**Automated Summarization of Human Protein Atlas Gene Pages Using Large Language Models**

Konstantinos Antonopoulos

DDLS course – Final Project

Stockholm, 2025-10-29

# **Abstract**

Large-scale omics resources such as the Human Protein Atlas (HPA) have mapped gene and protein expression across tissues, yet navigating this wealth of data remains cognitively demanding. Current portals lack concise, integrated overviews of gene function, localization, and disease relevance. Here, I develop an AI-based summarization agent that generates expert-style gene descriptions from structured HPA data using a locally deployable large language model. With advanced prompt engineering, the system generates concise, accurate summaries capturing tissue specificity, subcellular localization, and disease context. Automated and LLM-based evaluations confirm factual consistency and readability, limiting AI-hallucinations.

# **Background & Motivation**

The human proteome remains vast and only partially characterized, despite substantial progress in large-scale omics initiatives. Projects such as the Human Protein Atlas (HPA) have generated tissue, cellular-, and subcellular-level expression maps covering thousands of genes and proteins across human tissues and disease contexts, all of them publicly available [1–9]. However, the utility of these rich datasets is often inhibited by high cognitive load on users navigating multiple modalities and pages. It has become increasingly clear that summarizing multi-modal expression, localization and disease-association data into concise, biologically meaningful summaries rather than raw tables alone is critical to accelerate functional interpretation and translational discovery.

The HPA has made major progress toward accessibility by introducing per-gene summary pages and modality-specific overviews. Yet even these resources require users to scan through multiple sections and interpret tables to gain a coherent understanding of a gene’s biological role. For a biologist or clinician exploring unfamiliar genes, this can be time-consuming. A concise, automatically generated summary that integrates the most important information, such as tissue and cell-type expression, subcellular localization, and disease relevance, would provide an immediate overview and help users decide whether to explore the gene further.

Creating such summaries manually for over 20,000 genes is infeasible. Each gene integrates multiple layers of information across modalities, from RNA and protein expression to localization and disease profiling. This motivates the use of language models capable of handling structured and unstructured biological data. Recent advances in large language models (LLMs) have made it possible to run capable open-source models locally on personal hardware, reducing computational barriers while maintaining strong summarization performance [10]. These models can be fine-tuned or prompted effectively to produce short, factual, domain-specific text from complex inputs.

The goal of this project is therefore to design and evaluate an AI-based summarization agent that generates a concise, factual and integrated description for each HPA gene entry. The summaries will not replace detailed datasets but serve as a high-level entry point for researchers. They should be short enough to read in seconds, yet informative enough to highlight the key biological and pathological contexts of each gene. By combining structured HPA JSON data and domain-specific prompting, the system aims to emulate how a domain expert would concisely describe a gene to another researcher.

Beyond utility, this approach will also serve as a testbed for evaluating factual and abstractive summarization in biomedical data. It will examine how well current LLMs can synthesize multi-modal biological information, maintain accuracy and avoid hallucination. The resulting summaries could enhance user experience in the HPA portal by making gene exploration faster and more intuitive, representing a natural next step in the evolution of the Atlas from a data-rich to a user-centric knowledge platform.

# **Dataset Summary**

All data used in this project originate from the open-access Human Protein Atlas (HPA) portal [11]. The dataset was constructed using a custom script that, for each available gene, retrieved data from: https://www.proteinatlas.org/{gene\_ensembl\_id}.json. This file contains most HPA information per gene. Additional data from the Brain and Blood sections, present only in the web tables, were extracted through text mining and merged into the corresponding JSON entries. The resulting dataset is a dictionary with Ensembl gene IDs as keys and full gene-level JSON objects as values. No further preprocessing was required, as most fields consist of textual annotations describing various biological modalities.

The model was not fine-tuned. Instead, a few-shot prompting strategy was used: ten genes with manually curated expert summaries served as examples to illustrate ideal input–output pairs and guide the LLM’s generation behavior. These genes were chosen to represent a diverse range of biological contexts and data complexity.

Given the project’s time and resource constraints, summaries were generated for 100 representative genes selected from the full set of 20,162 available in HPA. To subset the genes, I first analyzed the dataset to identify key stratification fields, including the presence of PEA data, the number of diseases a gene is upregulated in, tissue specificity, cell type specificity and subcellular location (Supplementary Fig. 1-5A). I prioritized genes with PEA data and a broad distribution of upregulated diseases, ensuring that 75% of the sampled genes met these criteria. Genes without PEA data were included to make up the remaining 25% of the sample. Stratified sampling was performed based on these fields and any shortfall in the desired sample size of 100 genes was filled by randomly selecting from the remaining genes. This approach ensured that the final subset reflected the desired distributions while emphasizing genes with PEA data and diverse disease associations (Supplementary Fig. 1-5B). I also calculated the percentage of missing values in each category of variables-fields (Fig. 1). The full list of selected genes is provided in the Appendix.

# **Method Description**

## Baseline Approach

To generate the summaries, I utilized a locally deployable large language model (LLM) that can be run on a personal computer. Specifically, I selected GPT-oss-20b by OpenAI [10], as it offers a balance between lightweight deployment and competitive performance compared to larger models. For each gene, the model was provided with its corresponding input JSON along with the following prompt: "Summarize and comment the underlining trends (30-50 words)"

## Improved Approach

To generate the summaries, I utilized the same locally LLM, GPT-oss-20b. The process involved two steps. First, the model was prompted to produce a concise summary focusing on key biological aspects such as tissue and cell type specificity, subcellular location, diseases, pathways and secretome information. The prompt emphasized extracting relevant information directly from the input JSON data while maintaining scientific accuracy. In the second step, the model refined the initial summary by seamlessly integrating insightful comments to highlight underlying trends and connections between the biological features. This second step ensured the final summary was cohesive, scientifically accurate and stylized for a scientific audience.



**Figure 1. Percentage of missing values in the field-groups.** Groups consist of all variable-fields that are similar, for example Cancer prognostic group consists of all fields that have to do with cancer prognostics.

## Evaluation

The evaluation schema for assessing the quality of gene summaries integrates both automated metrics and a large language model (LLM)-based evaluator to ensure comprehensive and robust analysis. Automated metrics are designed to quantify specific aspects of the summaries, such as their lexical overlap with the input, factual consistency and conciseness [12]. For instance, ROUGE scores (ROUGE-1 and ROUGE-L) measure the overlap between the generated summary and the input, capturing recall-based alignment of n-grams and longest common subsequences, respectively. These metrics provide insight into how well the summary captures the key textual elements of the reference. Additionally, entailment scores assess the presence of key entities, such as tissues, cell types, and diseases, within the summary, while QA-based checks evaluate whether expected answers to structured questions derived from the gene data are present. A word count score ensures that summaries remain concise, penalizing overly verbose outputs.

Complementing these automated metrics, the LLM evaluator introduces a human-like layer of judgment, leveraging its ability to interpret and assess the summary holistically. The LLM is prompted to rate the factual accuracy and coherence of the summary on a scale of 1 to 5, accompanied by a concise comment explaining its reasoning. This evaluation captures nuances that automated metrics may overlook, such as logical consistency, subtle inaccuracies, or hallucinated information. Together, the automated metrics and LLM evaluation provide a balanced framework, combining quantitative rigor with qualitative insight, to ensure that the summaries are both factually reliable and effectively communicate the essential information.

# **Results**

The improved model showed slightly better performance across all metrics except QA match rate, which scored equally well in both models (Fig. 2A). QA match score and word count, the most important metrics, were consistently high, exceeding 80% for both models. Rouge-1 and Rouge-L scores were below 50% in both cases, as expected, given the input data format as JSON fields rather than full-text sections. The entailment score was around 50%, reflecting that while half of the important fields were captured, cases involving numerous tissues, cell types, or diseases often lacked full coverage. This is not necessarily problematic, as the model prioritizes conciseness while retaining essential information, as evidenced by the high QA match scores that evaluate critical details. The LLM evaluator scored similarly between the two models (Fig. 2B), with slightly lower scores for the improved version. Most discrepancies were due to mismatches in blood concentration units (e.g., ng/mL instead of ng/L), which can be easily corrected across all cases.

## A graph of a bar graph AI-generated content may be incorrect.

**A**

**B**

A graph of different shapes and sizes

AI-generated content may be incorrect.

**Figure 2. Model Comparison. A.** Automated summary metrics for each model. **B.** LLM evaluator scores for each model, with baseline scores represented in shades of blue and improved scores in shades of red.

I conducted a variance test by running the summary generation pipeline five times for the same gene using both the baseline and improved models. The automated metric results indicate high consistency across runs for both models, with a single outlier in the baseline model's word count metric, where one abstract was unusually long. Even though the metrics suggest that the improved and baseline models perform equally well, A/B testing with colleagues revealed a clear preference for the improved model. An example for comparison is the following:

**Baseline** (simple data repeat): HNMT shows strong liver and immune-cell expression, nucleoplasmic localization, and limited brain regional specificity. It’s a disease-variant gene, prognostic mainly in renal clear-cell and pancreatic adenocarcinoma, and is up-regulated in chronic liver disease, infections, and various cancers.

**Improved** (more concise, it also tries to connect the different modalities): HNMT, a cytosolic-nucleoplasmic histamine-degrading methyltransferase, is highly expressed in hepatocytes and monocyte-derived macrophages. Its specificity is linked to diseases where it is upregulated, such as chronic liver disease, viral hepatitis, HCC, and various infections.

# **Conclusion & Discussion**

This study demonstrates the feasibility of leveraging current lightweight large language models (LLMs) to generate abstractive summaries from structured JSON inputs. The structured nature of the data simplifies the summarization process, enabling even smaller-scale models to perform effectively. However, ensuring consistency in the output remains a significant challenge, particularly when requiring the LLM to adhere to strict formatting standards, such as generating evaluation scores. Addressing this issue may benefit from employing multiple LLMs in iterative workflows or adopting a mixture-of-experts approach to enhance reliability and robustness.

The evaluation of abstractive summaries presents another layer of complexity, as fully automating this process remains elusive. A practical solution in real-world applications could involve combining automated metrics with multiple LLM evaluators to flag problematic cases. Human reviewers could then focus on these flagged instances, ensuring quality control and correcting errors where necessary. This hybrid approach balances scalability with the need for precision.

Looking ahead, an intriguing avenue for exploration lies in advancing beyond simple summarization to enable integrative analysis. This would involve LLMs synthesizing data, applying logical reasoning, and identifying trends or insights. While preliminary attempts in this project show promise, further experimentation is required to refine this capability, ensuring outputs are both accurate and free from hallucinations. Such advancements could unlock new possibilities for data-driven discovery and interpretation.

# **Data & Code Availability**

All data used in this project are publicly available trough Human Protein Atlas portal: https://www.proteinatlas.org

All code used is documented and presented in: https://github.com/kantonopoulos/hpa-ai-summaries

# **Acknowledgments**

I would like to thank Evelina Sjöstedt and María Bueno Álvez for their help preparing some expert-summaries to provide the model with. I would also want to thank Mathias Uhlen for our interesting discussions regarding this project. I used ChatGPT for ideation regarding the project and how to orchestrate the LLM agent. The chat link is in the last appendix.

# **References**

1. Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, Mikes J, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science. 2019 Dec 20;366(6472):eaax9198.

2. Álvez MB, Bergström S, Kenrick J, Johansson E, Åberg M, Akyildiz M, et al. A human pan-disease blood atlas of the circulating proteome. Science. 2025 Oct 9;0(0):eadx2678.

3. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017 Aug 18;357(6352):eaan2507.

4. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single-cell type transcriptomics map of human tissues. Sci Adv. 2021 July;7(31):eabh2169.

5. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, et al. A subcellular map of the human proteome. Science. 2017 May 26;356(6340):eaal3321.

6. Sjöstedt E, Zhong W, Fagerberg L, Karlsson M, Mitsios N, Adori C, et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. Science. 2020 Mar 6;367(6482):eaay5947.

7. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

8. Jin H, Zhang C, Zwahlen M, von Feilitzen K, Karlsson M, Shi M, et al. Systematic transcriptional analysis of human cell lines for gene expression landscape and tumor representation. Nat Commun. 2023 Sept 5;14(1):5417.

9. Uhlén M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, et al. The human secretome. Sci Signal. 2019 Nov 26;12(609):eaaz0274.

10. OpenAI, Agarwal S, Ahmad L, Ai J, Altman S, Applebaum A, et al. gpt-oss-120b & gpt-oss-20b Model Card [Internet]. arXiv; 2025 [cited 2025 Oct 25]. Available from: http://arxiv.org/abs/2508.10925

11. The Human Protein Atlas [Internet]. [cited 2025 Oct 25]. Available from: https://www.proteinatlas.org/

12. Yan Z. Evaluation & Hallucination Detection for Abstractive Summaries. eugeneyan.com [Internet]. 2023 [cited 2025 Oct 25]; Available from: https://eugeneyan.com/writing/abstractive/

**Appendix**

## List of Subsetted Genes

'ENSG00000100450', 'ENSG00000030419', 'ENSG00000124767', 'ENSG00000104921', 'ENSG00000173714', 'ENSG00000105854', 'ENSG00000150540', 'ENSG00000172232', 'ENSG00000166145', 'ENSG00000173992', 'ENSG00000078098', 'ENSG00000149968', 'ENSG00000104267', 'ENSG00000266967', 'ENSG00000150782', 'ENSG00000095739', 'ENSG00000000938', 'ENSG00000108700', 'ENSG00000170909', 'ENSG00000168811', 'ENSG00000130234', 'ENSG00000188643', 'ENSG00000177106', 'ENSG00000103066', 'ENSG00000146013', 'ENSG00000145220', 'ENSG00000231924', 'ENSG00000162692', 'ENSG00000116809', 'ENSG00000125538', 'ENSG00000169429', 'ENSG00000173918', 'ENSG00000165973', 'ENSG00000122584', 'ENSG00000107821', 'ENSG00000108187', 'ENSG00000184613', 'ENSG00000142192', 'ENSG00000135047', 'ENSG00000007312', 'ENSG00000185985', 'ENSG00000164951', 'ENSG00000142798', 'ENSG00000151651', 'ENSG00000178498', 'ENSG00000100100', 'ENSG00000149564', 'ENSG00000148346', 'ENSG00000101405', 'ENSG00000170775', 'ENSG00000106541', 'ENSG00000120217', 'ENSG00000240505', 'ENSG00000204516', 'ENSG00000115602', 'ENSG00000079101', 'ENSG00000112116', 'ENSG00000219438', 'ENSG00000132026', 'ENSG00000158714', 'ENSG00000091972', 'ENSG00000132330', 'ENSG00000185070', 'ENSG00000089250', 'ENSG00000079112', 'ENSG00000100097', 'ENSG00000167419', 'ENSG00000164111', 'ENSG00000108691', 'ENSG00000136634', 'ENSG00000145494', 'ENSG00000185008', 'ENSG00000089902', 'ENSG00000127191', 'ENSG00000008277', 'ENSG00000139780', 'ENSG00000129925', 'ENSG00000277865', 'ENSG00000225190', 'ENSG00000182472', 'ENSG00000176256', 'ENSG00000211788', 'ENSG00000180957', 'ENSG00000234438', 'ENSG00000185115', 'ENSG00000105668', 'ENSG00000162777', 'ENSG00000113073', 'ENSG00000182601', 'ENSG00000166266', 'ENSG00000167968', 'ENSG00000274070', 'ENSG00000118017', 'ENSG00000134780', 'ENSG00000197785', 'ENSG00000099937', 'ENSG00000164953', 'ENSG00000103222', 'ENSG00000165650', 'ENSG00000089335'

## Supplementary Figures – Dataset Distributions

## 

**A**

**B**

## 

**Supplementary Figure 1.** **Number of genes with or without PEA data. A.** Distribution of PEA data across all genes in the complete dataset, where over 60% of genes lack PEA data. **B.** Distribution of PEA data in the subsetted sample, where 75% of genes are enforced to have PEA data to ensure representation in summaries and alignment with disease distributions.

## 

**A**

**B**



**Supplementary Figure 2.** **Number of genes upregulated in various diseases.** **A.** All genes. **B.** Subsetted genes.

## 

**A**

**B**



**Supplementary Figure 3.** **Number of genes and their tissue specificity category.** **A.** All genes. **B.** Subsetted genes.

## 

**A**

**B**



**Supplementary Figure 4.** **Number of genes and their cell type specificity category.** **A.** All genes. **B.** Subsetted genes.

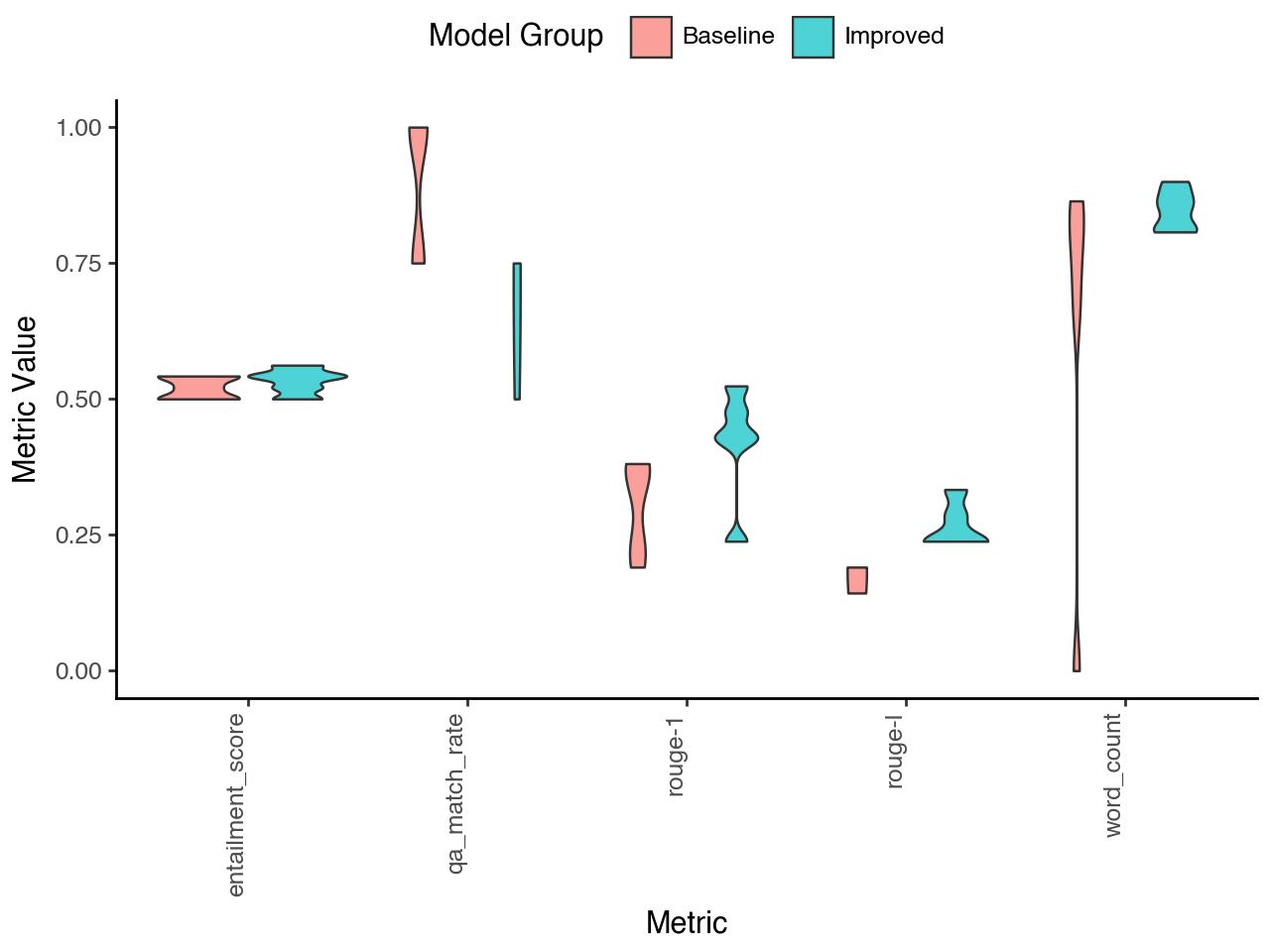
## 

**A**

**B**



**Supplementary Figure 5.** **Number of genes in a specific subcellular location.** **A.** All genes. **B.** Subsetted genes.



**Supplementary Figure 6.** **Comparison of variance in model outputs.** Comparison of model outputs for the same gene across five runs between baseline and improved models.

## ChatGPT Chat Link

https://chatgpt.com/share/6900c6b3-14c0-800d-a25a-6afa69e71da1