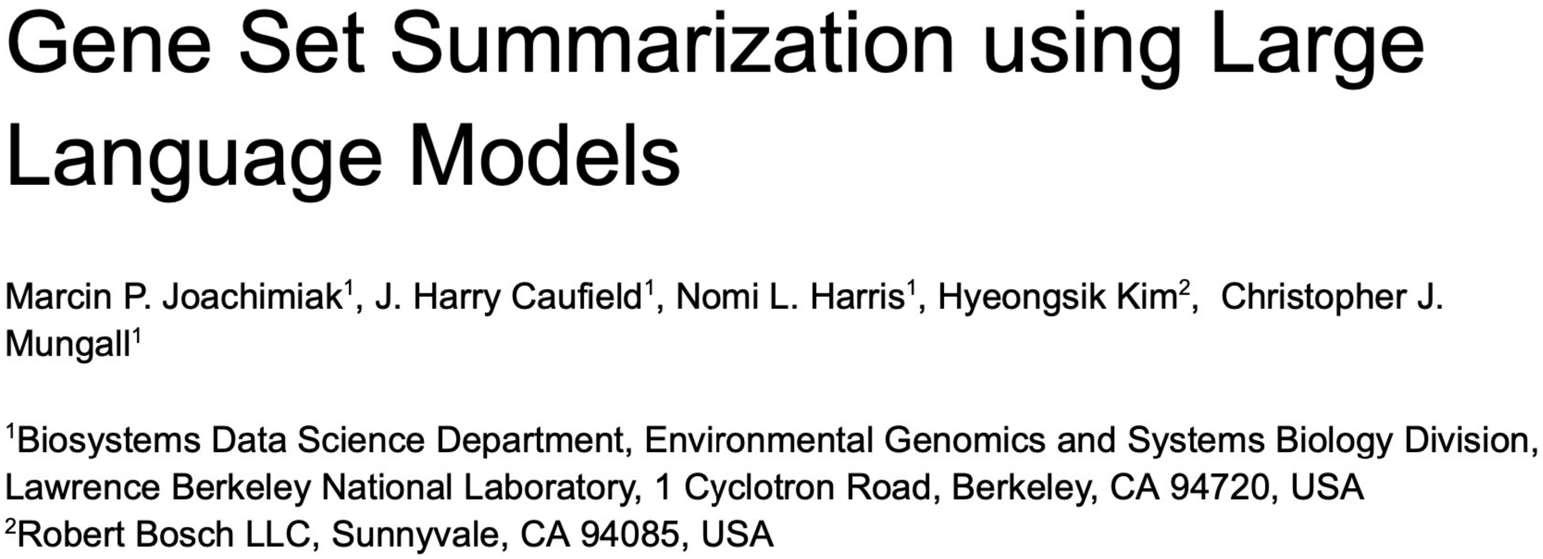
Journal Time

Sijia Huo

Sep 27, 2023



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]

[1]Biosystems Data Science Department, Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Robert Bosch LLC, Sunnyvale, CA 94085, USA

Joachimiak, Marcin P et al. “Gene Set Summarization using Large Language Models.” *ArXiv* arXiv:2305.13338v2. 25 May. 2023 Preprint.

Outline

- Statistical Gene Set Enrichment Analysis & Over-Representation Analysis
- SPINDOCTOR (with Different Summarization Approaches)

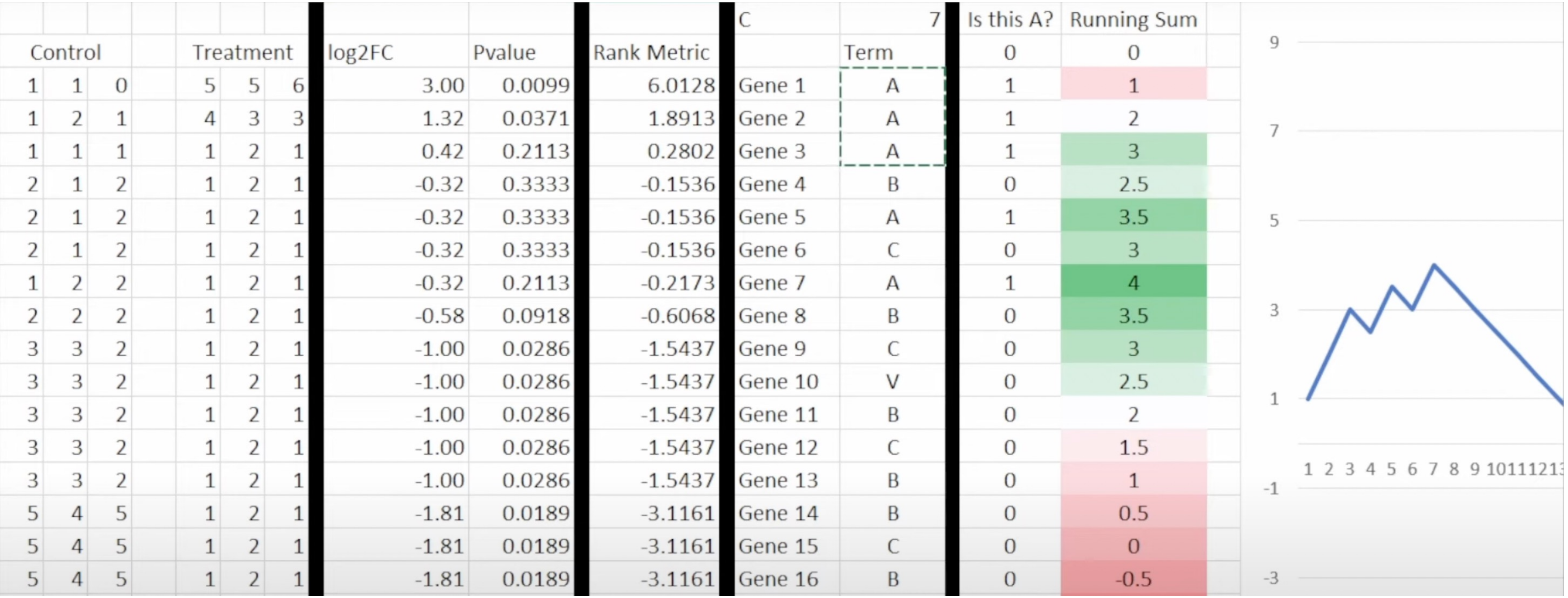
    No Synopsis

    Narrative Synopsis

    Ontological Synopsis

- Evaluation
- Results
- Discussion

Statistical Gene Set Enrichment Analysis (GSEA)

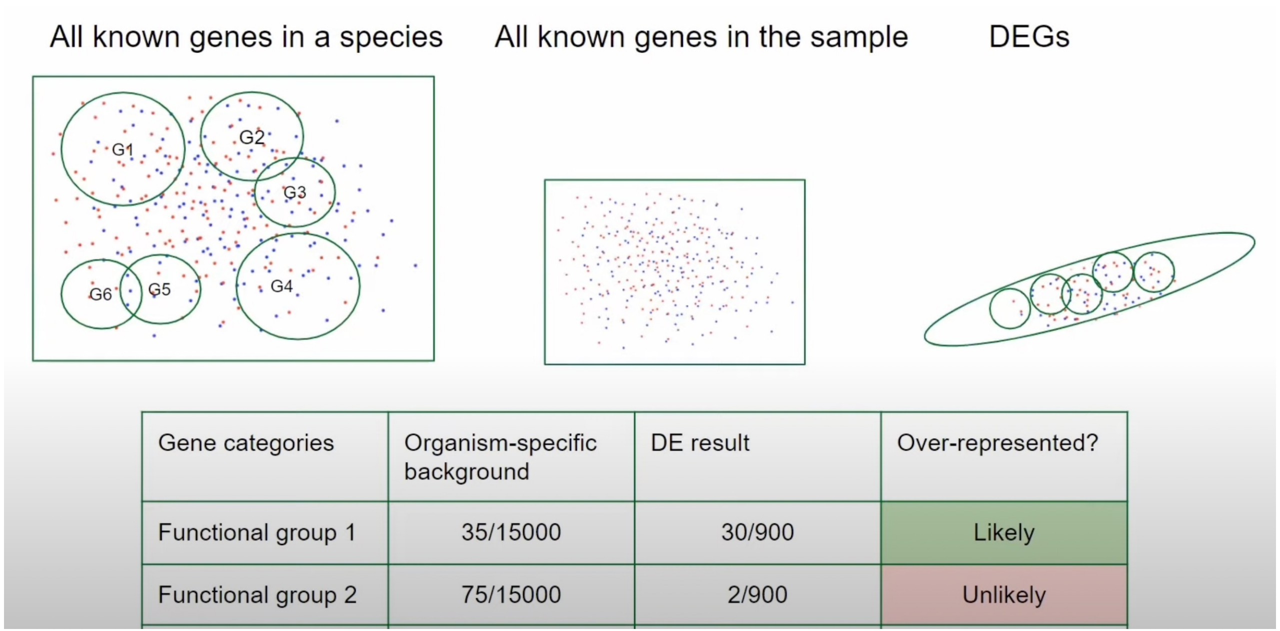


									C	7	Is this A?	Running Sum
Control			Treatment			log2FC	Pvalue	Rank Metric		Term	0	0
1	1	0	5	5	6	3.00	0.0099	6.0128	Gene 1	A	1	1
1	2	1	4	3	3	1.32	0.0371	1.8913	Gene 2	A	1	2
1	1	1	1	2	1	0.42	0.2113	0.2802	Gene 3	A	1	3
2	1	2	1	2	1	-0.32	0.3333	-0.1536	Gene 4	B	0	2.5
2	1	2	1	2	1	-0.32	0.3333	-0.1536	Gene 5	A	1	3.5
2	1	2	1	2	1	-0.32	0.3333	-0.1536	Gene 6	C	0	3
1	2	2	1	2	1	-0.32	0.2113	-0.2173	Gene 7	A	1	4
2	2	2	1	2	1	-0.58	0.0918	-0.6068	Gene 8	B	0	3.5
3	3	2	1	2	1	-1.00	0.0286	-1.5437	Gene 9	C	0	3
3	3	2	1	2	1	-1.00	0.0286	-1.5437	Gene 10	V	0	2.5
3	3	2	1	2	1	-1.00	0.0286	-1.5437	Gene 11	B	0	2
3	3	2	1	2	1	-1.00	0.0286	-1.5437	Gene 12	C	0	1.5
3	3	2	1	2	1	-1.00	0.0286	-1.5437	Gene 13	B	0	1
5	4	5	1	2	1	-1.81	0.0189	-3.1161	Gene 14	B	0	0.5
5	4	5	1	2	1	-1.81	0.0189	-3.1161	Gene 15	C	0	0
5	4	5	1	2	1	-1.81	0.0189	-3.1161	Gene 16	B	0	-0.5

- Rank genes based on fold change values, calculate enrichment score for each functional terms, then conduct hypothesis (permutation) test and adjust for multiple hypothesis testing.

Gene Set Enrichment Analysis| GSEA algorithm
https://www.youtube.com/watch?v=Tm0LhciYxk8

Over-Representation Analysis (ORA)



Gene categories	Organism-specific background	DE result	Over-represented?
Functional group 1	35/15000	30/900	Likely
Functional group 2	75/15000	2/900	Unlikely

- Determine whether a priori defined gene sets (functional group) are more present (over-represented) in a subset of "interesting" genes compared to the background gene lists. Use Fisher's exact test (Hypergeometric test).

Gene Set Enrichment Analysis (+ R tutorial)
https://www.youtube.com/watch?v=B7F7a9NcGS0

SPINDOCTOR

- Both GSEA and ORA make use of knowledge bases (KBs) that have two components: (1) an ontology, which provides a hierarchical logical organization of gene function descriptors; and (2) gene annotations, which associate genes with these descriptors.

- One of the leading system is Gene Ontology (**GO**)

- SPINDOCTOR investigate the ability of GPTs to interpret lists of genes, such as those yielded by gene expression experiments and GWAS. It reframe the task from a statistical enrichment one to a text summarization one.

- SPINDOCTOR take as input a gene set and producing as output (1) a list of ontology terms from GO, analogous to enriched terms in an over-representation analysis; and (2) a narrative summary that weaves together the different functions.

SPINDOCTOR – Prompt Example

I will give you a list of {{ taxon }} genes together with descriptions of their functions.
Perform a term enrichment test on these genes.
i.e. tell me what the commonalities are in their function.
Make use of classification hierarchies when you do this.
Only report gene functions in common, not diseases.
e.g if gene1 is involved in "toe bone growth" and gene2 is involved in "finger morphogenesis"
then the term "digit development" would be enriched as represented by gene1 and gene2.
Only include terms that are statistically over-represented.
Also include a hypothesis of the underlying biological mechanism or pathway.

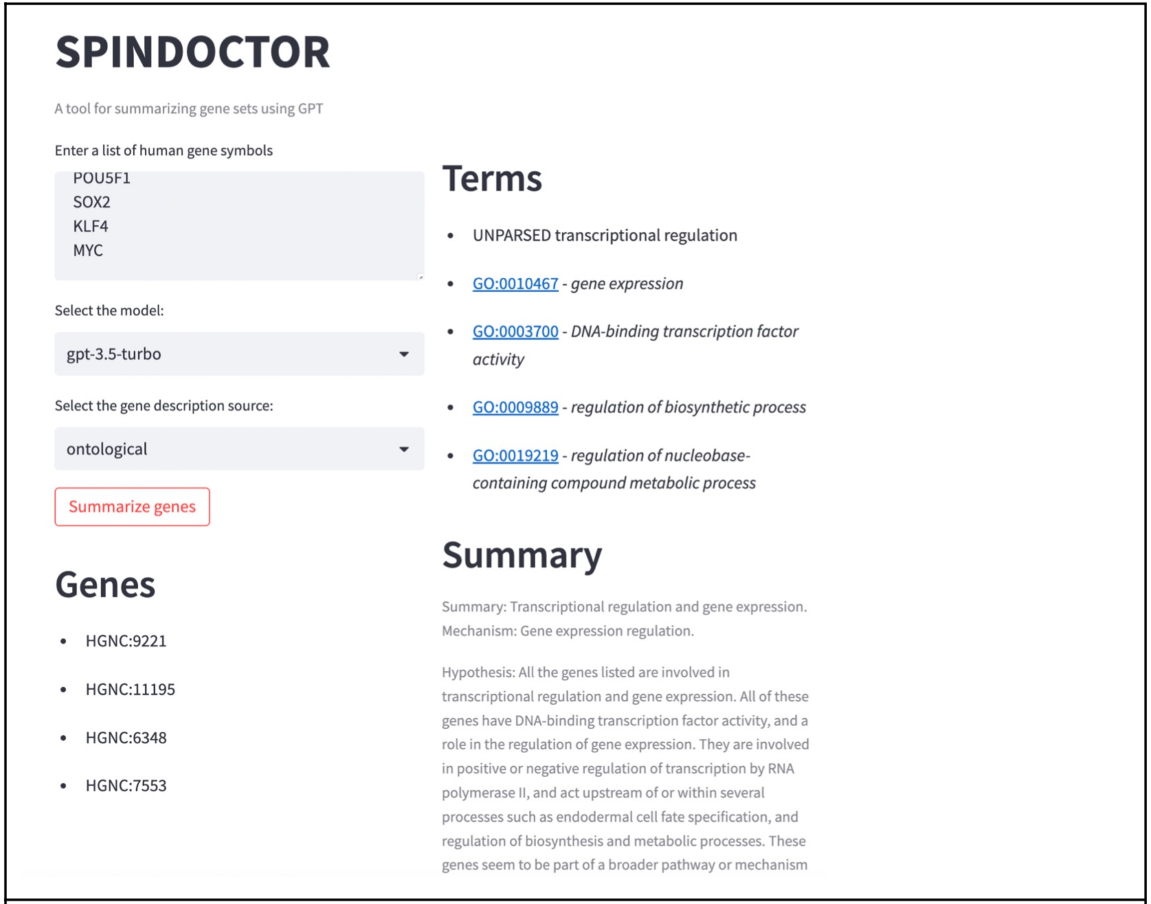
Provide results in the format

{{SUMMARY_KEYWORD}}: <high level summary>
{{MECHANISM_KEYWORD}}: <mechanism>
{{ENRICHED_TERMS_KEYWORD}}: <term1>; <term2>; <term3>

For the list of terms, be sure to use a semi-colon separator, and do not number the list.
Always put the list of terms last, after mechanism, summary, or hypotheses.

Here are the gene summaries:
{GENE_DESCRIPTIONS}

SPINDOCTOR – Interface



SPINDOCTOR

A tool for summarizing gene sets using GPT

Enter a list of human gene symbols

```
POU5F1
SOX2
KLF4
MYC
```

Select the model:

gpt-3.5-turbo

Select the gene description source:

ontological

Genes

- HGNC:9221
- HGNC:11195
- HGNC:6348
- HGNC:7553

Terms

- UNPARSED transcriptional regulation
- GO:0010467 - *gene expression*
- GO:0003700 - *DNA-binding transcription factor activity*
- GO:0009889 - *regulation of biosynthetic process*
- GO:0019219 - *regulation of nucleobase-containing compound metabolic process*

Summary

Summary: Transcriptional regulation and gene expression. Mechanism: Gene expression regulation.

Hypothesis: All the genes listed are involved in transcriptional regulation and gene expression. All of these genes have DNA-binding transcription factor activity, and a role in the regulation of gene expression. They are involved in positive or negative regulation of transcription by RNA polymerase II, and act upstream of or within several processes such as endodermal cell fate specification, and regulation of biosynthesis and metabolic processes. These genes seem to be part of a broader pathway or mechanism

SPINDOCTOR – Summarization Approaches

- SPINDOCTOR generates a structured prompt from the input gene list, containing textual summaries of genes from a list of sources (RefSeq, AGR, Automated Gene Description…)

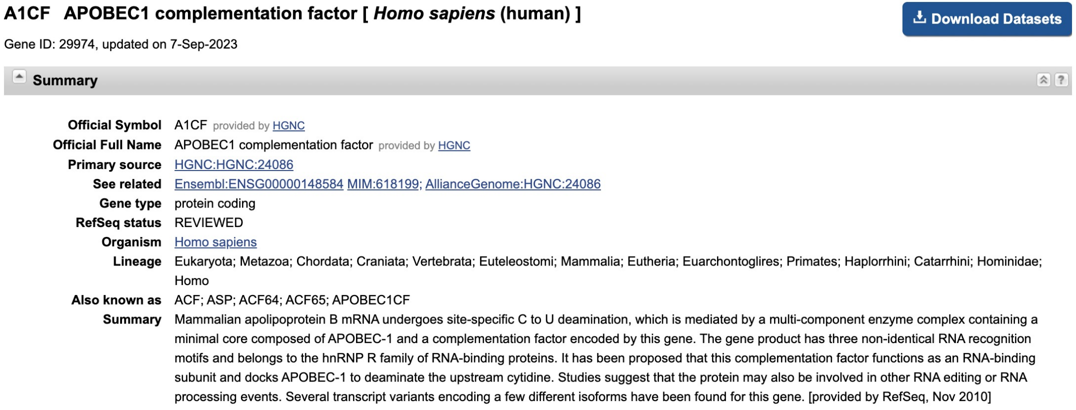
- SPINDOCTOR is intended for fine-tune LLMs such as GPT-3.5 models and successors (e.g. text-davinci-003, gpt-3.5-turbo, and gpt-4).

Synopsis	Source of synopses	Explicit Curation
No synopsis	Underlying Language Model (“latent knowledge base”)	Indirect
Narrative synopsis	RefSeq Gene Summaries	Textual summary
Ontological synopsis	Alliance of Genome Resources (AGR) Automated Gene Descriptions	GO annotations

SPINDOCTOR – Summarization Approaches

□ No Synopsis: Original GPT training Corpus

□ Narrative Synopsis: Narrative Gene Description from RefSeq



A1CF APOBEC1 complementation factor [*Homo sapiens* (human)]

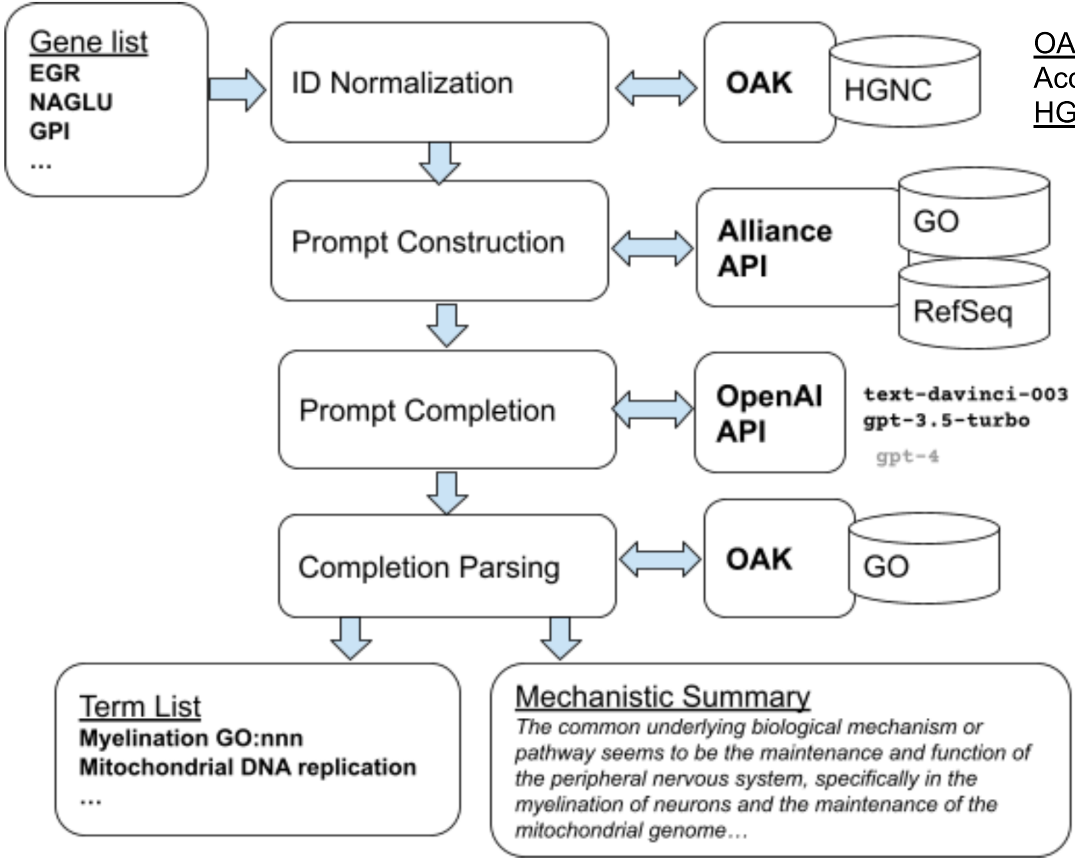
Gene ID: 29974, updated on 7-Sep-2023

Summary

Official Symbol	A1CF provided by HGNC
Official Full Name	APOBEC1 complementation factor provided by HGNC
Primary source	HGNC:HGNC:24086
See related	Ensembl:ENSG00000148584 MIM:618199; AllianceGenome:HGNC:24086
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	ACF; ASP; ACF64; ACF65; APOBEC1CF
Summary	Mammalian apolipoprotein B mRNA undergoes site-specific C to U deamination, which is mediated by a multi-component enzyme complex containing a minimal core composed of APOBEC-1 and a complementation factor encoded by this gene. The gene product has three non-identical RNA recognition motifs and belongs to the hnRNP R family of RNA-binding proteins. It has been proposed that this complementation factor functions as an RNA-binding subunit and docks APOBEC-1 to deaminate the upstream cytidine. Studies suggest that the protein may also be involved in other RNA editing or RNA processing events. Several transcript variants encoding a few different isoforms have been found for this gene. [provided by RefSeq, Nov 2010]

□ Ontological Synopsis (automated gene description): derived from curated ontological GO annotations; here “automated” refers to the ontology-to-text process.

SPINDOCTOR – Pipeline



The common underlying biological mechanism or pathway seems to be the maintenance and function of the peripheral nervous system, specifically in the myelination of neurons and the maintenance of the mitochondrial genome…

OAK: Python Library, Ontology Access Kit, for id normalization
HGNC: for human genes

SPINDOCTOR – Other Details & Implementation

- SPINDOCTOR truncates the length of each gene description proportional to total number of tokens relative to maximum token length (currently 4k for GPT-3.x models, and 8k or 32k for GPT-4) from the end of the string, assuming text at the beginning is more informative. Record this information loss as the truncation factor (TF), with 1 as no truncation and 0 as nothing left.

- SPINDOCTOR uses default configuration with the lowest temperature (max determinism)

- SPINDOCTOR explicitly avoids asking for GO identifiers but only GO terms to avoid GPT-3.5 models hallucinating “likely seeming” numeric identifiers.

- Both a command line interface and a web application interface is provided. The web application interface makes use of the streamlit framework, and currently must be executed locally.

Evaluation: Data

❑ Datasets: 70 human gene sets for evaluation, from multiple sources (e.g., MSigDB, GeneWeaver).

❑ Data Preparation: For each gene set, we generated an additional perturbed gene set simulating noise, where we dropped out 10% of genes and inserted random genes as replacements.

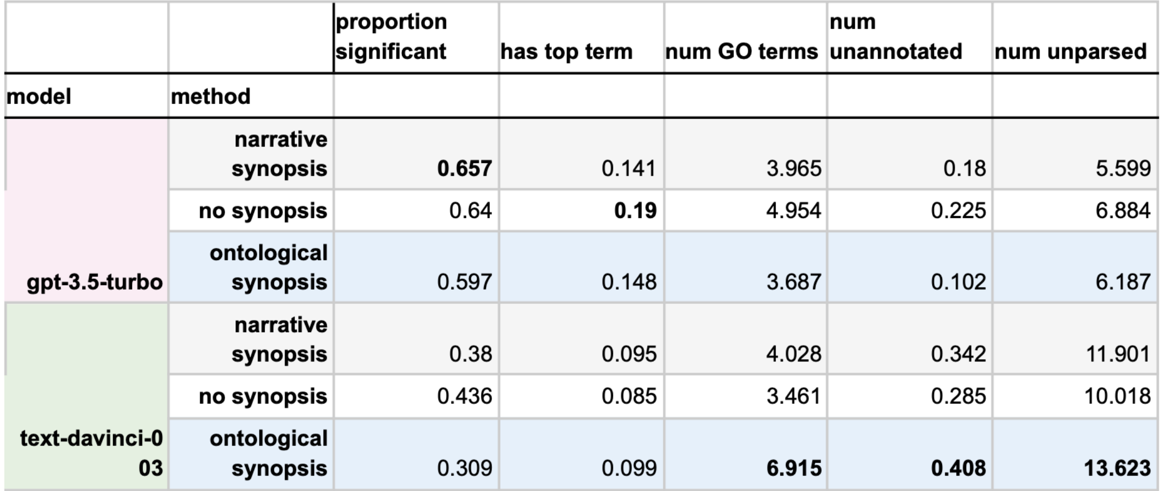
❑ Gold Standard: For each gene set, conduct standard gene set enrichment implemented in OAK, using hypergeometric tests and Bonferroni correction.

Evaluation: Metrics

Proportion of significant terms	How many GO terms returned by GPT are significant (p<0.05) in gold standard.
Has top term?	Are top GO terms in gold standard returned by GPT?
Number of GO terms in results	Measures number of terms from the prompt completion that could be grounded using the current GO vocabulary. (how "concise" the method is?)
Number of unannotated terms	GO terms that are neither directly nor indirectly used to annotate any of the genes in the gene set. (hallucination or may potentially reflect true gene function under-annotation)
Number of unparsed terms	The number of terms returned in the enrichment list that cannot be parsed (grounded) to a GO term identifier.

Results

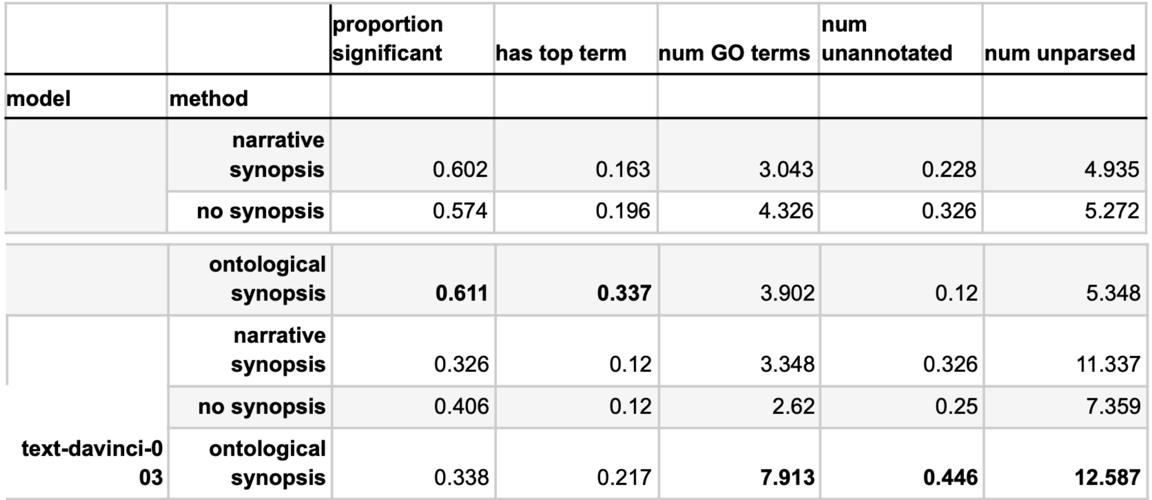
- ❑ Newer turbo model outperformed davinci.
- ❑ Model typically failed to return the top (most significant) term.
- ❑ Qualitative assessment of GPT summary: biologically plausible are often arbitrary and miss key terms that are often more informative.
- ❑ Sometimes the term returned by the GPT essentially means the same thing as the GO terms expected but can not be grounded.



		proportion significant	has top term	num GO terms	num unannotated	num unparsed
model	method					
gpt-3.5-turbo	narrative synopsis	**0.657**	0.141	3.965	0.18	5.599
	no synopsis	0.64	**0.19**	4.954	0.225	6.884
	ontological synopsis	0.597	0.148	3.687	0.102	6.187
text-davinci-003	narrative synopsis	0.38	0.095	4.028	0.342	11.901
	no synopsis	0.436	0.085	3.461	0.285	10.018
	ontological synopsis	0.309	0.099	**6.915**	**0.408**	**13.623**

Results: TF = 1

- ❑ For **smaller gene sets** with no input truncation, ontology-based synopses perform best.
- ❑ For the full range of gene sets, ranging in size up to 200 genes, the best approach is with no synopsis relying on the model's latent KB.
- ❑ Ontological synopses always yielded a low level of unannotated GO terms: avoiding hallucination or being to conservative.



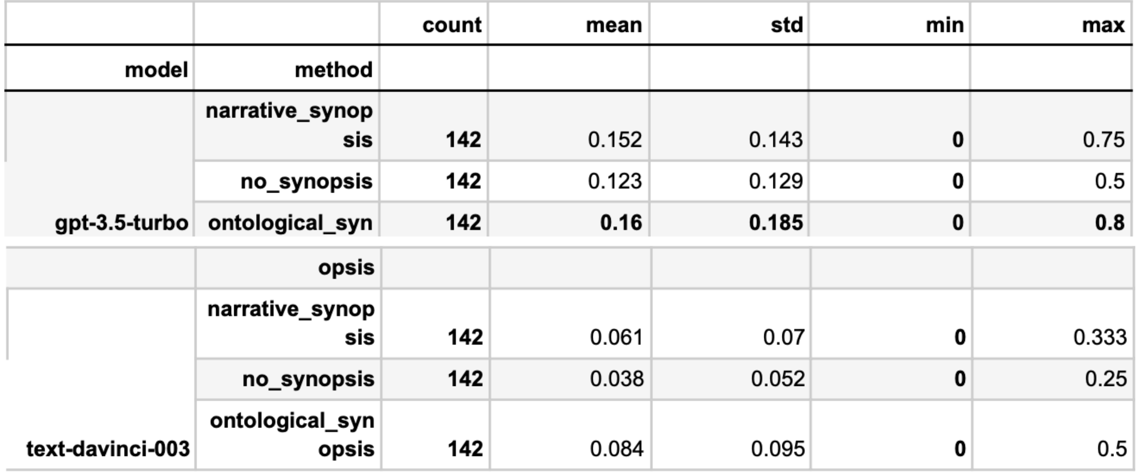
		proportion significant	**has top term**	**num GO terms**	**num unannotated**	**num unparsed**
model	**method**					
	narrative synopsis	0.602	0.163	3.043	0.228	4.935
	no synopsis	0.574	0.196	4.326	0.326	5.272
	ontological synopsis	**0.611**	**0.337**	3.902	0.12	5.348
	narrative synopsis	0.326	0.12	3.348	0.326	11.337
	no synopsis	0.406	0.12	2.62	0.25	7.359
text-davinci-003	**ontological synopsis**	0.338	0.217	**7.913**	**0.446**	**12.587**

Results: Stability of LLM (Ontology Terms)

- Measure the Jaccard similarity of the term sets of each run.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

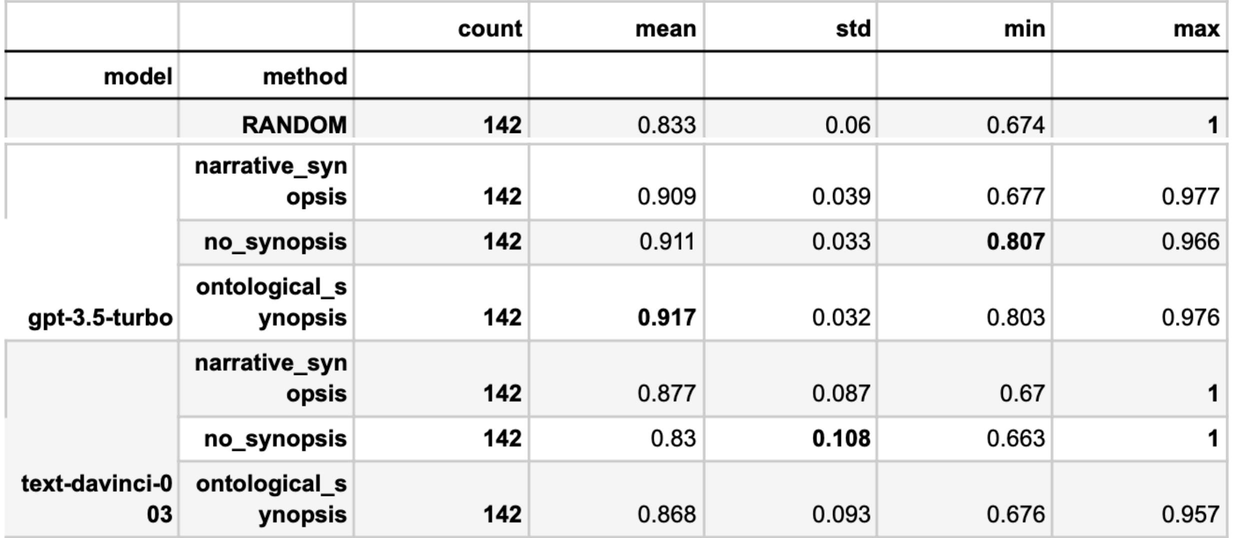
- There is a very low level of consistency across runs, with the most consistent being turbo with ontological synopses.



		count	mean	std	min	max
model	method					
gpt-3.5-turbo	narrative_synopsis	142	0.152	0.143	0	0.75
	no_synopsis	142	0.123	0.129	0	0.5
	ontological_synopsis	142	**0.16**	**0.185**	0	**0.8**
text-davinci-003	narrative_synopsis	142	0.061	0.07	0	0.333
	no_synopsis	142	0.038	0.052	0	0.25
	ontological_synopsis	142	0.084	0.095	0	0.5

Results: Stability of LLM (Narrative Summaries)

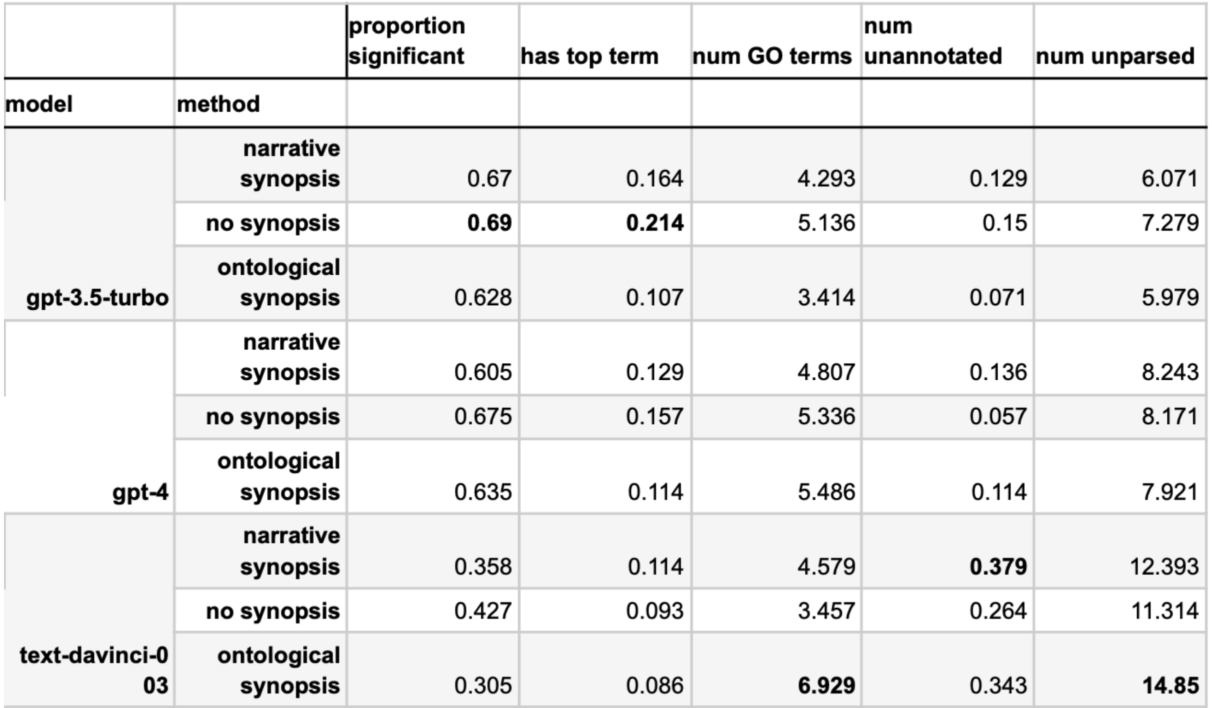
- ❑ Calculate the **cosine similarity of text embeddings** of descriptions using the OpenAI text-embedding-ada-002 model.
- ❑ Overall summaries generally varied quite widely, with turbo varying less widely than davinci.



		count	mean	std	min	max
model	**method**					
	RANDOM	**142**	0.833	0.06	0.674	**1**
gpt-3.5-turbo	**narrative_synopsis**	**142**	0.909	0.039	0.677	0.977
	no_synopsis	**142**	0.911	0.033	**0.807**	0.966
	ontological_synopsis	**142**	**0.917**	0.032	0.803	0.976
text-davinci-003	**narrative_synopsis**	**142**	0.877	0.087	0.67	**1**
	no_synopsis	**142**	0.83	**0.108**	0.663	**1**
	ontological_synopsis	**142**	0.868	0.093	0.676	0.957

Results: GPT4

- GPT-4 did not deliver major gains over the smaller turbo model.



		proportion significant	has top term	num GO terms	num unannotated	num unparsed
model	method					
gpt-3.5-turbo	narrative synopsis	0.67	0.164	4.293	0.129	6.071
	no synopsis	**0.69**	**0.214**	5.136	0.15	7.279
	ontological synopsis	0.628	0.107	3.414	0.071	5.979
gpt-4	narrative synopsis	0.605	0.129	4.807	0.136	8.243
	no synopsis	0.675	0.157	5.336	0.057	8.171
	ontological synopsis	0.635	0.114	5.486	0.114	7.921
text-davinci-003	narrative synopsis	0.358	0.114	4.579	**0.379**	12.393
	no synopsis	0.427	0.093	3.457	0.264	11.314
	ontological synopsis	0.305	0.086	**6.929**	0.343	**14.85**

Results: Hallucinations

❑ Aggregate all unannotated terms for all GPT results (these represent potential hallucinations). Then validate whether each term was descriptive for any gene in that gene set.

❑ Unable to detect any true hallucinations.

❑ Some summaries include reports of p-values (though not specifically asked for) that are fabricated (“sandbag” a researcher).

Results: Gene Symbols or In-Context Info

❑ To test whether the model was relying on gene symbols and its own latent KB of those genes, rather than the in-context information provided, swap out each gene description for a random gene description.

❑ The model uses the descriptions, and summarized these, ignoring the gene symbols.

Discussion: Limitations & Future Work

□ Due to constraints on the number of tokens in a single prompt, may not be feasible to provide **background genes.**

□ Hard to derive **statistics** to quantify the results.

□ Results are highly **non-deterministic**.

□ Inputs are **unordered gene sets**, not ranked lists (Like GSEA).

□ Do not make use of the **conversational abilities** of LLMs. (In the future, the users may be able to enter a dialog to transparently interact with multiple different biological KBs.)

Language models are not a shortcut to manual curation.

Thanks!!

Sijia Huo

Sep 27, 2023