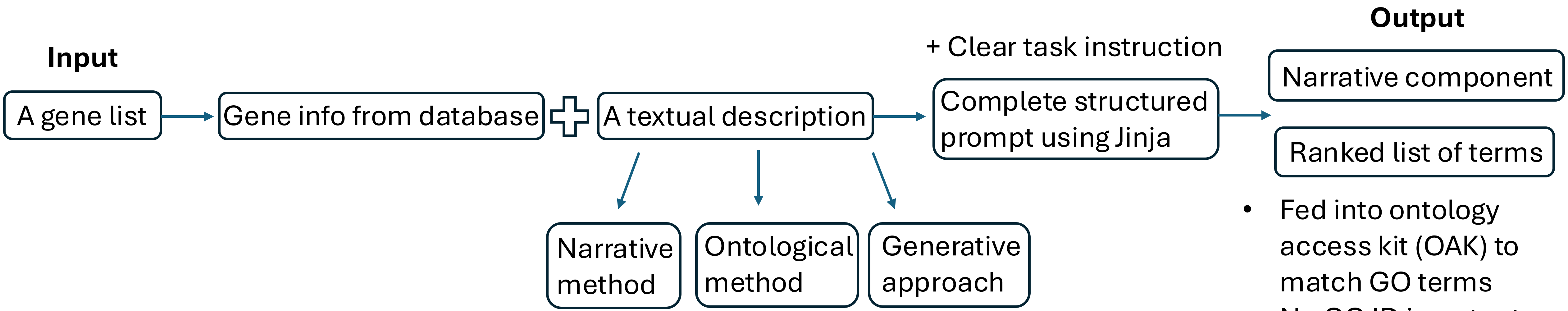
Token Length challenge

- When long description, truncate proportionally from back of the sentence
- Truncate factor: TF = 1.0, no truncation; TF = 0.25, only used ¼ of original description

Method

What is TALISMAN?
Terminological ArtificiaL Intelligence SuMmarization of Annotation and Narratives

How does it work?



- Fed into ontology access kit (OAK) to match GO terms
- No GO ID in output

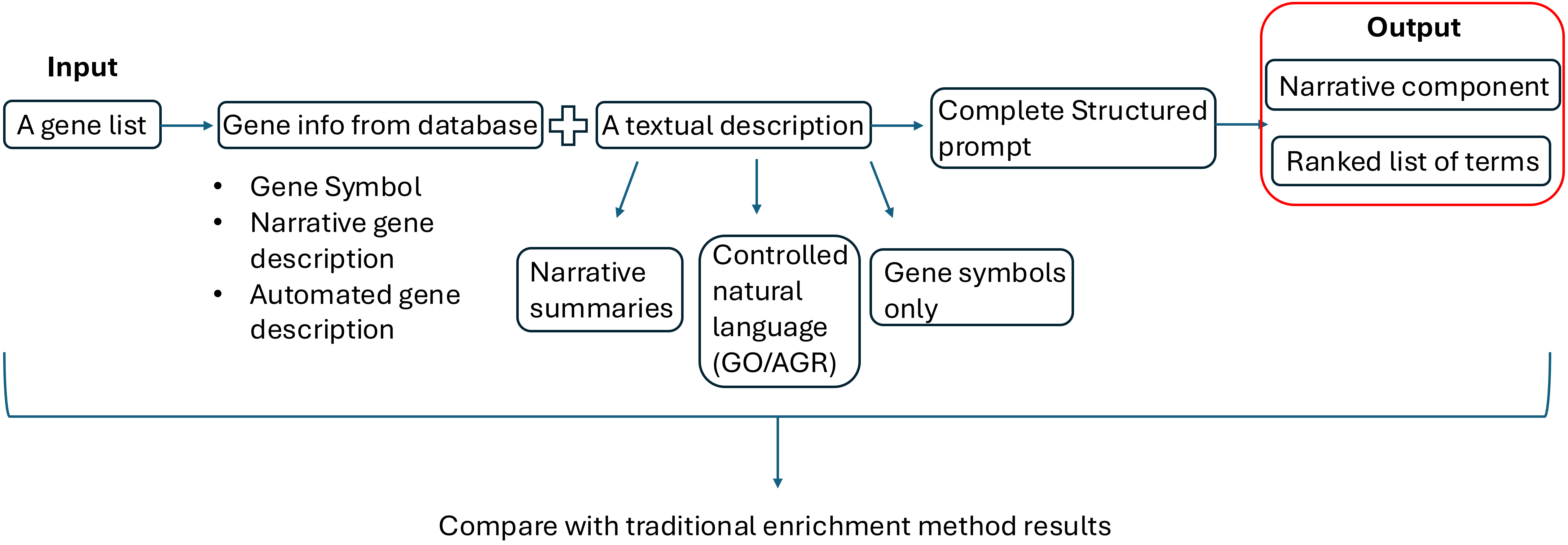
- Jinja: a template engine, combine fixed template with variable gene information

Variables here:
- Taxon (species, e.g. human/mouse)
- Gene description
- OAK: a toolkit that provides standardized access to ontologies

Method

What is TALISMAN?
Terminological ArtificiaL Intelligence SuMmarization of Annotation and Narratives

How does it work?



How TALISMAN is implemented, how it is used.

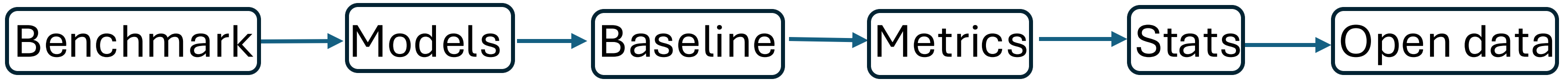
- What it is: Python tool for GPT-based gene function summaries
- Interfaces: Command line and local web UI
- Cost-savvy: Caches results to avoid paying twice
- No API? : Works via copy-paste with ChatGPT
- Use case: Fast, consistent narratives + term lists for genes

Benchmark design

Benchmark → Models → Baseline → Metrics → Stats → Open data

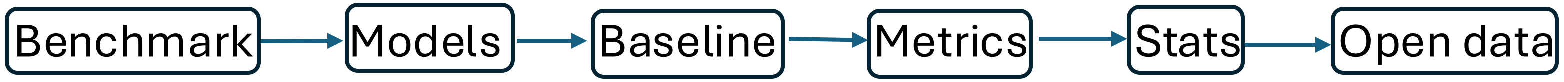
- built their own human gene sets (70)
- noise-injected versions for robustness
- Drop 10% + random genes

Benchmark design



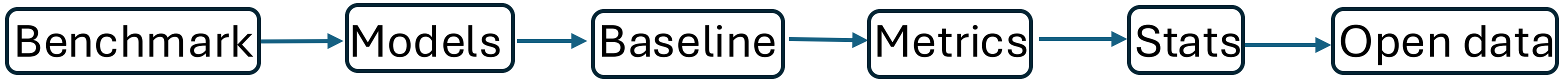
- 3 TALISMAN input strategies
- 3 generations of GPT: 3.0 / 3.5 / 4

Benchmark design



- Baseline: standard enrichment
- Account for GO hierarchy (parent/child terms count as matches)

Benchmark design



- Precision, Recall, F1
- Has hit / Has top hit
- Tested under different thresholds (n, p)

Benchmark design

Benchmark → Models → Baseline → Metrics → Stats → Open data

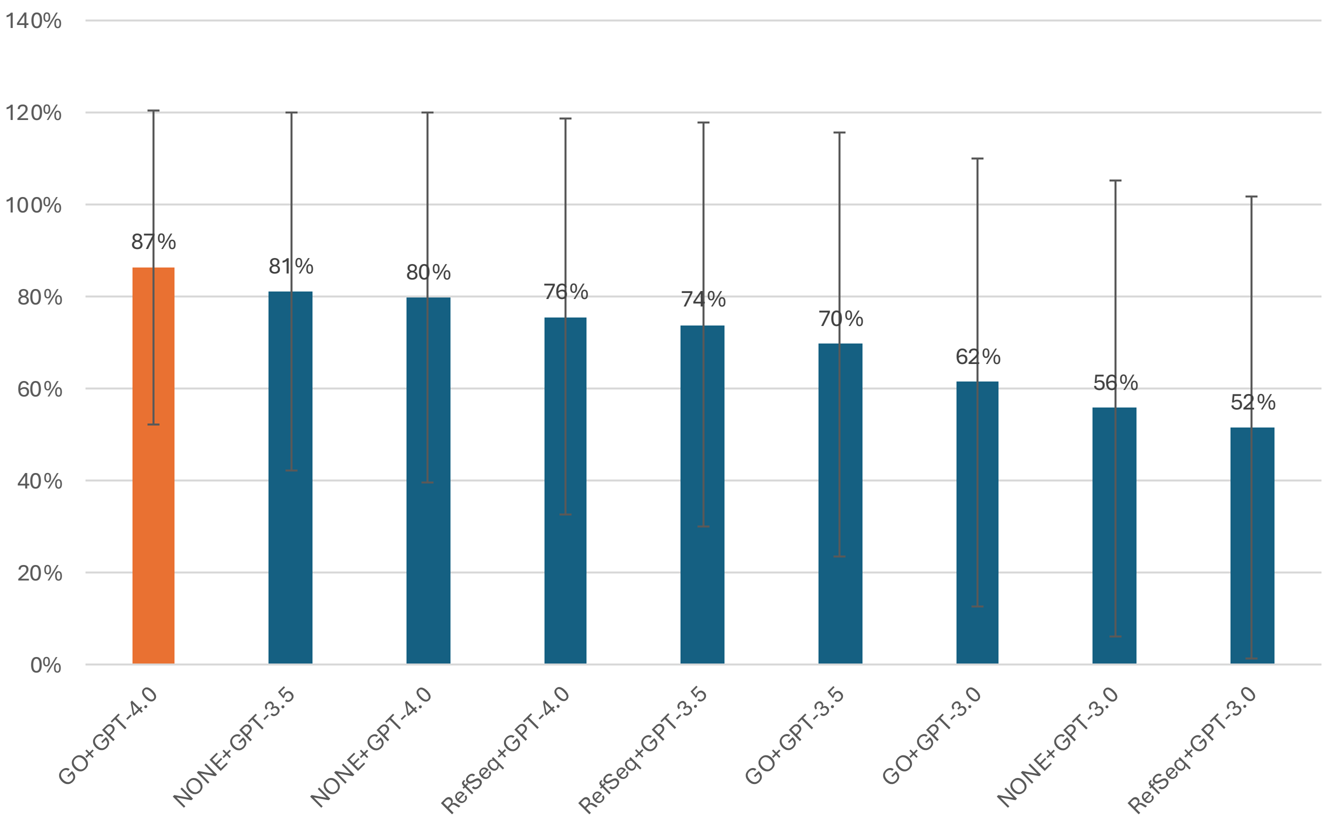
- Mann–Whitney exact test: used to compare the difference between two data distributions

1. run standard enrichment analysis to obtain a **gold standard**.
2. check whether TALISMAN (GPT) predictions include the gold standard's **top 1 term**.

Metric: proportion of runs with a "has top hit."

Across all experiments, how often GPT finds the key term?

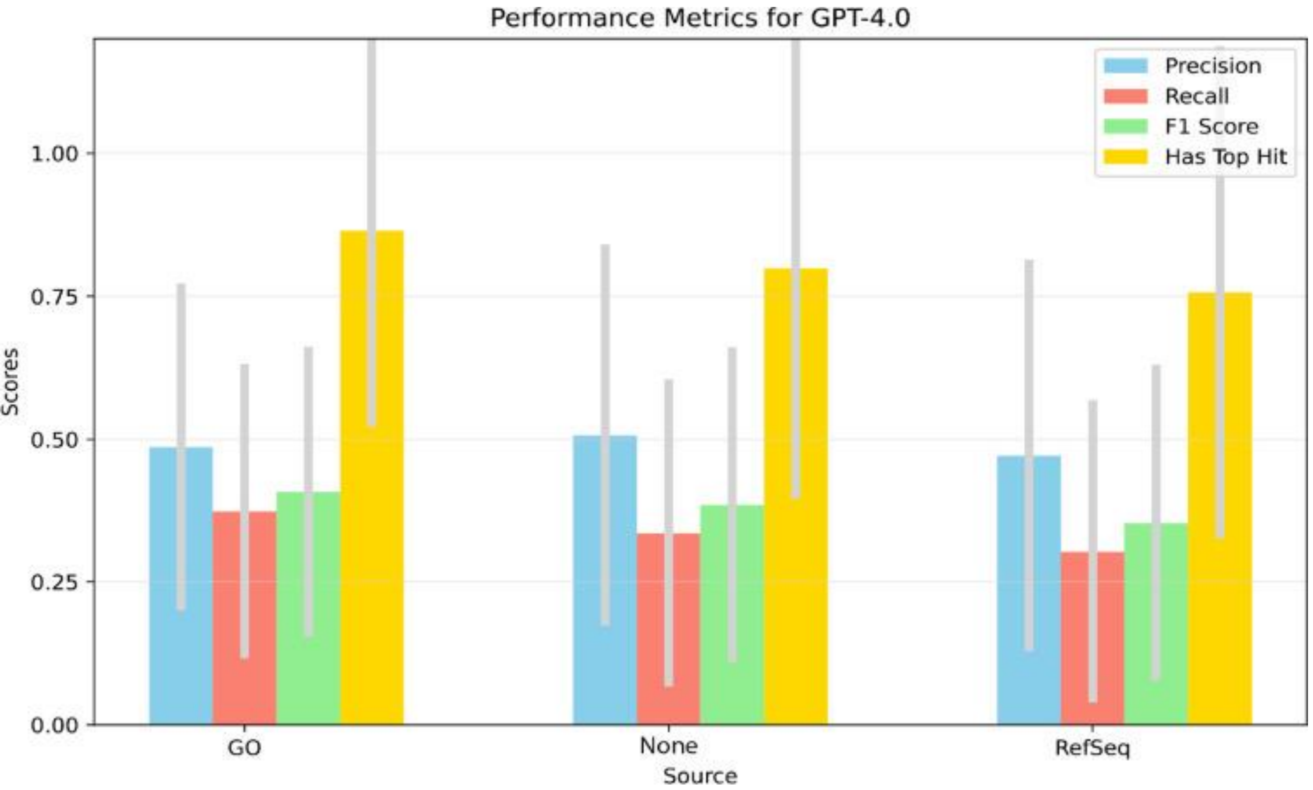
Table 1



- Metric: “Has Top Hit” = recovered the #1 GO term
- Best: GPT-4 + GO (~0.86) and more consistent
- Runner-up: GPT-3.5 + None (~0.81), strong without extra text
- Lagging: GPT-3.0 lower and more variable
- Takeaway: Model > source; but depending on the text source, models trade off precision and recall differently
- Variability remains

Which input source best balances precision and recall for GPT-4?

Figure 3

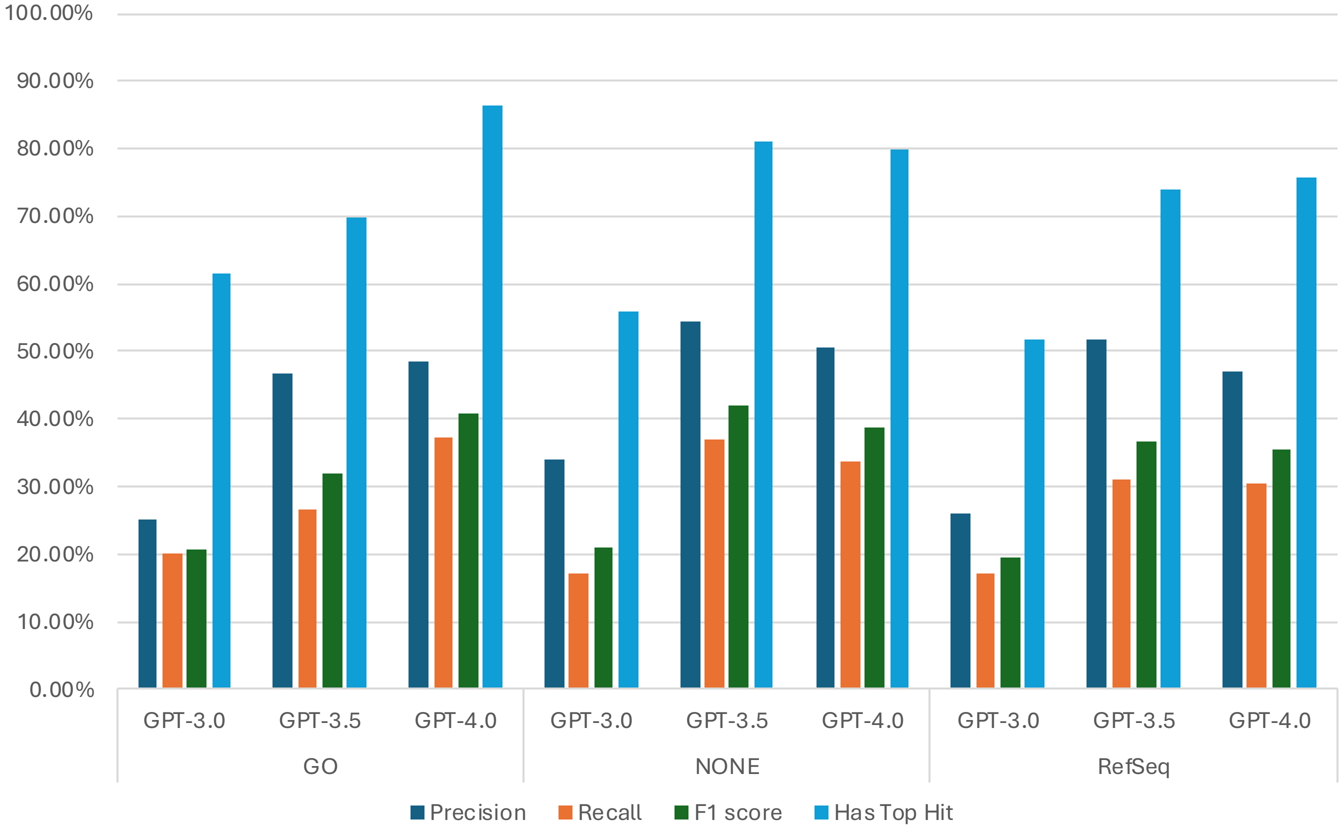


Precision: correct / predicted (fewer false alarms)
Recall: correct / true (fewer misses)
F1: harmonic mean of precision & recall

- Mean Precision / Recall / F1 over gene sets (top-10 gold, ontology closure)
- Recall & F1: GO descriptions highest → best coverage of enriched terms
- Precision: None (no synopsis) highest → most conservative/clean lists
- RefSeq: Middle of the pack on all three
- Trade-off: GO = higher recall but more false positives; None = higher precision but more misses
- Use case: Exploration ⇒ GO; Precision-critical ⇒ None; RefSeq ⇒ balanced narrative

Which model–source combo performs best on precision, recall, F1, and top-hit?

Table 2



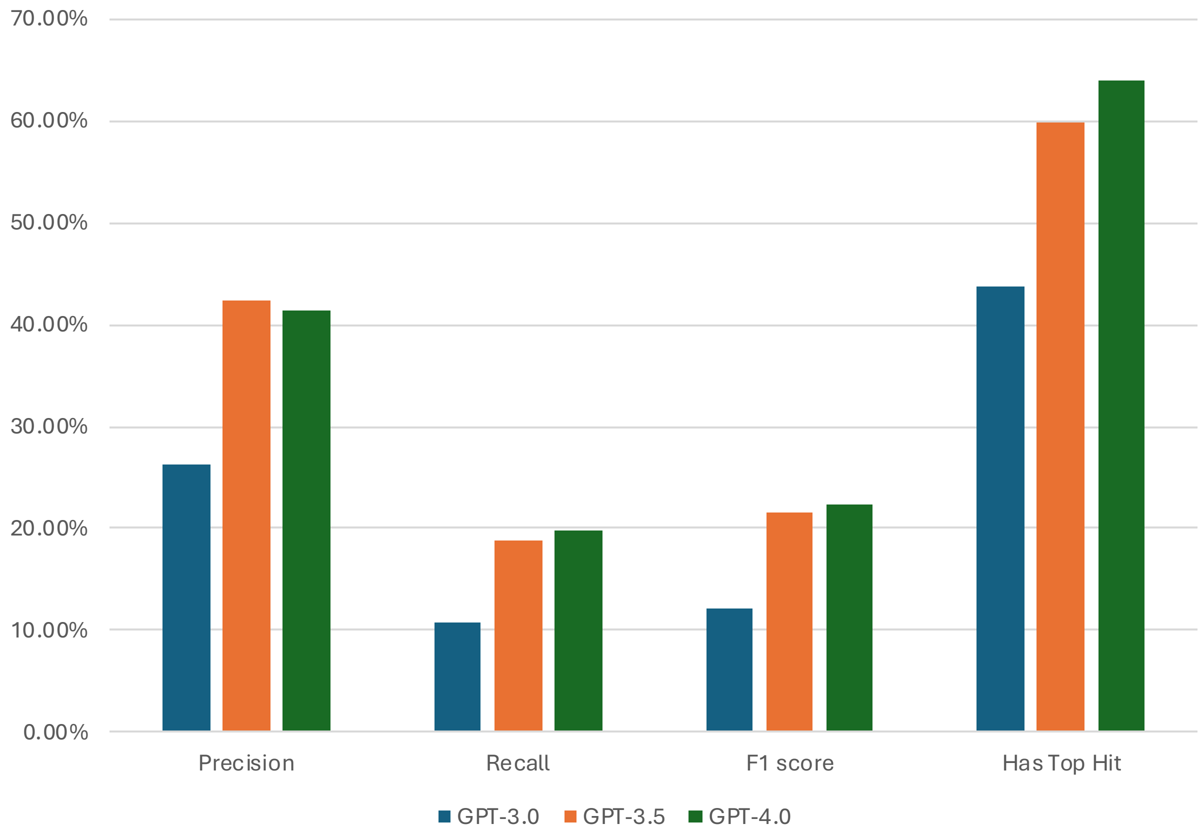
- Precision, Recall, F1, and Has-Top-Hit for each Model × Source combo
- Recall & Top-hit: GPT-4 + GO best
- Precision & F1: GPT-3.5 + None best
- Trade-off: GO ↑ recall, ↓ precision; None ↑ precision, ↓ recall (RefSeq ~ middle)

Overall: GPT-3.0 lowest, most variable

Precision: correct / predicted (fewer false alarms)
Recall: correct / true (fewer misses)
F1: harmonic mean of precision & recall

Which model performs best on average across sources?

Table 3



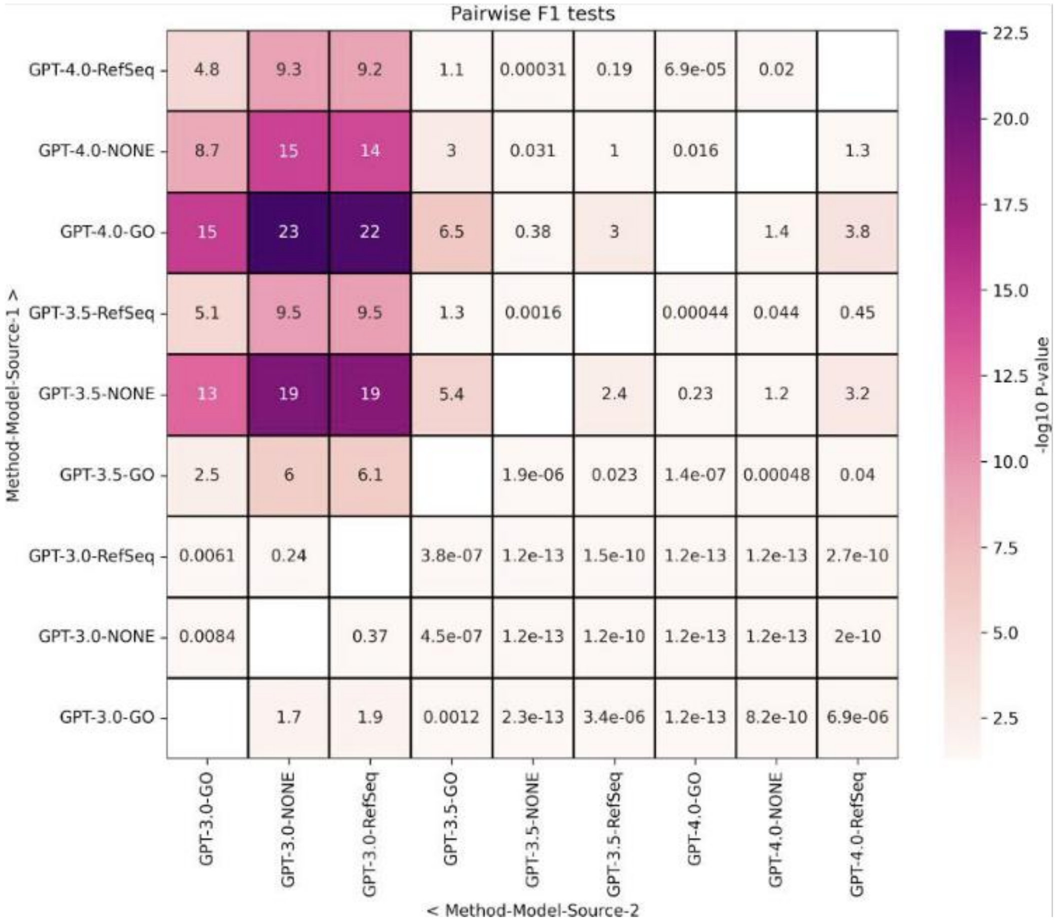
- Mean Precision, Recall, F1, Has-Top-Hit averaged over all sources/cutoffs
- GPT-4: Best Recall, F1, Has-Top-Hit; Precision slightly below GPT-3.5
- GPT-3.5: Best Precision; mid Recall/F1
- GPT-3.0: Lowest on all metrics

Takeaway: Prefer GPT-4 for coverage/F1; use GPT-3.5 when precision is paramount

Precision: correct / predicted (fewer false alarms)
Recall: correct / true (fewer misses)
F1: harmonic mean of precision & recall

Which model–source pairs are significantly different on F1?

Figure 4



- Pairwise Mann–Whitney (exact) tests on F1 between all Model × Source combos; cell value = −log10(p)
- Darker/bigger → more significant difference
- All GPT-3.5/4.0 ≫ GPT-3.0 (deep cells) → newer models clearly better
- Top performers: GPT-3.5-None and GPT-4.0-GO are significantly better than most others; 3.5-None vs 4.0-None often not significant

Note: Heatmap shows significance, not direction. Use Table 2 means to see who's higher

$-\log_{10}(\text{p-value})=1.3$

1.3 → p≈0.05 marginally significant

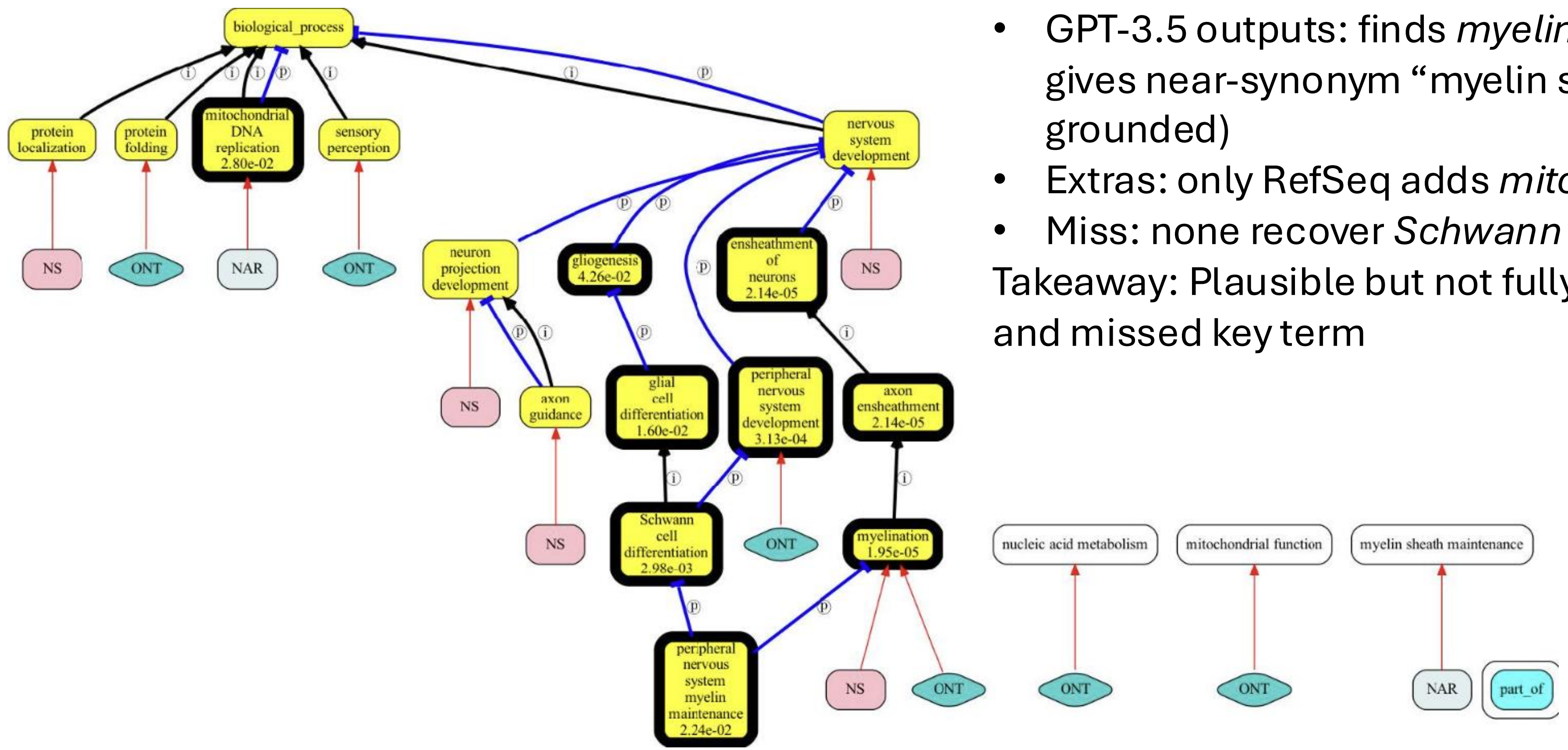
2 → p≈0.01

3 → p≈0.001

5 → p≈1e−5 very significant

Do GPT summaries recover the key GO terms for “sensory ataxia”?

Figure 6



- Gold standard top hits: *myelination*, *Schwann cell differentiation*
- GPT-3.5 outputs: finds *myelination* with GO and None; RefSeq gives near-synonym “myelin sheath maintenance” (not grounded)
- Extras: only RefSeq adds *mitochondrial DNA replication*
- Miss: none recover *Schwann cell differentiation*

Takeaway: Plausible but not fully aligned; synonym/grounding gaps and missed key term

What do GPT-4 summaries say across GO, RefSeq, and None inputs?

Table 4

Source	Summary	Mechanism
Ontological synopsis (GPT-4.0)	The provided genes are mainly involved in processes related to the nervous system, peripheral nerve function, and cellular maintenance functions.	These genes may contribute to the biological processes related to the nervous system development, cellular response regulation, and transportation of molecules within cells, interacting in various pathways.
Narrative synopsis (GPT-4.0)	Majority of the genes are associated with neuropathic conditions and myelin-related processes in the peripheral nervous system.	The underlying biological mechanism may be related to the formation, maintenance, and function of the myelin sheath in the peripheral nervous system and the regulation of cellular pathways that impact neuronal survival and function.
No synopsis (GPT-4.0)	Enriched terms associated with the given list of genes are mostly involved in the development and maintenance of the nervous system, cellular response, and transport processes.	These genes may contribute to the biological processes related to the nervous system development, cellular response regulation, and transportation of molecules within cells, interacting in various pathways.

- GPT-4 summaries for *sensory ataxia*; inputs = GO / RefSeq / None
- Common themes: nervous system, peripheral nerve, cellular maintenance/transport
- RefSeq: more myelin-specific; mentions neuropathic conditions; myelin formation/maintenance
- GO & None: broader/general wording; mechanisms nearly identical
- Note: prose ≠ statistics; phrases like “enriched” not p-values

Takeaway: readable narrative, input-dependent wording; use as complement to enrichment

Conclusion

- TALISMAN: LLM-based gene-set summarization (GO / RefSeq / None)
- Plausible narratives; not a replacement for statistical enrichment
- GPT-4 + GO → highest recall / top-hit
- GPT-3.5 + None → highest precision / F1
- Clear precision–recall trade-off (GO↑ recall, None↑ precision)
- Outputs non-deterministic; run-to-run variability
- Grounding gaps (synonyms/obsolete terms); missed key terms
- Hallucinations rare for terms; p-values fabricated if requested
- Benchmark provided: 70 sets + perturbed; open code/results

Future direction

- Hybrid pipeline: LLM summary + standard enrichment filtering
- Long-context / newer models; reduce truncation, improve stability
- Expanded benchmarks: more gene sets, organisms, modalities; effect sizes