Scalable Extraction and Curation of Glycan Biomarker Data from Biomedical Literature Using Large Language Models

by Cyrus Chun Hong Au Yeung

B.S. in Biochemistry and Chemistry, July 2022, The University of Hong Kong

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

May 18, 2025

Thesis directed by

Raja Mazumder
Professor of Biochemistry and Molecular Medicine

Robel Kahsay
Associate Professor of Biochemistry and Molecular Medicine

© Copyright 2025 by Cyrus Chun Hong Au Yeung
All rights reserved

Dedication

Dedicated to my parents. Their support gave me the courage to embrace change and uncertainty. This journey of studying abroad would not have been possible without them.

Acknowledgments

The author would like to express his sincere gratitude to those who have supported him throughout his thesis journey:

I am deeply grateful to my advisor, Professor Raja Mazumder, for his unwavering support and guidance. I especially appreciate his understanding and compassion during periods when I was managing personal challenges.

I extend my sincere appreciation to Professor Robel Kahsay – co-investigator of GlyGen – for his generous support. I thank him for providing the OpenAI API prototype script and offering weekly consultation hours to address my thesis questions. His guidance in programming and algorithmic thinking significantly helped me in solving problems.

I would also like to thank Jeet Vora, whose mentorship helped me navigate the planning and writing of this thesis. He has provided invaluable insights on glycobiology and data curation.

I am thankful to past and present members of the GW GlyGen team: Karina Martinez and Jingyue Wu, for offering training in backend development; and Luke Johnson, Urnisha Bhuiyan, and Kate Warner, for their efforts in data harmonization, which laid the groundwork for my project.

Abstract of Thesis

Extraction and Curation of Glycan Biomarker Data from Biomedical Literature Using Large Language Models

Automated extraction and curation of glycan biomarker data from biomedical literature are crucial for synthesizing discrete findings into generalized glycan-disease associations. This thesis presents a scalable named entity recognition (NER) and relation extraction (RE) pipeline utilizing state-of-the-art large language models (LLMs), specifically GPT-based models, to systematically identify glycans, glycoproteins, glycoenzymes, diseases, and experimental contexts from a corpus of 4,595 glycan biomarker-related PubMed abstracts, yielding ~300 distinct glycan biomarker entries. The pipeline accurately converts complex biomarker relationships described in unstructured natural language texts into structured data adhering to a custom-designed JSON schema, enabling efficient integration into downstream analytical processes such as clustering and embedding analyses. A high throughput of 55 curated abstracts per minute implies a potential replacement for manual annotation efforts traditionally required for glycoinformatics research. Distinct from previous text mining efforts which relied primarily on domain-specific NLP approaches, this method leverages the contextual understanding and flexibility of general-purpose LLMs. The pipeline offers unique advantages, including strict schema compliance and efficient parallel processing, hence facilitating large-scale biomedical literature mining without requiring significant local computational resources. By enabling comprehensive curation of glycan-disease relationships, this study provides a robust foundation for hypothesis generation and glycan-associated disease mechanisms.

Table of Contents

Dedication iii
Acknowledgements iv
Abstract of Thesis v
List of Figures viii
List of Tables ix
List of Symbols / Nomenclature x
Chapter 1: Introduction 1
1.1. Biological Importance of Glycosylation 1
1.2. Common Glycan Biomarkers in Disease 2
1.3. Challenges in Glycan Biomarker Research 4
1.4. Need for Advanced Computational Approaches 5
1.5. Emergence of Large Language Models (LLMs) 6
1.6. Objectives and Workflow of this Thesis 8
Chapter 2: Literature Review 10
2.1. Historical Perspectives on Glycan Biomarkers 10
2.2. Evolution and Standardization of Glycan Structure Notations 11
2.3. Existing Efforts in Glycan Data Integration 13
2.4. Biomedical Text Mining Methodologies 17
2.5. Named Entity Recognition (NER) in Biomedical NLP 19
Chapter 3: Methods 22
3.1. Literature Retrieval from PubMed 22
3.2. Unsupervised Abstract Clustering 24
3.3. Named Entity Recognition and Relation Extraction 24
3.4. Validation of Named Entity Recognition Accuracy 26
3.5. Fine-Tuning of Large Language Models 27
3.6. Standardization of Glycan Terminology 27
3.7. Generation of Embeddings for Glycan Terms and Abstract Sentences 28
3.8. Embedding-Based Clustering and Visualization 29
3.9. Computational Environment and Implementation 30

Chapter 4: Results 31
4.1. Search Query for Literature Retrieval 31
4.2. Characteristics of Retrieved Abstracts 32
4.3. Results of Abstract Clustering 33
4.4. JSON Schema and Prompt Design 35
4.5. Example Output of Named Entity Recognition and Relation Extraction 36
4.6. Summary of named entity recognition and relation extraction 37
4.7. Standardized Glycan Terminology Results 40
4.8. Visualization of Glycan Term Clusters 41
4.9. Outcomes of Model Fine-Tuning 43
Chapter 5: Discussions 44
5.1. Interpretation and Implications of Results 44
5.2. Comparison with Traditional (Non-LLM) NLP Approaches 46
5.3. Technical Challenges and Limitations 48
5.4. Unique Advantages of GPT Models in Biomedical Text Mining 50
5.5. Prompt Engineering as a Critical Factor 51
5.6. Future Directions and Opportunities 55
Figures 57
Tables 67
References 72
Appendices 80

List of Figures

Figure 1. Common disease-associated glycan structural changes. 57

Figure 2. Automated workflow for glycan biomarker extraction. 58

Figure 3. Publication trends for glycan-biomarker query and selected keywords. 59

Figure 4. Determining the optimal number of abstract texts. 60

Figure 5. K-means clustering plots (k = 3 and k = 6) of abstract texts. 61

Figure 6. Example glycan-disease relation extraction. 62

Figure 7. Summary of the extracted entities. 63

Figure 8. Top standardized glycan entities in the curated corpus. 64

Figure 9. UMAP projection of extracted glycan terms. 65

Figure 10. Training loss curve for GPT-4o model fine-tuning. 66

List of Tables

Table 1. Classification of glycan notations and nomenclatures. 67

Table 2. Representative terms for each abstract cluster (k = 3). 69

Table 3. Representative terms for each abstract cluster (k = 6). 69

Table 4. Summary of extracted glycan entities. 70

List of Symbols / Nomenclature

Glycobiology

ER – Endoplasmic Reticulum

GH – Glycoside Hydrolase

GPI – Glycosylphosphatidylinositol

GT – Glycosyltransferase

Common Monosaccharide Codes

Fuc – Fucose

Gal – Galactose

GalNAc – N-acetylgalactosamine

GlcNAc – N-acetylglucosamine

Man – Mannose

Neu5Ac – N-acetylneuraminic acid (the most common sialic acid in humans)

Sia – Sialic acid (a broad family term for acidic 9-carbon sugars)

Bioinformatics and Genomics

CAZy – Carbohydrate-Active Enzymes Database

DOID – Disease Ontology Identifier

HGNC – HUGO Gene Nomenclature Committee

IUPAC – International Union of Pure and Applied Chemistry

KEGG – Kyoto Encyclopedia of Genes and Genomes

LINUCS – Linear Notation for Unique description of Carbohydrate Sequences

MeSH – Medical Subject Headings

OMIM – Online Mendelian Inheritance in Man

PDB – Protein Data Bank

SNFG – Symbol Nomenclature for Glycans

WURCS – Web3 Unique Representation of Carbohydrate

Machine Learning and Natural Language Processing

API – Application Programming Interface

BERT – Bidirectional Encoder Representations from Transformers

EHR – Electronic Health Record

GMM – Gaussian Mixture Model

GPT – Generative Pre-trained Transformer

JSON – JavaScript Object Notation (a file format for structured data)

LLM – Large Language Model (e.g. GPT-4, BioBERT)

NER – Named Entity Recognition

NLP – Natural Language Processing

PCA – Principal Component Analysis

POS – Part-of-Speech

RE – Relation Extraction

TF-IDF – Term Frequency-Inverse Document Frequency

UMAP – Uniform Manifold Approximation and Projection

WSS – Within-Cluster Sum of Squared Errors (alternatively: Within-Cluster SSE)

Chapter 1: Introduction

1.1. Biological Importance of Glycosylation

The molecular foundations of life have long been explored through the lenses of DNA and proteins, but only recently have complex sugars emerged as essential regulators of cellular function and communication. Glycosylation is a ubiquitous post-translational modification wherein these sugars, also known as glycans, are covalently linked to proteins or lipids. This type of modification occurs in nearly all living organisms and is essential in enriching the structural and functional diversity of biomolecules[1]. The predominant types of glycosylation include N-linked glycosylation, where glycans attach to the nitrogen atom of asparagine residues, and O-linked glycosylation, where glycans attach to the oxygen atom of serine or threonine residues. Glycans can also link to other amino acids such as tryptophan or lipid molecules such as phosphatidylinositol.

Unlike the linear templates of DNA and RNA, glycan structures are branched and non-template-driven. The biosynthesis of glycans is instead orchestrated by a complex network of enzymes involving glycosyltransferases (GTs) and glycosidases (GHs)[2]. These enzymes sequentially add and trim monosaccharide units. This gives rise to an extraordinary range of molecular configurations, suggesting that glycans may embed large-scale molecular information[3]. The biosynthetic pathways are compartmentalized within the endoplasmic reticulum (ER) and Golgi apparatus[4]. While N-linked glycans are initially synthesized in the ER and mature in the Golgi apparatus, O-linked glycans are predominantly synthesized in the Golgi apparatus[4]. The expression of specific glycan structures is cell type- and tissue-specific[5], influenced by the expression levels of glycosylation-related genes.

Glycosylation plays multiple pivotal roles in human biology. It assists in proper protein folding and enhances stability, preventing aggregation and facilitating quality control mechanisms within the ER[6]. Glycans on cell surfaces mediate interactions through recognition by glycan-binding proteins (lectins), influencing processes like immune responses and inflammation. Pathogens exploit host glycan structures for attachments and entry, while alterations in glycosylation can aid in immune evasion[6]. Additionally, glycosylation modulates signaling pathways critical for embryonic development and cell differentiation, such as the Notch signaling pathway[7].

1.2. Common Glycan Biomarkers in Disease

Glycan alterations are one of the earliest biochemical changes that occur when signaling networks governing development and cell-type specification become dysregulated. Again, since glycosylation is not template-driven, subtle changes in the expression or localization of individual GTs and GHs propagate into reproducible, disease-specific glycan signatures that can be captured analytically and translated into clinical biomarkers. There are several high-quality reviews[7-10] covering well-established glycan biomarkers with a primary focus on cancers; this section covers a high-level overview of the two major classes of diagnostic glycan biomarkers, namely the mucin-type O-linked glycans and N-linked glycans (Figure 1).

In the scope of O-linked glycans, diseases are commonly accompanied by premature chain termination[9]. Loss of core extension or capping often exposes the monosaccharide GalNAcα1-O-Ser/Thr (Tn antigen) and its sialylated derivative (STn antigen), or leaves the disaccharide Galβ1-3GalNAcα1-Ser/Thr (T antigen) and its

sialylated derivative (ST antigen) on mucins such as MUC1 and MUC2[11]. These truncated glycans form epitopes that are virtually absent in healthy adult epithelia yet are overexpressed in breast, colon, stomach, and ovary carcinomas[8], linking to invasion, chemoresistance, and poor prognosis[12]. Apart from cancer, the accumulation of Tn on IgA1 is also the hallmark of IgA-nephropathy[13], while diffuse STn expression in colonic mucins accompanies disease severity in ulcerative colitis[14].

Glycan extension represents a second group of O-linked biomarkers: terminal sialyl-fucosyl epitopes that decorate mucins and select glycosphingolipids[9]. Sialyl-Lewis a (CA 19-9) and sialyl-Lewis X (generated by the sialyltransferase ST3Gal and fucosyltransferases FUT3/6) mediate leukocyte trafficking but are utilized by malignant and infected cells to engage selectins, enhance motility, and induce metastases[15]. Notably, CA 19-9 is currently the only FDA-cleared serological test for pancreatic cancer and is widely used to monitor upper GI malignancies[16].

While O-glycan biomarkers often involve the types of glycan cores, N-glycan biomarkers usually reflect the degree of processing in the ER. Oncoproteins and acute-phase glycoproteins report increases in β1,6-GlcNAc branching, terminal α2,6-sialylation, and core α1,6-fucosylation[10]. These modifications enhance the hydrodynamic profile of the glycoprotein, prolonging its half-life and modulating receptor clustering[17]. In hepatology, the core-fucosylated α-fetoprotein isoform outperforms total AFP for early hepatocellular-carcinoma detection[18], while fucosylated haptoglobin and complement protein C3 have emerged as monitoring biomarkers[19]. Additionally, highly branched and hypersialylated N-glycans on prostate-specific antigen

(PSA), periostin, or integrins enhance the specificity of existing protein assays across different cancers[20,21].

1.3. Challenges in Glycan Biomarker Research

The repeated appearance of a small set of glycan motifs in cancer, infectious, and metabolic disorders shows that abnormal glycosylation is a common biochemical thread[22]. Turning these motifs into reliable biomarkers requires a side-by-side comparison of disease-specific changes and the assignment of each change to its signaling or biosynthetic pathways. Progress is slowed by fragmented nomenclature systems, incompatible nomenclatures, and a lack of glycan functional annotations that link a structure to the enzyme, compartment, or glycoprotein it involves[23].

The same glycan may be written as a monosaccharide symbol such as GlcNAc, a composition like Hex(5)HexNAc(4)Fuc(1), an IUPAC string[24], a LINUCS[25] or GlycoCT[26] record, a WURCS[27] line, an SNFG[28] glyph, a GlyTouCan ID[29], a composition-based N-glycan code such as FA2G2, or a text string such as "core-fucosylated bi-antennary complex N-glycan". Antigen names (sialyl-Lewis X, Tn), glycolipid tags (GM3, Gb3), and well-circulated abbreviations (sLe^x, TF) further enhance the variability of describing glycan structures and motifs. Databases and software rarely accept more than one or two notations, forcing researchers to translate manually and often losing stereochemical detail or branch order. Because most vocabularies stop at the structure, they offer little help in mapping motifs to pathways or signaling events. These obstacles create a strong case for text mining pipelines that can harvest glycan terms from the literature, reconcile the diverse notations, and enrich each entry with metadata such as protein attached, glycoenzyme, and cell type.

Automated extraction and normalization will let investigators integrate biomarker evidence across fields, highlight missing annotations, and prioritize motifs for experimental follow-up. Table 1 lists the principal naming systems now in use and demonstrates how the coexistence of numerous notations and the gaps in pathway mapping may hamper glycan biomarker research.

1.4. Need for Advanced Computational Approaches

The sheer number of glycobiology literature and the existence of more than a dozen notation systems make manual curation of glycan biomarkers labor-intensive. Early efforts at harmonizing glycan structural representations such as GLYDE[30] and the GlycoCT-centered GlycomeDB integrator[31] solved a significant part of the formatting problem, but the majority of structures are still left stranded in free-text articles. In addition, with thousands of PubMed papers mentioning “sialyl-Lewis X” or “core-fucosylated”, scalable text mining pipelines become desirable.

Traditional named entity recognition (NER) tools built for genes or chemicals miss many mentions of glycans because a single molecule may appear as “FA2G2S2”, “Neu5Acα2-6Galβ1-4GlcNAc…”, “Hex(5)HexNAc(4)Fuc(1)”, or “core-fucosylated bi-antennary N-glycan”. Studies using PubAnnotation[32] (a text annotation ecosystem) demonstrate that dictionaries tuned for glycobiology significantly increase recall, although rule-based pipelines still struggle when novel abbreviations appear[33]. Deep-learning language models that treat glycans as tokens, such as Glycowork[34], offer a way to recognize unseen glycan representations by learning substructure patterns directly from sequence strings.

Beyond finding the term, curators need sentence-level evidence that a motif is a biomarker, plus links to its disease context and the glycoenzyme involved. General biomedical relation-extractors such as CoINNER[35] and BioContext[36] recover gene-chemical-disease triples, but they lack glycan ontologies and therefore normalize "Tn antigen" as a generic chemical. Recent language model frameworks that fine-tune BERT-family encoders for biological interaction mining[37] demonstrate near-manual accuracy and can be adapted to glycan-disease-enzyme relations once a suitable training corpus is assembled.

Combining notation-agnostic NER and transformer-based relation extraction will allow researchers to keep pace with the literature and input high-confidence glycan-disease linkages into biomarker discovery pipelines. Advanced NLP-based computational approaches with high curation standards are becoming an indispensable part of efforts to turn the fragmented textual record into a map of glycan biomarkers.

1.5. Emergence of Large Language Models (LLMs)

With the advancement of LLMs, pre-trained language models have become central to biomedical NLP. Models pre-trained on billions of text corpus of webpages or biomedical resources provide contextualized language understanding that can be fine-tuned for various tasks. These biomedical LLMs have demonstrated excellent performance on classic text mining tasks including NER, relation extraction (RE), and question answering (QA)[38]. For example, BioBERT[38] and PubMedBERT[39] achieved top-tier accuracy on gene/disease NER and protein-chemical interaction extraction by having learned biomedical terminology and syntax from biomedical corpora like PubMed.

In parallel to domain-specific models, general-purpose LLMs like ChatGPT[40] (based on models including GPT-4 developed by OpenAI) and Llama[41] (or LLaMA, a group of models developed by Meta AI) have emerged, showcasing the ability to perform multiple MLP tasks through prompting alone. ChatGPT, based on the GPT series, was trained on a broad internet corpus (including some biomedical text) and can follow natural language instructions to solve problems. While ChatGPT is not specialized for biomedical research, it offers remarkable flexibility. For example, it can be given definition of task(s) or a few examples to generating answers or structured outputs without additional training. This opened new possibilities for biomedical text mining. Instead of developing a separate trained model for each text mining task, one can prompt ChatGPT to perform the tasks simultaneously. For instance, one can supply ChatGPT with a PubMed abstract and ask it to "identify all glycan-related entities and their relationships to diseases", essentially a NER-RE task, and request the output in a structured data format. In particular, recent GPT-based models can output schema-guided results, such as JSON records, when instructed with a predefined schema[42].

In the context of glycobiology, a glycan biomarker schema was designed in this thesis project to capture key entities and relations of interest. The schema included fields for glycan names, associated glycoproteins (the proteins expressing the glycan biomarkers), glycoenzymes (enzymes that alter the expression of the glycan biomarkers), relevant diseases or conditions, sample types (tissue or cell context), and evidence sentences in the PubMed corpus. The model, given instructions and optionally examples, will attempt to parse the abstract and fill in the correct information under each field. An advantage of using OpenAI's GPT models in the way is that it can flexibly handle

complex, nested information and output it in a machine-readable JSON without explicit programming of extraction rules. Moreover, it can incorporate both context and reasoning, such as excluding the enzyme PNGase F since it is a glycoenzyme used in glycomics experimental pipelines instead of biologically causing the change in glycan expression.

General LLM models like GPT-4 were pre-trained on general web data before a cut-off date and hence may lack some most-updated domain knowledge or precision that domain-specific models have, while careful prompt engineering is required for it to generate desired output and to minimize hallucinations[43] (incorrect outputs not grounded in the text). However, LLMs feature few-shot adaptability[40], where they can quickly adapt to new schemas or extraction targets with just a few example demonstrations, whereas traditional models would require retraining on a labeled dataset.

1.6. Objectives and Workflow of this Thesis

This thesis project leveraged OpenAI's chat completions API for NER and RE in the glycobiology literature. The ability of GPT-based models to generate consistent structured outputs was harnessed, outputting results as JSON objects conforming to the glycan biomarker schema. Each JSON object encodes an array of putative biomarker relationships in a given abstract. This approach enables the aggregation of extracted data across many publications. The structured outputs can be directly fed into downstream analytical workflows, which include clustering glycan biomarkers by disease or constructing a network of glycan-disease interactions. A detailed workflow of the pipeline is outlined in Figure 2.

In the following chapters, this thesis will provide a general overview of common glycan nomenclatures identified from glycobiology literature and glycan terms and their relationships to diseases in respective of altered metabolic pathways. Then, in the Methods and Results section, we will explore how applying LLM models in a schema-driven information extraction setting can accelerate knowledge discovery in glycobiology, a domain that benefits greatly from improved text mining due to its complex nomenclature and expanding literature. This sets the foundation for glycan-focused text mining and the development of a pipeline for automated text mining for glycan biomarkers.

Chapter 2: Literature Review

This chapter aims to give a more detailed summary of glycan biomarker research, covering its conceptual origins, the evolving systems used to represent glycan structures, and the efforts to harmonize these representations across databases. On the other hand, the progression of computational techniques is covered, particularly in natural language processing (NLP) which enables precise and scalable identification of biological entities and their contextual relationships. Together, these developments underscore the opportunity to apply language models to automate the extraction and curation of glycan-disease associations from biomedical text, leading to the conceptualization of this project.

2.1. Historical Perspectives on Glycan Biomarkers

Key biomarkers consisting of carbohydrate structures were revealed in the early discoveries in immunology. The ABO blood group system, identified by biologist Karl Landsteiner in 1901, was the first human blood group antigen system to be recognized[44]. We now know that the ABO blood types (A, B, O(H)) correspond to the specific glycan structures on red blood cells, although at the time of discovery, the chemical nature of the antigens was unknown. It was only decades later when the structural and genetic basis of the A, B, and H antigens were elucidated[45], highlighting that these early nomenclatures ("A" and "B" antigens) were not descriptive of the molecular structures.

In many cases, the glycan-based biomarkers were named based on the phenotype, the order of discovery, and the name of the discoverer. By the late 20th century, attention turned to the potential of glycans as disease biomarkers. A landmark example is Carbohydrate Antigen 19-9 (CA 19-9), identified by Koprowski et al. in 1979 using a

monoclonal antibody that recognized colorectal cancer cells[46]. CA 19-9 was the first widely used glycan-based tumor marker; it was later determined to be the "sialyl Lewis A" tetrasaccharide on a glycoprotein[47]. Notably, the name "CA 19-9" comes from the hybridoma clone number and does not convey the glycan structure. This was typical of early glycan biomarkers. In fact, CA 19-9 had been routinely used as a pancreatic cancer marker long before its structure was unveiled[46]. Similarly, other glycan-based antigens were given trivial names (e.g. Thomsen-Friedenreich antigen or T antigen) that lacked structural description. These historical cases show the first glycan biomarkers were recognized and named in the clinic, but their nomenclature was often *ad hoc* and disconnected from their glycan structures.

2.2. Evolution and Standardization of Glycan Structure Notations

The term "glycobiology" was coined much later in the late 1980s when the advance in analytical technology supported the systematic studies of glycan structures and functions[48]. As the glycobiology matured, formal systems were developed to name and depict complex carbohydrate structures. "IUPAC Nomenclature of Carbohydrates (Recommendations 1996)" provided a systematic textual representation of glycan structures. The IUPAC system assigns each monosaccharide a short code and specifies linkages with numeric locants[24]. In the extended format, each residue is prefixed by its anomericity (α/β) and stereochemistry (D/L), while prefixed by its ring size (*p* for pyranose/*f* for furanose). The glycosidic linkages are specified. For example, the IUPAC extended notation for raffinose is:

α-D-Gal*p*-(1→6)-α-L-Glc*p*-(1↔2)-β-D-Fru*f*

[24]. While precise, this format is notoriously cumbersome for routine use. The simplified condensed IUPAC format omits certain repetitive details to create a more compact representation. Taking the raffinose example again, its condensed notation is:

Gal(α1-6)Glc(α1-2β)Fru*f*

[24], which is shorter though still requires substantial field knowledge to interpret. To further simplify the notation, the short form is offered which further assumes the linkage positions of anomeric carbons. The IUPAC system is rigorous and hence unambiguous, but its complexity limited its everyday use by researchers, prompting the development of more user-friendly notations.

To further improve human readability, scientists developed symbolic representations of glycans. The Symbol Nomenclature for Glycans (SNFG) emerged as a community standard for drawing glycans with colored shapes[28,49]. In SNFG, each monosaccharide is depicted as a distinct symbol (square, circle, triangle, etc., with their colors indicating sugar type). For example, N-acetylglucosamine (GlcNAc) is depicted as a filled blue square, and mannose (Man) is depicted as a filled green circle. This visual alphabet, combined with linkage lines, allows complex glycans to be depicted graphically and in a standardized way. The illustration of a glycan structure represented by the SNFG can be comprehended quickly and intuitively. First published in 2015[49], SNFG has been widely adopted in publications and databases to ensure consistent communication of glycan structures, making glycobiology data more accessible and interchangeable across fields.

In parallel, the burgeoning glycoinformatics field necessitated computer-readable formats for glycans. Several notations have been developed to encode glycans in open-

text formats that capture the complexity of experimentally derived structures and facilitate software parsing. One widely used format is GlycoCT, introduced in 2008 as a unifying sequence representation for glycans[26]. GlycoCT utilizes a connection table (CT) approach – it encodes monosaccharide building blocks and linkages in a tabular format with separate sections for residues, connections, and optionally, repeating units and undetermined substructure linkages. This allows for complex branched glycans with missing structural information to be represented precisely. Another modern machine-readable format is WURCS (Web3 Unique Representation of carbohydrate Structures), introduced in 2014 to address the need for a unique text encoding for each glycan structure[27,50]. Unlike some earlier notations, WURCS enforces canonical ordering, i.e. using deterministic rules to sort residues, to ensure that the same glycan structure will always generate the same unique linear string.

Both GlycoCT and WURCS are machine-readable and can capture incomplete or ambiguous glycan structures, making them suitable for experimental data and computational analysis. These formats are integral to glycobioinformatics tools, enabling software and databases to exchange glycan data reliably.

2.3. Existing Efforts in Glycan Data Integration

The diversity of glycan names and notations has driven initiatives to integrate glycan information across different domains and systems. Major efforts have been made in the development of a centralized resource for glycan structures. First introduced in 2016, GlyTouCan[51] is an international glycan structure repository that assigns a unique accession number (GlyTouCan ID) to every glycan submitted by curators or users. In the

current 3.0 version, GlyTouCan allows researchers to submit structures in any common format translatable to WURCS and drawn SNFG representations[29], and the repository will immediately generate a stable identifier (e.g. G17689DH) that represents that structure in a canonical form. The assignment of a unique identifier can then be used as a reference across databases and publications. By design, this abstracts the notation details, as two laboratories will refer the same glycan to the same GlyTouCan ID even though one uses IUPAC naming and the other uses WURCS. Such abstraction greatly facilitates integration – for example, if a glycoinformatics database links a unique glycan ID to a glycoprotein, while a clinical dataset links the same ID to a disease biomarker, one can interrogate cross-disciplinary glycan information and disease linkages via the shared accession. Notably, GlyTouCan is capable of handling glycan ambiguity by assigning different IDs to structural variants: it can register glycans with partial information, including undefined linkage positions and absent linkage information[29], and still provide a unique identifier. The GlyTouCan repository has become a foundational resource for the standardization of glycan identifiers and serves as an infrastructure for storing glycan structures.

Another high-level integration effort is GlyGen[23], a comprehensive, integrative knowledge platform for glycoscience research. Funded as an NIH Common Fund program, GlyGen aims to harmonize glycan-related data from diverse experimental and bioinformatic resources into a unified, standardized resource and provide a unified web portal (https://www.glygen.org). GlyGen cross-references and aggregates information from a wide array of domain-specific and cross-disciplinary external databases. Key resources include chemical structure databases (e.g. UniCarbKB[52], ChEBI[53]), protein

resources (e.g. UniProtKB[54], PDB[55]), pathway databases (e.g. CAZy[56], KEGG[57]), and biomedical resources (e.g. OMIM[58], GlyCosmos[59]). It then links these data via common identifiers including GlyTouCan ID to enable the integration of dispersed glycan-related information, providing a unified and comprehensive overview of glycan structures, biosynthetic pathways, protein interactions, and disease associations. By serving as a centralized knowledgebase akin to a "glycan-equivalent" of UniProtKB, GlyGen bridges the gap between fragmented datasets, allowing researchers to query and explore glycans in a holistic platform. The GlyGen portal[23] (first released in 2018) provides tools for searching and browsing their integrated data. In addition, the GlyGen portal provides tools for searching and browsing integrated data (https://www.glygen.org/quick-search/). A user can query a glycan and find which proteins and organisms it is found in or conversely query a gene or protein and see what glycans are attached. In this process, reconciling different glycan notations (IUPAC form, SNFG images, and GlycoCT/WURCS strings) is required to achieve this integration. Efforts like GlyGen to host an interoperable framework and facilitate information sharing are essential as they serve as a translator between different glycan notations.

Glycans are frequently represented using diverse textual nomenclatures across multiple databases, as well as through textual motifs and structural descriptors in research articles and experimental datasets. Despite progress in standardization, the diversity of natural language terms for glycans poses another challenge, especially for text mining and data retrieval. Glycans are often referred to by different names or descriptions in the literature. Taking the Thomsen-Friedenreich antigen again as an example, the O-linked disaccharide Galβ1-3GalNAcα forms the core structure of the antigen; in different papers

this glycan may be called “TF-antigen”, “T antigen”, “Gal-GalNAc” (although it misses linkage specification), or “T cryptantigen”. While “TF-antigen” and “T antigen” are abbreviated terms that originated from its discovery in early serological studies, structurally, it is sometimes simplified to “Gal-GalNAc” [60], emphasizing its monosaccharide composition. In functional contexts, the term “T cryptantigen” arose from its cryptic expression on healthy red cells, which becomes exposed during pathological conditions such as sepsis – a process known as “TCA activation”[61,62]. Similarly, a complex N-glycan might be described in texts as “triantennary N-glycan with core fucosylation” without attributing it to a formal structure.

This variability in glycan names creates ambiguity for text mining algorithms. A natural language processing (NLP) system or a named entity recognition (NER) tool must recognize that phrases as different as “Lewis X antigen”, “CD15”, and “SSEA-1” all are glycan entities (and indeed they refer to the same glycan motif, Galβ1-4(Fucα1-3)GlcNAc), or that the composition term “G2S2” denotes a biantennary N-glycan that contains terminal sialic acid residues[63]. Mapping these terms to a structured representation like a GlyTouCan ID or GlycoCT descriptor is non-trivial, especially if the text mining system’s dictionary is incomplete, where it will miss mentions of glycan entities. New terms and new aliases of known terms will likely emerge, where authors might coin a new name for an unusual glycan in a specific study, or use an alternative notation. Another challenge is that some glycan terms overlap with general biochemical terms or other entities, confounding standard NER systems. The term “core fucose” exemplifies the semantic challenges, as it specifically refers to a fucosylated GlcNAc residue within N-glycans. However, the words “core” and fucose” may be interpreted as

independent entities when decontextualized, leading to potential loss of information. Likewise, the term “T antigen” may be misinterpreted as “T cell antigen”, illustrating the problems of semantic disambiguation in text mining. For this reason, much of the knowledge linking glycans to functions or diseases is still manually curated by experts reading papers.

These ambiguities necessitate the development of disambiguation rules and the incorporation of ontologies to enhance the accuracy of glycan term extraction. The development and application of advanced text mining algorithms and ontology mapping are essential for further harmonizing and unifying glycan-related information. These efforts are critical for transforming unstructured, literature-derived knowledge into broader, more integrative insights in glycobiology.

2.4. Biomedical Text Mining Methodologies

Biomedical text mining refers to the automated extraction of information from unstructured biomedical texts, such as journal articles or clinical notes. A complete workflow generally goes through four major stages in a pipeline: text gathering, text pre-processing, feature/semantic informatic extraction, and annotated text analysis[64]. The process starts with data acquisition, where a relevant subset of textual data is gathered from data sources such as PubMed (for abstracts and full-text papers) or clinical databases (for electronic health records, EHRs). Then, the raw text is pre-processed for analysis. Pre-processing may include sentence segmentation (splitting paragraphs into sentences) and tokenization (breaking each sentence into individual words or other smaller units called “tokens”). Each token can then be annotated with linguistic features

through part-of-speech (POS) tagging and syntactic parsing to derive grammatical structure[64]. Additionally, cleanup and normalization turn words into lowercase, convert words to their base form (lemmatization), and map terms to standardized vocabularies. Once the text is linguistically enriched, information extraction tasks are performed to retrieve specific biomedical knowledge; such extraction task often involves NER for identifying mentions of key biological and medical entities (e.g. genes, diseases, chemicals) and classifying them into predefined categories[65]. Recognized entities are often mapped to entries or entity identifiers in authoritative knowledgebases or ontologies – for example, linking a gene name to its HGNC[66] gene symbol, or a disease term to a Disease Ontology[67] Identifier (DOID)[68]. The final extraction step is relation extraction (RE), which seeks to detect meaningful relationships between the recognized entities in the annotated text, such as a gene-to-disease association reported in a sentence.

The results of NER, normalization, and RE can then be integrated into structured knowledgebases. Integrated data can then yield machine-readable facts such as triplets of entities and their relations that allow querying and downstream analysis. A text mining pipeline might populate a database of protein-disease interactions by processing thousands of abstracts. Many toolkits and platforms support similar workflows: general NLP software like spaCy and Stanza provide functionalities for tokenization, POS tagging, and biomedical text parsing. Biomedical NLP platforms like cTAKES (for EHRs) and PubTator (for PubMed articles) provide pre-trained pipelines that identify domain-specific entities and link them to databases[69], allowing curators to annotate literature with high precision.

Through text mining pipelines, unstructured text (i.e. a corpus) from sources like PubMed or EHRs is transformed into structured data that enables downstream analyses and knowledgebase integration.

2.5. Named Entity Recognition (NER) in Biomedical NLP

The application of NLP in biomedical NER has significantly evolved over the past few decades. Early approaches to biomedical NER were rule-based or dictionary-based. Biocurators would have to construct extensive lexicons of domain-specific terms and design pattern-matching rules to capture variants of those terms in the text. Such an approach was effective for precise matches, but it was ungeneralizable and required considerable manual labor for curating rules. Starting in the 2000s, the field shifted toward statistical machine learning methods[65]. Supervised learning models, notably hidden Markov models (HMMs), support vector machines (SVMs), and conditional random fields (CRFs), were used to train on labeled corpora such as the GENIA[70] and the BioCreative challenge[71] datasets. These supervised models were trained to identify entities based on contextual features (such as neighboring words or POS tags) and achieved better precision and coverage than rule-based systems. For example, the CRF-based models along with other machine learning tools on JNLPBA[70] (an annotated biomedical dataset) showed decent performance in recognizing gene names through training on thousands of labeled examples.

In the last decade, deep learning approaches have revolutionized NER in biomedicine. Neural network models automatically derive features from data, reducing the reliance on manual feature crafting. Early neural network-based NER models for

biomedical text mining used recurrent neural networks (RNNs), particularly bi-directional Long Short-term Memory (BILSTM) networks[72] often coupled with a CRF output layer[73]. These BiLSTM-CRF models substantially improved biomedical NER performance by capturing long-range context in sequences[38]. Building on BiLSTM, the transformer-based language models emerged as a well-recognized leap forward.

Bidirectional Encoder Representations Transformers (BERT)[74] developed in 2017 was a highly impactful general-purpose language representation model that could be fine-tuned for NER with minimal task-specific architecture. Applying BERT to biomedical text, however, was challenging because BERT's original training data (i.e. Wikipedia, BookCorpus) didn't include biomedical terminology, leading to vocabulary mismatches and poor handling of domain-specific jargon. Domain-adapted BERT models like BioBERT[38] were thereby developed to address such issues. It took the BERT base model and continued pre-training it on biomedical corpora (PubMed abstracts and PMC abstracts), yielding significant gains. BioBERT significantly outperformed the original BERT and previous models on biomedical NER, and alongside several pretrained-transformer models, drove the development of biomedical NER from rule-based and dictionary-based approaches to data-driven machine learning. This included the most recent deep learning with pre-trained transformer models that leverage vast unlabeled corpora. This evolution has greatly improved NER accuracy and scalability, enabling the extraction of entities with much less manual effort.

Recently, the use of large language models (LLMs), introduced in Chapter 1.5, represents a significant advancement in biomedical natural language processing. Trained on extensive textual corpora including biomedical texts, these models have demonstrated

strong performance in tasks such as NER, frequently achieving state-of-the-art results on benchmark datasets. Their development underscores the feasibility of leveraging LLMs for the automated, structured extraction of biomedical entities and relationships, with promising applications in the field of glycobiology.

Chapter 3: Methods

3.1. Literature Retrieval from PubMed

To systematically retrieve relevant glycan biomarker-related studies, a targeted search strategy on PubMed (https://pubmed.ncbi.nlm.nih.gov) was implemented. Advanced search syntax and field-specific query tags were utilized by following the PubMed User Guide (Version: February 25, 2025)[75]. The search aimed to capture articles describing glycan structures, glycosylation processes, and their roles as biomarkers across diseases and conditions. Review articles were excluded to prioritize primary research. The query was executed via PubMed's search interface to return a list of citations sorted by best match (as of November 8, 2024). Using PubMed's native download functionality, all citation results (with a 10,000 citations limit) were saved as a PMID (unique PubMed identifier)-format .txt file. The txt file contained a list of newline-separated PMIDs corresponding to the matched citations.

To efficiently and reproducibly retrieve structured metadata and textual content of the identified articles, the author implemented a Python-based data retrieval module leveraging Entrez Programming Utilities (E-Utilities), specifically the "EFetch" API endpoint[76]. E-utilities are a suite of eight computational tools supplied by the NCBI to enable data retrieval automation mining from Entrez databases, including PubMed, PMC, and Gene[77,78]. They operate via URL-based API calls, supporting batch searching of records in structured formats. The API call approach was used to gain programmatic access to PubMed records in XML format, enabling both batch processing of mass records and targeted extraction of key fields for downstream curation steps in the pipeline.

The PMID list generated from the initial search served as the input of the Entrez API workflow. PMIDs were processed in sequential batches of 1,000 to comply with the API's rate limits. XML responses were parsed and extracted the relevant fields: "PMID", "ArticleTitle", "AbstractText", "KeywordList", "ChemicalList", and "MeshHeadingList" from the "MedlineCitation" and "Article" objects. The "KeywordList" field contained author-provided keywords, and the "ChemicalList" and "MeshHeadingList" were objects containing entity and identifier properties encompassing protein names, enzyme names, disease names, etc. Multiple values under each field were provided as pipe-separated objects. Each article's extracted data was saved as an individual .txt file, structured as:

PMID: [ID]

Title: [ArticleTitle]

Abstract: [AbstractText]

Keywords: [Keyword1|Keyword2|…]

Chemical names: [Descriptor1|Descriptor2|…]

MeSH names: [DescriptorA|DescriptorB|…]

The per-citation data structure facilitated parallel processing in subsequent entity curation steps. The fetched literature records in the txt format were fed into a custom NLP pipeline for glycan entity recognition and mapping. Concurrently, the data retrieval module simultaneously parsed the records into a single .tsv table for unsupervised text clustering analysis.

3.2. Unsupervised Abstract Clustering

The PubMed corpus was initially preprocessed by programmatically removing unusual line terminators from the retrieved .tsv file. Columns related to keywords, MeSH unique identifiers, and chemical registry numbers were excluded. Pipe-separated objects were replaced with spaces. The “Title” and “Abstract” fields were merged into a single text string per record to form the input for textual clustering.

Text data were vectorized using Term Frequency-Inverse Document Frequency (TF-IDF). Unsupervised clustering of the vectorized corpus was performed using K-means clustering to group similar records. The optimal number of clusters (k) was determined by calculating the Within-Cluster Sum of Squared Errors (WSS, or Within-Cluster SSE) and silhouette scores across a range of k values. Principal Component Analysis (PCA) was employed to reduce the dimensionality of the TF-IDF vectors, facilitating visualization of the clustering outcomes. The top 10 terms with the highest TF-IDF weights for each identified cluster were extracted for further analysis.

3.3. Named Entity Recognition and Relation Extraction

The automated entity extraction component of the text mining pipeline was implemented via a Python script for processing fetched literature records. In the first stage, NER (of glycan entities, glycoproteins, glycoenzymes, associated disease terms, and sample details) and glycan-disease RE were simultaneously achieved by leveraging LLMs. Then, metadata including keywords, chemical names, and MeSH names were utilized to validate and supplement entities recognized in the first stage.

To ensure consistency in output the extracted data was formatted following a JSON schema that rigorously defined required fields. The schema guaranteed each extraction returns a JSON glycan biomarker record containing standardized information without hallucinated or missing fields; any unmapped or missing values were returned as empty strings. Query instructions guided the extraction process by first specifying the rationale and rules for placing recognized entities in their corresponding fields. Additionally, it explained how to split and map multiple entities, including the use of pipe-separated values when necessary. The query also specified the context required to distinguish between similar entities – for example, ensuring that only protein names are mapped to the glycoprotein entity, while glycoenzyme names are handled separately.

Integration of the OpenAI API formed the backbone of the pipeline. The script employed the Chat Completions endpoint to identify previously unreported entities and to resolve ambiguities in multiple entity mappings. This allows the system to leverage the understanding for contextual disambiguation. Parameters of the OpenAI API calls (i.e. model, temperature, maximum tokens) are configurable. In this pipeline, the following API call parameters were used:

Model: "gpt-4o-2024-08-06", "gpt-4o-mini-2024-07-18"

Temperature: 0

System content: "You are a biocurator expert in named entity recognition for glycobiology and biomedical literature. "

Response format type: "json_schema"

Strict: "True"

The pipeline performed NER and RE tasks on PubMed records in parallel batches. Parallelization was implemented using Python's "concurrent.futures" module for asynchronous API calls on independent records. Error handling features were designed to manage API rate limit errors, connection timeouts, and JSON decode errors. Those ensure the records parsing module to reliably and efficiently output entities in a standardized JSON file, strictly conforming to the provided JSON schema, for each PubMed record. Structured output JSON files were then integrated into the downstream JSON parser, rendering the JSON files into a tab-separated biomarker table.

3.4. Validation of Named Entity Recognition Accuracy

A total of 500 PubMed abstracts were randomly drawn from the master dataset (4,595 abstracts). Quality control was done manually by a curator. The curator reads the abstract text and the corresponding NER JSON output. For each extracted entity, one at a time, the curator records the performance of the NER pipeline with a single-key tag:

"tp" = true positive. Entity correctly identified and correctly typed.

"fp" = false positive. Identified entity does not correspond to the target entity type.

"tn" = true negative. Abstract contains no entity of the target field.

"fn" = false negative. Entity present in the sentence but missed by the system.

"pr" = partially right. Boundary or subtype error that still conveys some correct information.

The validation decisions are recorded for metric calculation.

3.5. Fine-Tuning of Large Language Model

A supervised fine-tuning procedure was applied to the base model "gpt-4o-mini-2024-07-18" using OpenAI's fine-tuning API. All chat completion records generated through model "gpt-4o-2024-08-06" in the NER-RE pipeline were supplied as training examples. The curated chat completions inputs and outputs were formatted into the required .jsonl format, where each line represented a single interaction in a chat-like structure. The "user" and "assistant" rules under each .jsonl message were then used to mimic prompt-response pairs, guiding the supervised training.

The fine-tuning job used a batch size of 9, a learning rate multiplier of 1.8, a random seed of 42, and a total of 3 epochs. No validation file was provided during the training. Model checkpoints were automatically generated during training for version control. The final model was saved with the suffix "bioNER".

3.6. Standardization of Glycan Terminology

Glycan entity terms were extracted from the glycan biomarker table generated through the NER pipeline. All unique contents under the "glycan_entity" field containing relevant glycan terms (such as names, motifs, substructures, and epitopes) were compiled into a consolidated list with counts to each instance in the glycan biomarker table. The compiled list of unique glycan terms was processed using OpenRefine to address issues of synonymy and notational inconsistency while grouping semantically identical terms. A key collision approach consisting of a combination of fingerprinting and n-gram fingerprinting (n-gram size = 2, 3) functions was utilized to cluster terms with minor variations in nomenclatures. Following key collision, the nearest neighbor method was

implemented. The distance between term clusters was calculated using the PPM (Prediction by Partial Matching) function using a radius of 1.0 and a block size of 6. Lexically dissimilar but chemically identical terms (such as N-glycans and N-oligosaccharides) were standardized manually.

During the process of key collision and nearest neighbor clustering, each term in every cluster was manually verified before grouping. To prevent any false positives, any terms not describing the exact glycan structure were removed from their cluster to ensure the integrity of the clusters. The key collision approach was reiterated by re-clustering the clustered terms until there were no clusters containing chemically identical glycan terms.

3.7. Generation of Embeddings for Glycan Terms and Abstract Sentences

The standardized glycan terms were processed to enable quantitative analysis and visualization. Preprocessing involved replacing Unicode characters (e.g. α, β) with their spelled form equivalents (i.e. alpha, beta) and filtering uninformative terms through a custom stop word list. Space characters in selected glycan motif terms were replaced to avoid information loss on tokenization – for example: “Lewis a” was processed to “Lewis(a)” to avoid being split into “Lewis” and “a”. Preprocessed terms were transformed into numerical features using the TF-IDF vectorization module while preserving hyphenated constructs. The dimensionality reduction algorithm was configured with a neighborhood size of 30. A controlled number of 800 tokens were generated to ensure the biological relevance of term features.

Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction. Cluster identification was performed using a Gaussian Mixture

Model (GMM) to accommodate for the anticipated overlap in the feature clusters. The identified clusters were then visualized in a two-dimensional projection. The most distinctive features associated with each cluster were identified, and the corresponding glycan terms before tokenization were retrieved to analyze the biological context of each cluster.

3.8. Embedding-Based Clustering and Visualization

The non-redundant list of glycan entity values from the glycan biomarker table was imported. Entries ≤ 40 characters were removed to avoid overly complicated compound sentences. Each string was converted to lowercase, Greek symbols were replaced with their Latin names, and all punctuation except hyphens and commas was removed with a regular expression. A domain-specific stop-word list (including “glycoprotein”, “structure”, “protein”, etc.) suppressed tokens that convey little structural information. No generic English stop-word list was applied.

Glycan terms were vectorized with TfidfVectorizer. The parameters for vectorization were max_features = 800, min_df = 2, max_df = 1.0, and with a custom stop-word set. No stemming was applied.

UMAP reduced the TF-IDF vectors to two dimensions for visual inspection. Parameters were n_neighbors = 30, min_dist = 0.3, metric = 'euclidean', and random_state = 42. The 2-D embeddings were clustered with a Gaussian Mixture Model (GaussianMixture, n_components = 2, random_state = 42). For each cluster, the mean TF-IDF vector was compared with the global mean to obtain a contrast score. Tokens were ranked by this contrast and the top ten terms were retained.

3.9. Computational Environment and Implementation

All development and testing of the NER-RE pipeline were conducted on the GlyGen developer server running CentOS Linux 7 (Core). The Python environment was based on Python 3.9.6. Core packages essential for the functioning of the pipeline and their versions are listed below:

numpy (1.26.4), pandas (1.5.3), scikit-learn (1.5.1), nltk (3.9.1), regex (2024.11.6), joblib (1.4.2), biopython (1.84), beautifulsoup4 (4.12.3), openai (1.61.1).

Chapter 4: Results

4.1. Search Query for Literature Retrieval

A rigorous search query was designed to obtain a comprehensive yet focused glycan biomarker dataset. The custom PubMed search string combined glycomics-related MeSH terms, title/abstract keywords, and biomarker-focused MeSH terms using Boolean logic:

```
((("glycomics"[MeSH Terms] OR "glycosylation"[MeSH Terms] OR
"glycan"[Title/Abstract] OR "glycans"[Title/Abstract])
AND "biomarkers"[MeSH Terms])
NOT (Review[Publication Type])
```

The MeSH terms “glycomics” and “glycosylation” were selected to search for glycobiology studies indexed under controlled vocabulary, while keywords “glycan” and “glycans” ensured the inclusion of non-index articles containing these terms in free-text fields. “biomarkers[MeSH Terms]” connected with a “AND” logic restricted results to studies explicitly linking glycans to studies that reported their applications as biomarkers. “NOT Review[Publication Type]” filtered out flagged review articles to focus on experimental research, although a small number of review articles were able to get past this filter.

4.2. Characteristics of Retrieved Abstracts

The systematic retrieval of glycan biomarker-related literature from PubMed yielded a corpus of 4,595 citation records, with publications ranging from 1981 to 2024 and a median publication year of 2011. This dataset represents a compilation of glycobiology studies linking to diseases, enabling comprehensive analysis of glycan-disease entity relationships

Custom PubMed query with manual validation of a subset (n = 500) revealing that 97.6% of retrieved articles addressed glycan structures or glycosylation processes. The exclusion of review articles reduced redundancy, with only 0.40% of initially retrieved records flagged as non-primary research during post-processing.

Temporal analysis revealed an average of 50% increase in the number of glycan biomarker publications from the 2000s to the 2010s and from the 2010s to the 2020s, plateauing in the 2020s (Figure 3). Articles were distributed across 904 different journals, with *J Biol Chem* (6.3%), *Glycobiology* (3.9%), *J Proteome Res* (2.5%) ranking at the top. Geographic distribution highlighted contributions from the United States (50.9%), England (24.2%), and the Netherlands (8.2%).

Data retrieval workflow implementing PubMed E-Utilities successfully fetched all 4,595 records including abstract text and metadata. 97 records contained an empty abstract and were removed. Two PMIDs (10467338, 9645584) sharing the same article title were removed from the dataset regardless of having an empty abstract field. 97.9% of records contained full abstracts.98.5% were labeled as journal articles and 7.1% were labeled as comparative studies. The top substance names were “Polysaccharides” (26.2%), “Biomarkers” (22.3%), and “Biomarkers, Tumor” (16.1%). The top MeSH

terms were “Humans” (80.6%), “Glycosylation” (55.2%), and “Animals” (37.0%). The top keywords were “glycosylation” (2.4%), “Glycosylation” (2.0%), and “biomarker” (1.6%).

4.3. Results of Abstract Clustering

Textual analysis of the PubMed corpus using Term Frequency-Inverse Document Frequency (TF-IDF) revealed clusters through unsupervised K-means clustering TF-IDF was selected based on its ability to effectively reduce noises from common and non-informative terms – giving higher weights to terms appearing frequently in a document while penalizing terms that appear frequency across documents (such as stop words like “the” and “and”). Within-Cluster Sum of Squared Errors (WSS, or Within-Cluster SSE) and silhouette scores were utilized to determine the optimal number of clusters (k) obtained from K-means clustering.

The Within-Cluster SSE and silhouette scores corresponding to different k values were analyzed (Figure 4). The elbow plot, depicting Within-Cluster SSE, showed the sum of squared distances from each point to its cluster centroid. The “elbow” point of an elbow plot was identified as a sharp decrease in the rate of Within-Cluster SEE reduction, suggesting an optimal number of clusters. To analyze the quality of clustering, the silhouette score plot was utilized. The silhouette score plot measures how similar an object is to its cluster compared to other clusters, with higher scores indicating better-defined clusters. A noticeable elbow was identified at k = 3, with a relatively high silhouette score. In addition, another non-profound elbow at k = 6 was identified with a local silhouette score maximum. Both clusterings are visualized in Figure 5.

When partitioned into three clusters (Table 2), cluster 0 (n = 3,173 abstracts) predominantly contained terms associated with tumor antigens, antibodies, and receptor activities, reflecting a general theme of tumor immunology and glycoprotein interactions. Cluster 1 (n = 1,220 abstracts) contained terms associated with disease such as hepatocellular carcinoma (HCC: “hcc”), with mentions of common cancer biomarkers including prostate specific antigens (PSA: “psa”) and immunoglobulin G (IgG: “igg”). Cluster 2 (n = 202 abstracts) contained termed associated with immunological responses against HIV, showing common tokens of neutralizing antibodies, CD4 (“cd4”, corresponding to CD4 cells subject to HIV viral infection), and viral envelope (“env”) proteins.

A more granular classification of the PubMed abstracts was given by partitioning the corpus into six clusters (Table 3). Cluster 0 (n = 1,789 abstracts) included general terms associated with tumor biology and immunological experiments, highlighting the Tn antigen (“tn”, highly expressed in cancers). Cluster 1 (n = 1,394 abstracts) included serum biomarker terms featuring N-glycans, likely linked to glycomics experiments. Cluster 2 (n = 704 abstracts) featured glycan structural terms (“acid”, “lewis”) related to glycoproteins (“receptor”, “domain”). Cluster 3 (n = 269 abstracts) contained exclusively terms of HIV virology. Cluster 4 (n = 245 abstracts) contained specialized immunoglobulin terms, while cluster 5 (n = 194 abstracts) focused specifically on mucin (MUC1), a heavily O-glycosylated protein known to be aberrantly glycosylated in epithelial tumor cells.

The textual analysis revealed the thematic context of the PubMed corpus, with k = 3 providing broad categorization and k = 6 providing additional resolution into

specialized themes covering both N- and O-linked glycan biomarkers across diseases including cancer and HIV infection.

4.4. JSON Schema and Prompt Design

The extraction task required a single, machine-readable output that captures every glycan-disease relationship mentioned in an abstract and aligns supporting information such as specimen type or causal enzyme. To achieve this goal, a dedicated JSON schema was created (Supplementary Information 2). Keywords, MeSH subject headings, and chemical names provided by PubMed are mapped back into the schema so that controlled identifiers accompany the free-text entities detected in the body of the abstract.

The schema is organized around a top-level array called glycan_biomarkers. Each entry in the array represents one glycan structure that changes in one direction in association with one disease. Mandatory fields include “glycan_entity”, “glycan_entity_change” and “disease_name”. Examples were provided for edge cases, such as excluding analytical enzymes (e.g. PNGase F) and chromatography peaks (e.g. GP1, GP2) from the glycoenzyme field. A second group of fields records context: specimen_type, organism, glycoprotein_entity_name, and “medical_intervention”. Where more than one value can occur, the schema allows a pipe-separated list, for example, “breast cancer cells|breast cancer tissues” in specimen_type. A final group accommodates ontology mapping. Values such as “disease_name_mapped” and “specimen_type_mapped” store MeSH or other identifiers harvested from the PubMed metadata. This structure ensures that every extracted statement is self-contained yet interoperable with external databases.

A carefully designed prompt (Supplementary Information 1) guides the language model through the extraction. The first section states the objective and lists the thirteen entity types that must be recognized. Step-by-step instructions then explain how to populate each JSON field, when to leave a field empty, and how to split multiple values with the pipe character while preserving their original order. The prompt instructs the model to favor precision over sensitivity, ensuring that only entities supported by explicit textual evidence are recorded.

4.5. Example Output of Named Entity Recognition and Relation Extraction

To illustrate the performance of our NER-RE pipeline, Figure 6 shows the output for PMID 30889355 (Title: "Comprehensive N-Glycome Profiling…"). The system correctly identified the glycan entity high-mannose glycans, normalized the change term elevated to our controlled vocabulary label increased level, and mapped the two specimen types (breast cancer cells and breast cancer tissues) under a single "glycan_biomarkers" entry. The pipeline concatenated these specimen types with a pipe character, yielding the unified entity (only three fields are shown for demonstration purposes):

glycan_entity: high-mannose glycans
glycan_entity_change: increased level
specimen_type: breast cancer cells|breast cancer tissues

This demonstrates the model's ability to aggregate evidence across multiple sentences to produce one cohesive glycan-disease record.

Other glycan entities in the abstract (core-fucosylated glycans, bisected glycans, and sialylated glycans) were likewise extracted with correct boundary spans and changed labels ("increased level" for core-fucosylated; "decreased level" for bisected and sialylated). The high precision of relation extraction is evidenced by the absence of spurious or mistyped triplets in this example: every glycan change is paired with the appropriate disease context without false positives.

Disease names and specimen types can be standardized by linking the extracted "breast cancer" label to its MeSH identifier (D001943) and mapping "mammary glands, human" (D008099) to a unified specimen ontology. This mapping ensures that all MeSH and chemical index terms are converted into unique PubMed identifiers, promoting consistency across the glycan biomarker table.

This demonstrates the pipeline's ability to (1) recognize glycan entities and their direction of change across multiple sentences, (2) normalize glycan entity change variants ("elevated" to "increased level"), (3) merge evidence from cells and tissues into a single biomarker entry, and (4) align disease and specimen mentions with controlled vocabularies.

4.6. Summary of Named Entity Recognition and Relation Extraction

The text mining pipeline extracted glycan biomarker information from the corpus of 4,595 PubMed records. 97 abstracts without available abstracts were removed, where 2 of the removed abstracts contained identical article titles. The filtered 4,498 abstracts yielded a structured dataset comprising 6,500 – 6,800 biomarker records depending on the LLM model, user prompt, and JSON schema. 56.6% of the extracted relations were

well-annotated glycan biomarkers, which are defined as “informative glycan structures that show directional changes in observed levels in a cell/tissue sample, associated with a disease or condition”. The NER process was guided by a structured query prompt (Supplementary Information 1), engineered to extract glycan-biomarker relationships while attempting to standardize recognized entities by mapping with standardized ontology names. The query prompt incorporated three core principles to guide the extraction of data into a structured data frame.

First, the query defined 13 biological entities across five categories (“assessed_biomarker_entity”, “associated_disease”, “biomarker_role”, “sample”, and “evidence”), with explicit inclusion/exclusion criteria to prioritize precision. For instance, the query specified that the “glycan_entity” field should require direct structural descriptors (e.g. “core-fucosylation”) and exclude non-glycan terms (e.g. glycoproteins, phosphorylation, and unassigned MS peaks). The optional “glycoprotein_entity_name” and “glycoenzyme_entity_name” provided additional information relating to the localization of the glycan expression and associated biosynthetic pathways.

Next, to enhance interoperability, extracted entities were semantically cross-referenced with provided metadata such as the standardized chemical names and MeSH names. Multiple mapped chemical names and MeSH names in the same field were provided by pipe-delimited strings while maintaining the same order as the reference field strings. This supported subsequent mapping of the biochemical, anatomical, and disease entities to their canonical identifiers supplied by authoritative ontology databases.

Instructions for multiple entities handling mandated the splitting of multi-entity relationships into discrete objects under the “glycan_biomarkers” array to avoid

conflating indirect associations. For example, an abstract describing plasma samples of AMI patients was reported with “decreased level of galactosylation” and “increased level of sialylation” should generate two separate entries, each linked to a “glycan_entity” with different “glycan_entity_change” keys. The requirement to include verbatim sentences with the glycan entity term ensured the traceability of the source text.

On successfully recognizing the entities, each record was generated following the predefined JSON schema (Supplementary Information 2) which mandates 13 fields. Names of the fields and their descriptions are included in the JSON schema. Two model configurations were compared: the full ChatGPT 4o model and its streamlined counterpart – 4o-mini. The 4o model demonstrated robust reasoning capabilities and adhered closely to the detailed prompt instructions, while the 4o-mini model, though faster, occasionally struggled with overly complex instructions.

Figure 7 shows the distribution of top entities identified by the NER-RE pipeline across the six major entity fields: disease, species, glycoprotein, glycoenzyme, biomarker role, and specimen. The overwhelming majority of identified species entities correspond to human studies, which underscores the suitability of these data for biomarker-focused research. Conversely, disease terms and specimen types exhibit substantial heterogeneity, with various entities distributed across many distinct categories. Such diversity highlights the complexity involved in mapping and harmonizing these terms. This shows a necessity for standardized terminologies and sophisticated approaches to integrate and interpret biomarker data effectively. In addition, it is also shown that the glycan entities identified are highly heterogeneous (Table 4), as expected from the number of notations circulating in the glycobiology field.

4.7. Standardized Glycan Terminology Results

Figure 8 presents the 30 most frequent standardized glycan terms extracted from the PubMed corpus. Broad descriptors occupy the top of the list: N-linked glycan (n = 744 occurrences), followed by the generic label Glycan (n = 315) and the well-known epitope Sialyl Lewis X (n = 229). Other high-frequency entries include O-linked glycan (n = 215), Tn antigen (n = 188), and Lewis X (n = 124). The middle tier of the bar chart is occupied by key terminal modifications, namely Fucosylated glycan (n = 116), Sialyl-Tn antigen (n = 112), Sialylated glycan (n = 109), and Sialic acid (n = 105); and the core structures T antigen (n = 104) and High mannose (n = 86). Branching and capping variants such as Core-fucosylated (n = 74), Sialyl Lewis a (n = 72), Alpha-2,6-linked sialic acid (n = 64), GPI anchor (n = 60) and Core 2 (n = 53) form the top 15.

Beyond canonical motifs, the chart reveals a smooth tail of lower-frequency yet relevant terms. These include monosaccharide names (GalNAc, n = 37), linkages (Alpha-2,3-linked sialic acid, n = 37), complex-type labels (Complex-type N-glycan, n = 24), glycosylation reactions (Galactosylation, n = 25), and glycan-protein associations (e.g. HNK-1, n = 26). In total, 27 of the top 30 entries are true glycan descriptors, reflecting the effectiveness of the NER-RE pipeline in capturing diverse glycan motifs.

Three of the top 30 bars are classified as false positives (grey bars): MUC1 (n = 42), Mac-2 binding protein glycoform (n = 33), and Advanced glycation end products (n = 27). Although these entries are not discrete glycan structures, they bear a direct relation to glycosylation. MUC1 is a mucin glycoprotein whose aberrant O-glycosylation is a hallmark of several cancers. The Mac-2 binding protein glycoform refers to a glycoprotein isoform with disease-specific glycan alterations. Advanced glycation end

products arise from non-enzymatic sugar addition to proteins and are studied alongside enzymatic glycan modifications in metabolic disorders. Their appearance in the top ranks highlights that our text-mining workflow also surfaces glycosylation-related biomarkers beyond pure glycan motifs, providing a richer context for downstream analysis.

4.8. Visualization of Glycan Term Clusters

Projecting the 800-dimensional TF-IDF vectors into two dimensions with UMAP revealed two well-separated clusters (Figure 9). The larger orange cluster spans a broad, diffuse region, while the smaller blue cluster forms a tight local knot. Both clusters were generated by Gaussian-mixture modeling (n = 2) and annotated with their top three distinguishing tokens.

A random sampling of 10 terms in each cluster is listed below:

Cluster 1:

'glycoforms of fetuin', 'bisected type N-glycans', 'branched core-2 structures', 'a-sialo N-glycan', 'N-glycan of CCP-2', 'NeuAc alpha 2-3Gal in N- and O-glycans', 'sialylated antigens', 'asialylated oligosaccharides', 'mannose 6-phosphate residues', 'extrinsic sialylation', 'sialic-acid links alpha(2,6)Gal/GalNAc', 'sialylated sulfated glycan', 'GD3 ganglioside', 'N-glycan-linked 6-sulfo sialyl Lewis(X)', 'sialylated core-1', 'cell surface glycosylation', 'core-fucosylated Hsp90', 'oxidized low density lipoprotein', 'bisected complex-type N-glycan', 'high branching'

Cluster 2:

'betaGal-betaGalNAc-containing glycans', 'sulfa-Tn-antigens', 'tetrasialotransferrin', 'SARS-CoV-2 spike glycans', 'GD1a', '9-O-acetyl-N-acetylneuraminic acid', 'synaptophysin glycosylation', 'digalactosylation', 'mono-glucosylation', 'IgG2G2S', 'asialofetuin', 'IgG4G1S', 'K13', 'GT1a-like oligosaccharides', 'Glycodelin A', 'TRA-1-81', 'arabinoxylan oligosaccharides', 'IGP58', 'GAG-NREs', 'AMCA'

The first, more expansive cluster is dominated by terms associated with canonical N-glycan and O-glycan motifs, sialylation, and Lewis-type epitopes. Its top tokens, such as “n-glycan”, “sialylated” and “fucosylated”, reveal a theme of branching, capping, and antigenic decoration. Sample entries from this cluster such as “bisected complex-type N-glycan”, “sialylated core-1, and “high branching” underscore the cluster’s focus on structural variants of N-glycans, mucin-type O-glycans and their sialo-/fucosyl-termini across diverse protein backbones.

By contrast, the second, compact cluster gathers terms that resemble IUPAC fragments, enzyme-mediated modifications, and glycoprotein identifiers. Its top tokens (“acid”, “alpha-galactose” and “agalactosylation”) point to line-notation features and site-specific chemical changes. Representative entries such as “9-O-acetyl-N-acetylneuraminic acid”, “betaGal-betaGalNAc-containing glycans”, “SARS-CoV-2 spike glycans”, “synaptophysin glycosylation” and “IgG2G2S” showed the heterogeneity of these terms, explaining the cluster’s tight grouping: they share uncommon substrings, hyphenated IUPAC conventions or protein-specific labels not

found in the broader glycan lexicon and thus share little to no tokens with the rest of the other glycan terms.

The two clusters highlight a split between general glycan-type descriptions (cluster 1) and highly specific chemical or protein-linked variants (cluster 2). This separation provided valuable insights, that embedding-based clustering can distinguish broad structural classes from niche, complex nomenclatures. This implies a scalable route to organize thousands of glycan terms in place of manual curation.

4.9. Outcomes of Model Fine-Tuning

The larger foundation model "gpt-4o-2024-08-06" was successfully distilled into a smaller, cost-effective variant using the "gpt-4o-mini-2024-07-18" base model in a supervised fine-tuning job. This approach enabled the reduction of computational work in future NER while preserving the model's capacity to accurately identify glycan biomarkers and associated biological entities.

A total of 4,595 examples, derived from the glycan biomarker NER pipeline using the GPT-4o model, were used to train the distilled model. A total of approximately 20.2 million tokens were processed. Supervised training was successful, achieving a final training loss of 0.0056 and outputting the fine-tuned model "ft:gpt-4o-mini-2024-07-18:personal:bioner:BKscp2Ph". The low training loss indicated the model learned the structure of the training data effectively.

Figure 10 showed the training loss decreased rapidly within the initial steps of fine-tuning and stabilized afterward, indicating consistent learning without divergence.

Chapter 5: Discussions

5.1. Interpretation and Implications of Results

Biological and Clinical Relevance of Extracted Glycan Biomarkers

The annotated table produced by our pipeline captures far more than a simple list of glycans and diseases. Each record links a glycan structure to its reported change of abundance, the tissue or cell type in which the change was observed, the carrier glycoprotein when specified, and any enzyme defect or therapeutic intervention that authors associated with the alteration. This multi-layer annotation turns the table into a versatile query platform. A researcher interested in neoglycosylation events can filter for entries where the glycan_entity_change field records an acquisition of structure. A clinician studying leukemia can retrieve all reports of high-mannose loss or sialylation gain in patient serum by combining the disease and specimen filters. The same flexibility allows exploratory searches for diseases linked to a single glycan motif or for motifs that change in opposite directions across different tissues.

Because the table is machine-readable and mapped to standard identifiers, it can be loaded directly into biomarker knowledge bases and pathway tools. Linking the glycan-disease pairs to genomic, transcriptomic, and metabolomic layers will close a long-standing gap in multi-omics studies where the glycosylation status of proteins is usually ignored. Pathway enrichment analyses that rely only on gene or protein names miss the regulatory influence of terminal fucosylation on immune evasion or of core branching on receptor activation. Including our glycan annotations will reveal hidden routes of cell adhesion, immune modulation, and protein folding that are essential in cancer and metabolic disease. More than half of secreted and membrane proteins carry at

least one glycan chain, so the addition of glycan nodes is expected to increase the biological coverage of network models by a meaningful fraction.

Impact on Biomedical Text Mining and Biomarker Discovery

The work also demonstrates a practical route for large-scale biomedical text mining in the era of reliable transformer models. The user only provides a JSON schema that lists the desired entity types and a short prompt that guides the language model, while the rest of the workflow runs automatically. No manual rule writing or expert curation of seed vocabularies is required, yet the system still recovers domain-correct relations such as high-mannose reduction in breast cancer tissue or core fucosylation gain in hepatocellular carcinoma. Because the extraction logic is schema-guided, the same codebase can be repurposed for lipid mediators, post-translational modifications, or microbiome metabolites simply by substituting the entity list and adjusting the relation labels.

Beyond automated curation, the consolidated table supports data-driven hypothesis generation. When multiple papers use different names for the same glycan class, the table harmonizes them under one identifier and reveals recurring links to the same disease that individual authors may have overlooked. Conversely, distinct glycan structures that appear in separate publications on a shared developmental syndrome can be grouped and inspected for common biosynthetic steps. Such patterns point investigators toward specific glycosyltransferases or lectins that warrant experimental follow-up. By turning thousands of free-text observations into a searchable graph, the

pipeline shortens the path from scattered evidence to testable biomarker panels and therapeutic targets.

5.2. Comparison with Traditional (Non-LLM) NLP Approaches

Advantages of LLMs in Named Entity Recognition

Early biomedical NER systems relied on dictionaries and hand-crafted rules. These approaches achieve acceptable precision on well-studied entity classes (e.g. genes) but their recall drops sharply when the vocabulary is sparse or newly coined[79]. Machine-learning upgrades that appeared in the BioCreative and i2b2 challenges improved recall yet still required thousands of annotated sentences for each domain and struggled with overlapping entities[80,81]. Review papers note that traditional models often misidentify polymorphic terms or abbreviations because they view context in a fixed left-to-right window and cannot reason over the full sentence[82].

LLMs overcome these limits by considering texts preceding and following the current word being processed – hence the property "bidirectional" in Bidirectional Encoder Representations from Transformers (BERT). In this NER-RE pipeline, GPT identifies novel glycan phrases such as "N-glycan-linked 6-sulfo sialyl Lewis X" the first time they appear since the surrounding words signal that the phrase describes a saccharide rather than a cell type. The same contextual awareness prevents false positives; using the same example from Chapter 2.3, a sentence about "T cell antigens" is not being confused with the Thomsen-Friedenreich "T antigen" because the immune-cell context shifts the embedding away from glycan space. Unlike conventional systems, this project pipeline requires no task-specific training data, yet it can even map disease

mentions to Disease Ontology identifiers with reasonable accuracy, a function that earlier rule sets accomplished only through explicit lookup tables[83].

The ontology prediction module was once incorporated in the pilot pipeline but subsequently removed to conserve tokens and to eliminate the possibility of hallucinations. However, the success in accurately assigning Disease Ontology Identifiers (DOIDs) to recognized disease entities showcased the depth of the latent biomedical knowledge pre-stored within the model weights.

The superiority of LLMs in Relation Extraction

Relation extraction has long been the bottleneck in text mining. Classic pipeline systems first call a named entity tagger and then apply pattern rules or shallow classifiers to pairs of entities in the same sentence. BioCreative benchmarks show that these methods rarely exceed an F1 of 55 % on inter-sentence chemical-disease or gene-phenotype relations[84]. Recent deep networks that incorporate document graphs raise the score but still miss roughly half of the cross-sentence links, largely because they cannot integrate distant context effectively[85]. Glycobiology poses an even harder challenge. A typical glycomics study reports several structures that rise or fall across multiple specimens and disease stages. The glycan names and disease labels can be several sentences apart, which breaks sentence-bound rule engines.

Transformer-based LLMs handle this complexity by representing entire passages in one embedding space. In this thesis, the NER-RE pipeline correctly linked glycan terms and diseases even though the pair were sentences apart. One example would be that system was able to map glycan terms “Sialyl Lewis X”, “monoantennary glycans” etc.

with the disease term “lung cancer”, although the glycan and disease terms come from two different sentences (PMID: 21214223. This ability matches the findings of recent research that multi-head attention models capture long-range biomedical relations more effectively than traditional sequence classifiers[86]. Benchmarks on chemical-protein tasks confirm the trend: ensembles that include a large language model rank at the top of the “DrugProt leaderboard”, outperforming carefully engineered feature pipelines[87,88].

Traditional NER may still outperform LLMs on the recognition of ultra-rare biochemical tokens. However, when relation extraction quality is considered, the balance shifts strongly toward transformer architectures. LLMs deliver complete biomarker triples with fewer false links, smaller annotation overhead, and no bespoke rule maintenance. These gains explain why our glycan pipeline, built on GPT without domain training, surpasses earlier pattern-based attempts at glycan literature mining and makes large-scale, continuously updated curation feasible for the first time.

5.3. Technical Challenges and Limitations

Precision vs. Recall: Handling False Negatives in LLM-based NER

Through manual inspection, the NER-RE pipeline demonstrated robustness against typical LLM hallucinations, with no entity identified that did not exist in the provided abstracts and metadata. However, precision errors did occasionally occur due to misinterpretations by the model when extracting glycan entities. For instance, in a scenario where the abstract described an "increased glycosylation of MUC1," the pipeline mistakenly extracted the phrase "glycosylation of MUC1" entirely as the glycan entity, subsequently repeating "MUC1" in the associated glycoprotein field. Ideally, a more

generic glycan term such as "glycans" should populate the glycan entity field, clearly distinguishing it from associated proteins. This type of misplacement reflects the importance of clearly specified curation rules to enhance precision. Nevertheless, excessively detailed rules (e.g. excluding non-primary research findings or enumerating glycan terms in multiple notation forms) may reduce recall, as stricter definitions will inadvertently cause more valid yet variably expressed entities to be omitted.

Token Limitations and Complexity in Prompt Engineering

A notable challenge identified during pipeline development concerns prompt complexity and associated token limits, particularly when using the more compact GPT-4o-mini model. Smaller models inherently possess fewer parameters, limiting their internal capacity for multi-step reasoning tasks. It was observed that, after reaching a certain prompt complexity, such models fail to exhibit further improvements in precision despite additional refinements to the instructions. This limitation arises because smaller models must manage their internal reasoning with higher token efficiency, leading to less internal computation per output step. As a result, complex or heavily structured instructions yield inferior model precision, forming a trade-off between computational efficiency and prompt sophistication.

Another significant token-related challenge is enforcing adherence to structured JSON schemas. It was reasoned that the comprehensive JSON schema employed in this project requires substantial token allocation to ensure outputs strictly conform to predefined structural requirements. Each entity field includes precise definitions and accepted

values, accompanying elaborate instructions in the prompt. While this structured approach guarantees data consistency, it substantially increases the token count of each API call. This constraint underscores an inherent limitation of employing detailed schemas with language models, where precision and structural rigor must be carefully balanced against available token budgets given by the selected model.

5.4. Unique Advantages of GPT Models in Biomedical Text Mining

Continuous Improvements and Releases of Better Language Models

The rapid pace of development in GPT models provides a significant advantage for biomedical text mining. Even as this thesis is being prepared, OpenAI released GPT-4.1[89], an advanced version with substantial improvements in reasoning and context comprehension, capable of handling context lengths up to 1,000,000 tokens. This rapid progression ensures continuous enhancement of language understanding, precision in named entity recognition, and accuracy in relation extraction. Importantly, GPT models uniquely support strict adherence to structured output schemas. As of late 2024, GPT remains the only major large language model series enforcing strict JSON schema adherence through the "strict=true" parameter for structured outputs[90]. While alternative models like Claude, Llama, Gemini, and DeepSeek support structured outputs, they do not guarantee complete conformity to user-provided schemas. This capability of GPT enables consistent, reliable extraction of structured data directly into user-defined formats, streamlining integration with biomedical knowledge bases and analytical tools.

Additionally, GPT models benefit from rapid expansion of integrated capabilities, including storage options (vector stores)[91] and tool integration (function calling)[92]. These

enhancements facilitate even more sophisticated text mining tasks such as extracting knowledge directly from complex article figures (image analysis)[93] or accurately mapping entities to user-supplied ontology datasets (file search)[94], greatly expanding the scope of automated knowledge curation in biomedical contexts.

Speed and Low Requirement for Local Computational Resource

Another distinct advantage of GPT models is their speed and minimal reliance on local computational resources. Using GPT models via the OpenAI API eliminates the need to store large model weights locally, significantly reducing hardware requirements. This cloud-based solution offers enhanced reliability, scalability, and ease of integration into automated pipelines. The OpenAI API supports parallel queries, as demonstrated in this project, enabling processing of large-scale text mining tasks with high throughput. Moreover, the availability of higher-tier API usage grants generous rate limits, minimizing bottlenecks and ensuring continuous processing efficiency, essential for real-time or large-scale biomedical literature curation tasks.

5.5. Prompt Engineering as a Critical Factor

The Complexity of Prompt Design and Token Management

Prompt engineering involves strategically crafting textual instructions provided to LLMs to guide effective responses. These carefully constructed prompts shape the model's behavior, enabling accurate, relevant, and context-sensitive outputs. Prompt engineering is emerging as a distinct research discipline within natural language processing, driven by its importance in harnessing the full potential of advanced language

models. For example, studies have demonstrated how carefully designed prompts can significantly improve the accuracy of question-answering tasks or enhance a model's capacity to extract structured data from unstructured texts. Recent research on prompt engineering introduced the “chain of thought” prompting technique, demonstrating that explicit step-by-step instructions dramatically improve language models’ reasoning abilities[95]. OpenAI documentation emphasizes that prompt engineering and chaining (subdividing tasks across multiple prompts) often provide a more practical and cost-effective solution to problems such as QA, compared to expensive model fine-tuning[96].

Effective prompt engineering also includes careful token management, as prompts and responses are subject to token limits imposed by the service providers. Precise yet concise prompts not only conserve tokens but also optimize model performance.

Case Studies and Practical Insights from Prompt Refinement

In the NER-RE pipeline developed in this project, prompt engineering played a central role. The pipeline’s prompt comprises two core parts: explicit instructions (the "query text," detailed in Chapter 4.4), and the abstract content, which includes the article title, abstract, and metadata (keywords and MeSH terms). Initially, the prompt provided to the LLM was minimalistic, omitting detailed definitions of each JSON schema field. While this simplified approach had some early success, several significant problems emerged, highlighting the necessity of detailed prompt engineering.

One recurring issue involved accurate entity recognition but improper field assignment. For instance, protein names frequently appeared in the glycan entity field when abstracts discussed glycosylation changes on glycoproteins. Similarly,

glycoenzymes responsible for synthesizing glycan structures, though correctly recognized, were erroneously placed in the associated glycoprotein field. Another critical challenge arose in relation extraction due to insufficient instructions regarding the structure and granularity of biomarker entries. Early prompts did not clarify that a glycan biomarker entry should include exactly one glycan structure, one associated change (such as increased or decreased level), and one disease. Furthermore, the pipeline initially struggled with entries describing identical glycan structural changes across multiple specimen types, requiring explicit guidelines that these should be consolidated into a single biomarker record with specimen types presented as pipe-separated values.

Prompt refinement involved enhancing instructions with clear definitions and structured guidance embedded both directly in the prompt and in the JSON schema. The JSON schema definitions themselves form part of the prompt engineering strategy, explicitly specifying acceptable values, such as restricting the glycan entity change field to a predefined vocabulary (e.g., increased level, decreased level, altered ratio, etc.), and clearly defining biomarker roles. Through iterative refinement, providing explicit instructions and structured definitions significantly improved the accuracy of both named entity recognition and relation extraction tasks, underscoring prompt engineering's critical role in maximizing the practical utility and reliability of LLM-driven pipelines.

5.6. Future Directions and Opportunities

Construction of Large-Scale PubMed Glycan Embeddings

The next step for this project would be to move from sentence-level extraction

toward continuous representation of the entire PubMed glycobiology corpus. In recent years, text-embedding models have become a cornerstone of natural language processing. An embedding is a numerical vector that captures the semantic context of a word, phrase, or sentence. When two texts have similar meanings, their vectors lie close together in the high-dimensional space; when their meanings differ, the vectors are far apart.

Generating a large-scale embedding index for PubMed would allow rapid, cost-effective discovery of hidden glycobiology knowledge. Once all abstracts are vectorized with an embedding model (e.g. text-embedding-3-large), any newly curated glycan term can be projected into the same space. A nearest-neighbor search then retrieves synonyms or closely related motifs even when the wordings differ. The same principle applies at the sentence level. A sentence that reports core fucosylation in hepatocellular carcinoma will theoretically reside close to other sentences describing similar changes, even if they mention a different protein or use different terminology. This property lets us surface new articles that discuss the same biomarker mechanism but escape earlier keyword searches. In practice, term embedding searches reduce the need for handcrafted queries and speed up literature screenings by several orders of magnitude.

Use of Domain-Specific Embeddings and Ontology Mapping

General embeddings perform well, yet domain-tuned models promise finer discrimination of biochemical nuance. BioBERT and its successors are trained on millions of PubMed sentences and therefore learn specialized associations, for example, that Man9GlcNAc2 might be related to high-mannose glycan. Adopting such domain-specific models should tighten the clusters we observed in UMAP space and help

separate genuine glycan motifs from chemically unrelated expressions that share surface tokens. Another step toward interoperability is ontology mapping. Each entity in the glycan biomarker table can be linked to a stable identifier in a controlled vocabulary. Disease names can align with the Mondo or Disease Ontology resources, specimen types with Uberon for tissues or Cellosaurus for cell lines, and enzymes with UniProt or CAZy. For glycans, however, a dedicated ontology is still missing. The GlyGen GlycoMotif dataset (https://glycomotif.glyomics.org) can serve as the backbone for a new controlled vocabulary that records the biosynthetic origin of each motif and the enzymes/cellular localization involved in its assembly. Incorporating these identifiers will allow seamless integration of the table into pathway resources and multi-omics dashboards, turning the dataset into a bridge between glycomics and other molecular layers.

Transition Toward Locally Hosted Open-Source BioNLP Models

The present study relies on proprietary GPT models for both NER and RE. They excel at broad semantic understanding but carry cost and rate constraints, and they occasionally miss rare glycobiology terms that were under-represented in their pre-training data. Open-source alternatives such as BioBERT, SciBERT, and PubMedBERT can run on local hardware and incur no usage fees. Although their reasoning ability is more limited, they often show higher recall for specialized terminology because their vocabularies include uncommon biochemical tokens. A hybrid workflow is therefore proposed. A domain model first labels candidate glycan, enzyme and disease spans with high recall. The GPT model then revisits the same sentences, is now aware of the suggested spans, and performs relation extraction with its stronger language

understanding. This two-pass strategy combines the vocabulary coverage of BioBERT with the contextual reasoning of GPT, improving overall accuracy while reducing the number of expensive GPT calls.

Ontology mapping can be integrated through a retrieval-augmented generation (RAG) strategy. Taking disease entity as an example: once the domain model identifies a disease mention, a lightweight retriever queries an indexed subset of the Disease Ontology and supplies the best-matching entry. This information is passed to the language model as additional context, allowing the resulting JSON to contain the curated identifier instead of free text. The same retrieval step can align tissues and cell lines with Uberon, Cellosaurus and, when available, link glycan entities to a glycan ontology based on GlyGen's GlycoMotif dataset. Coupling retrieval and generation in this manner produces a self-contained platform capable of mining the rapidly growing glycobiology literature at scale.

Figures

Figure 1. Common disease-associated glycan structural changes.

Representative motifs associated with disease, categorized by structural context. Symbol representations following SNFG notations were created with GlycoGlyph[97].

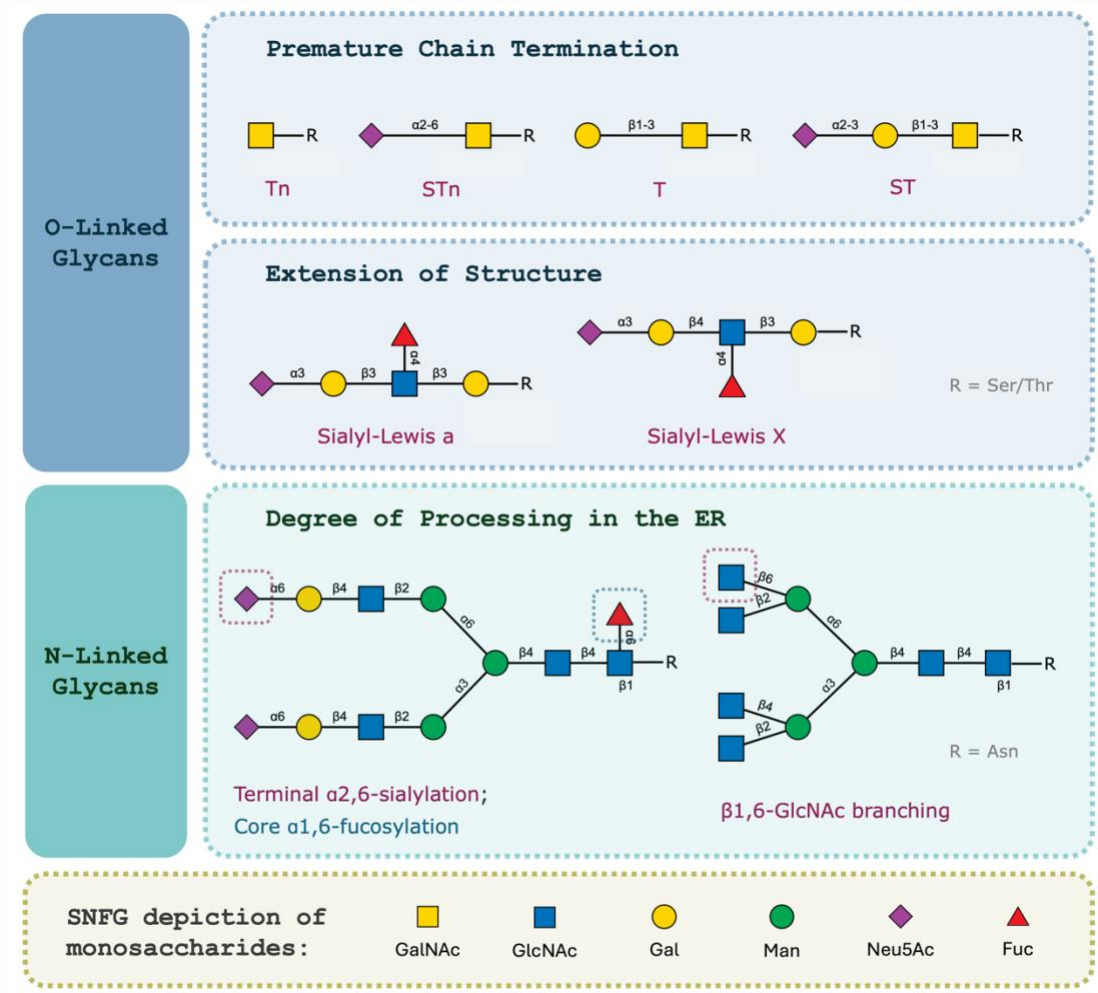


Figure 2. Automated workflow for glycan biomarker extraction.

Schematic showing successive steps from query generation and Entrez retrieval through GPT-based NER, ontology-assisted entity mapping, and conversion of the JSON output into a biomarker table; external prompt-schema files and dictionary resources are indicated alongside the relevant stages.

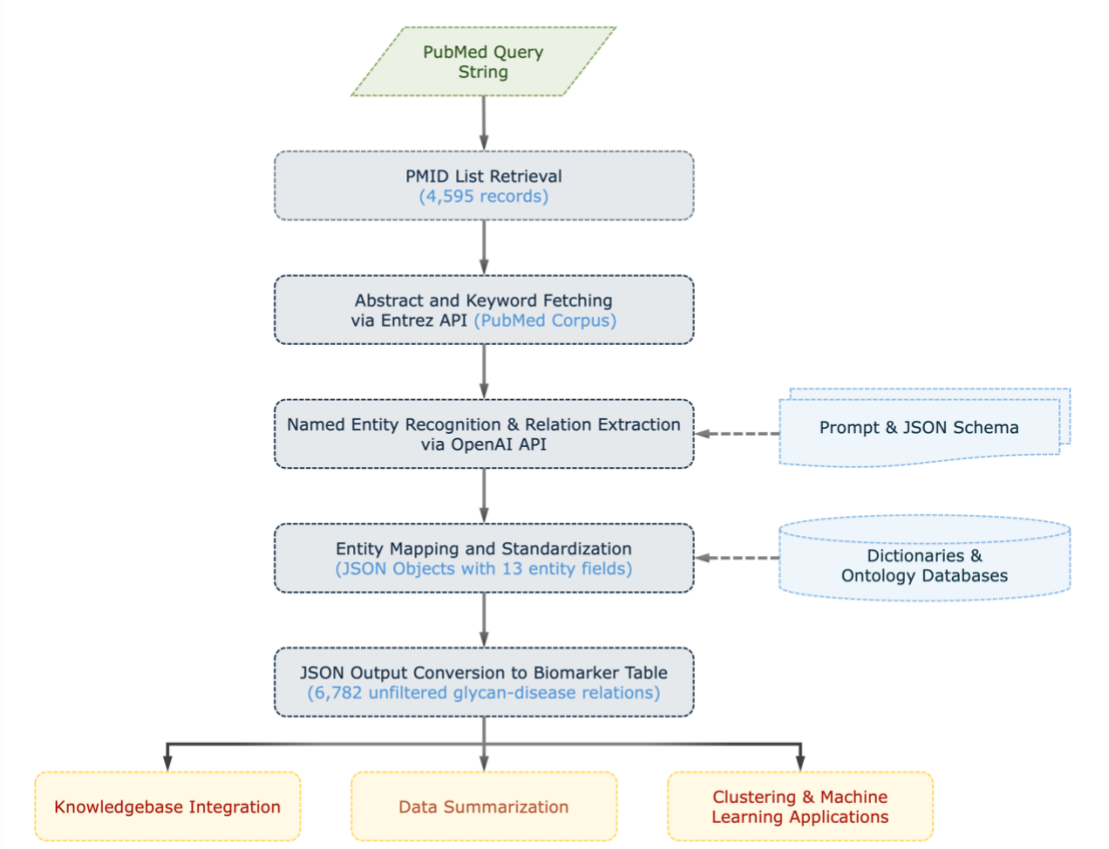


Figure 3. Publication trends for glycan-biomarker query and selected keywords. PubMed yearly publication counts for the custom glycan-biomarker search query and three selected glycobiology keyword sets.

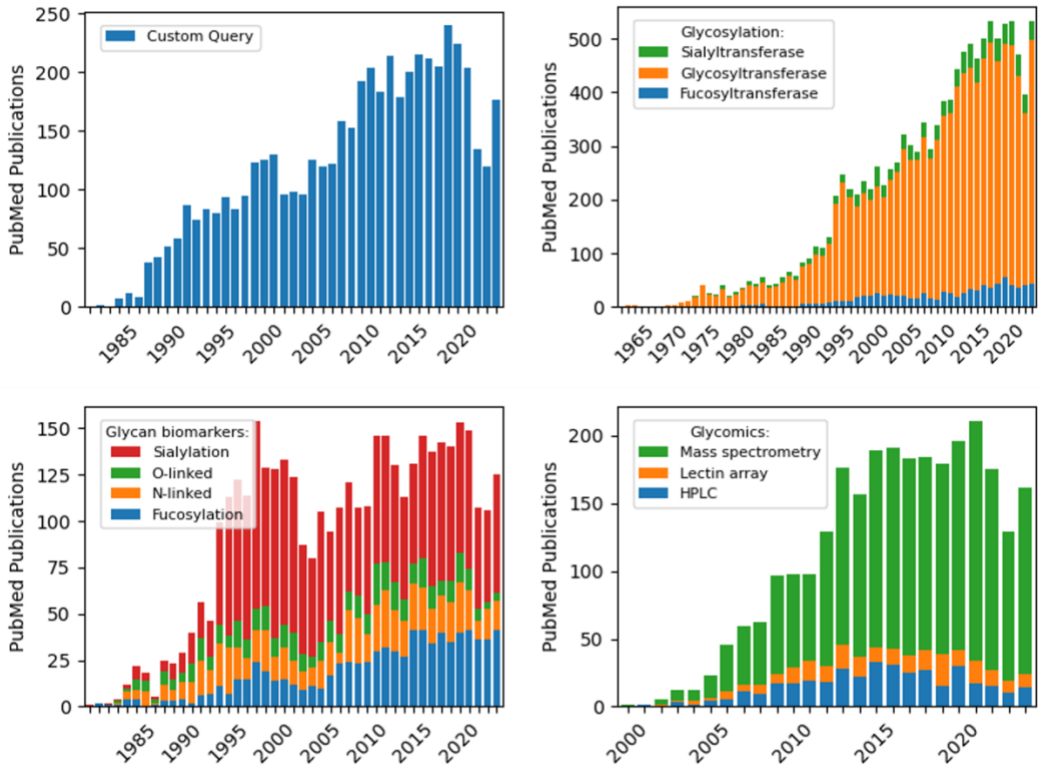


Figure 4. Determining the optimal number of abstract texts.

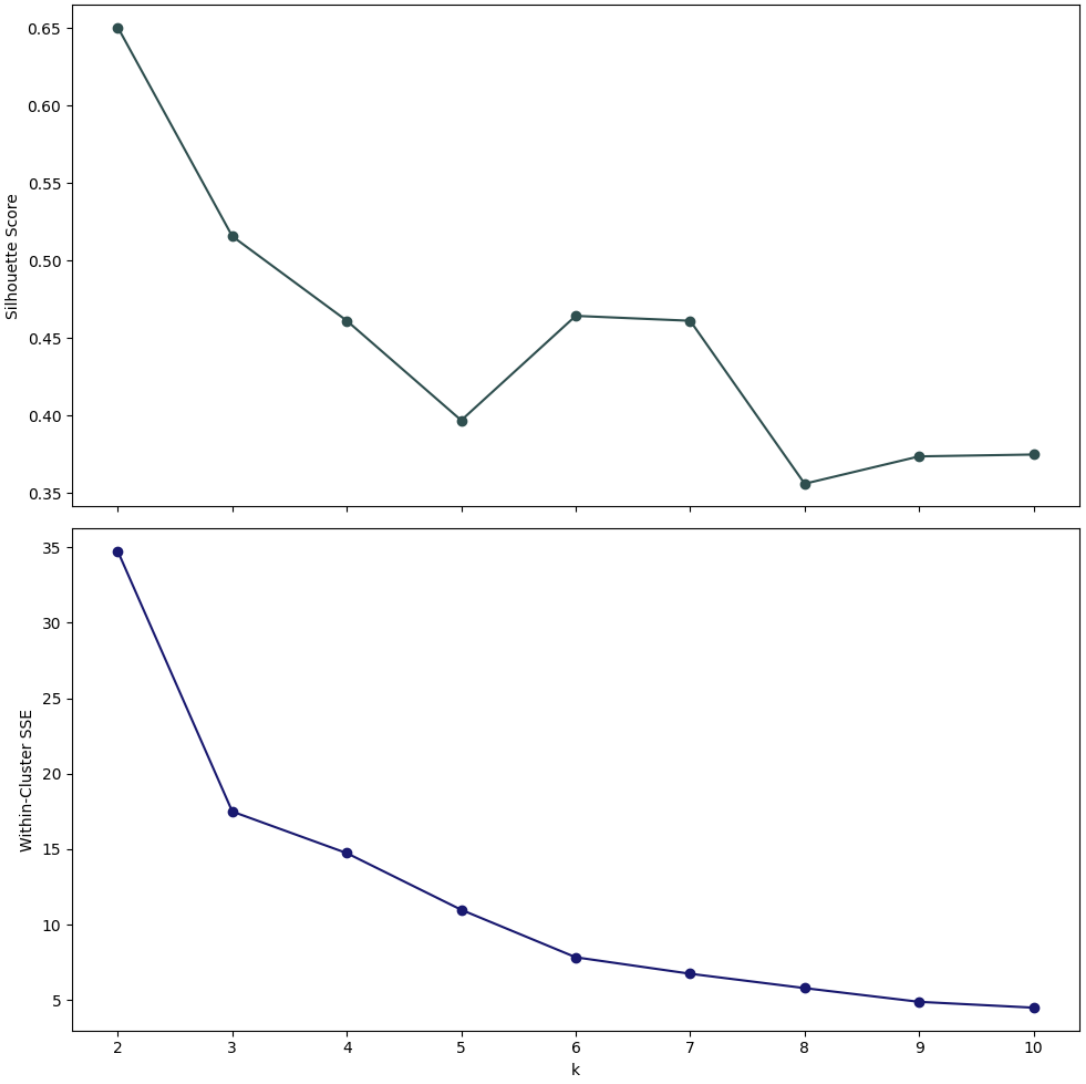


Figure 5. K-means clustering plots (k = 3 and k = 6) of abstract texts.

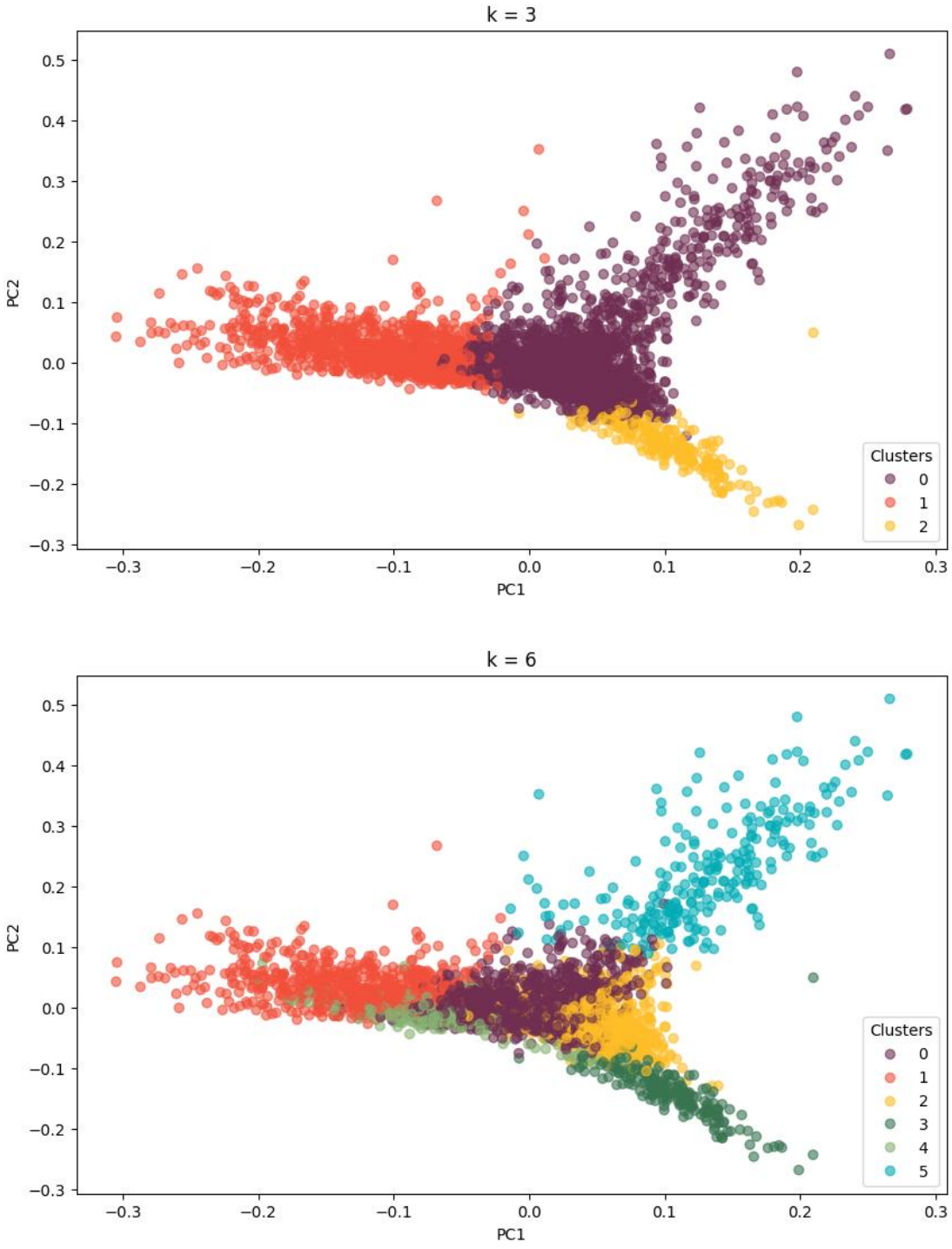


Figure 6. Example glycan-disease relation extraction.

Illustration of the GPT-4o-based NER-RE workflow, highlighted abstract text is aligned with the structured JSON produced by the model. Green and red arrows represent positive and negative correlations between a glycan-disease pair in a specimen, respectively. Bidirectional grey arrows link mapped metadata (keywords, chemical names, and MeSH names) to diseases and specimens.

Figure 7. Summary of the extracted entities.

The top 7 entities for each extracted field are labeled. All other values are grouped and displayed as “others”.

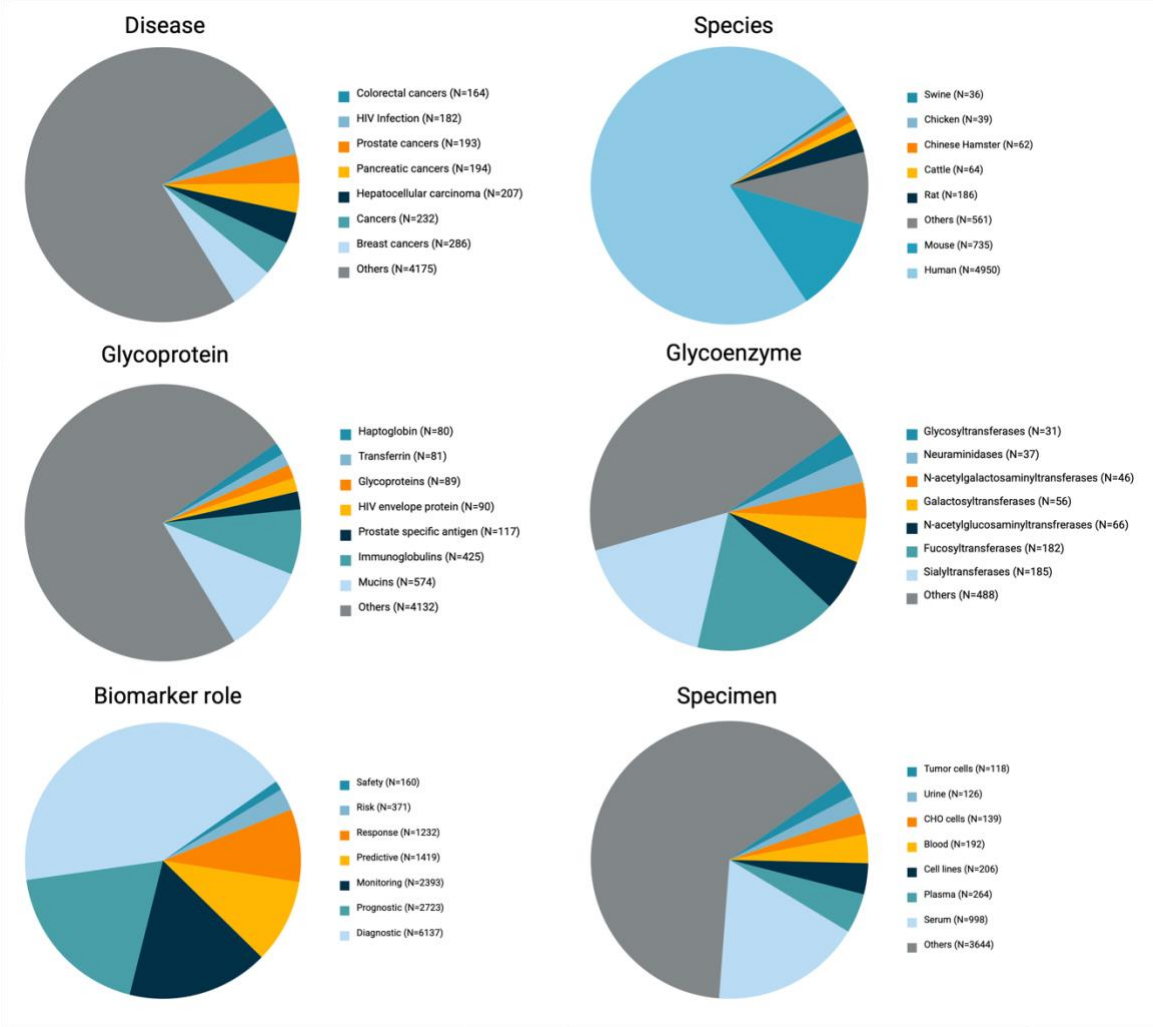


Figure 8. Top standardized glycan entities in the curated corpus.

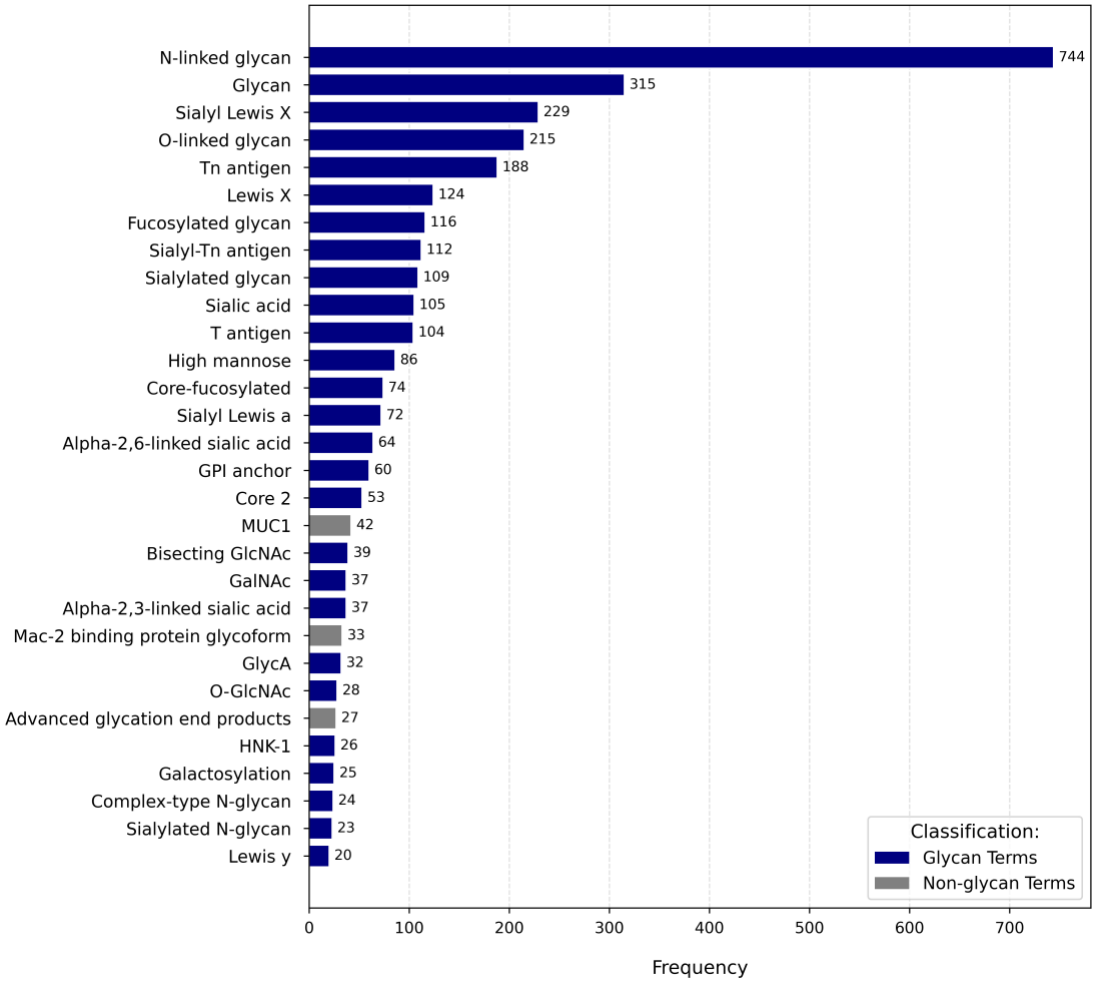
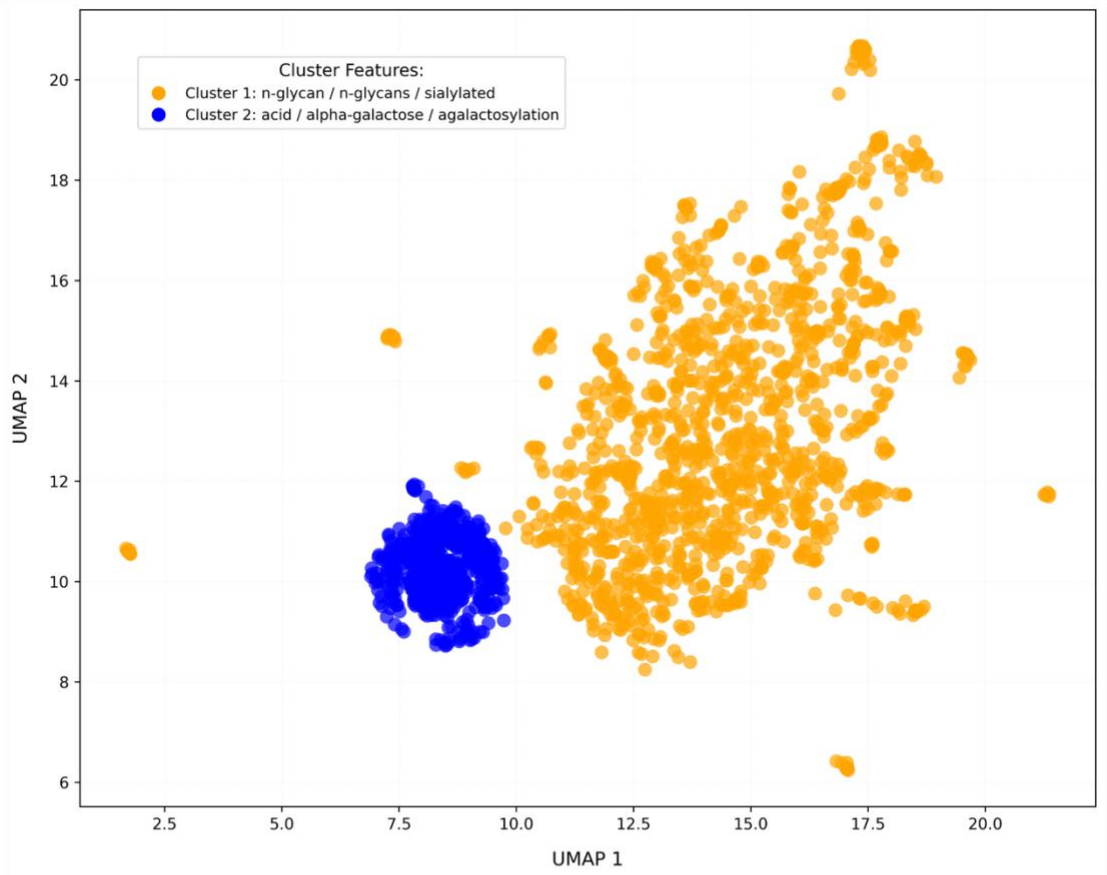


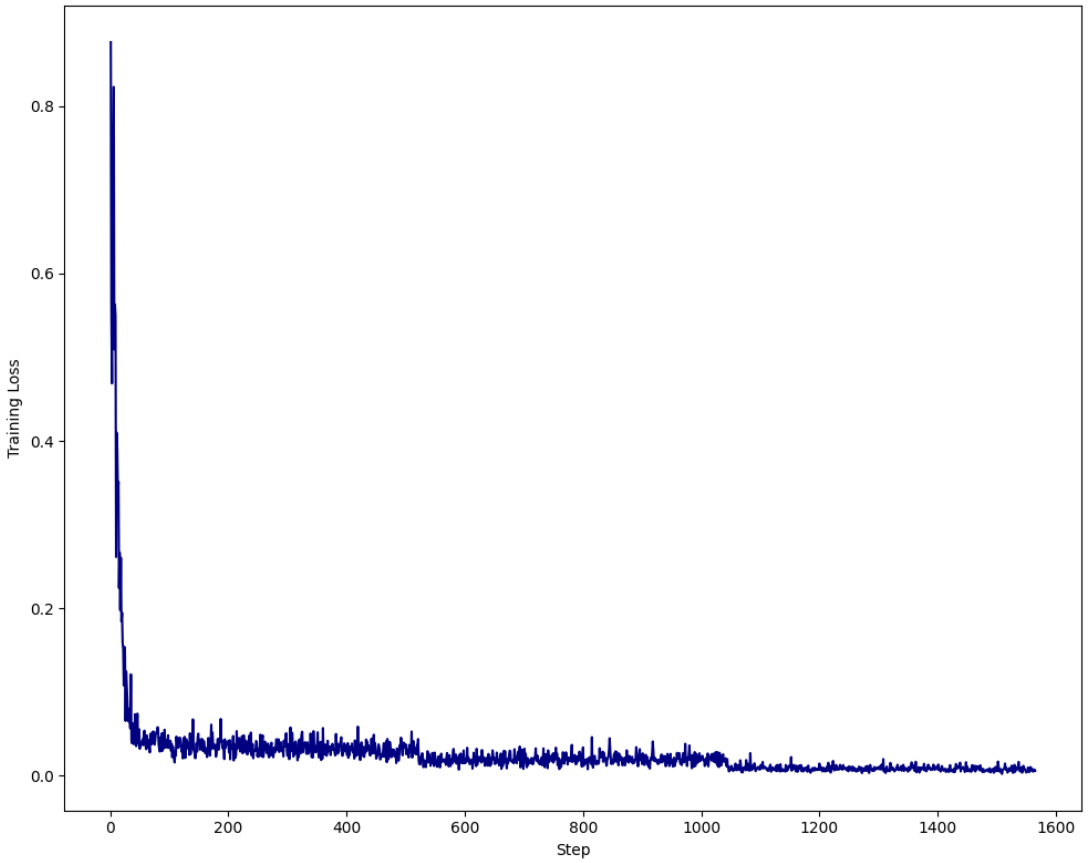
Figure 9. UMAP projection of extracted glycan terms.



Most frequent tokens in cluster	
Cluster 1	n-glycan, n-glycans, sialylated, lewis, fucosylated, antigen, o-linked, o-glycan, o-glycans, sialyl
Cluster 2	acid, alpha-galactose, agalactosylation, acetylation, sites, alpha-l-fucose, alpha-dystroglycan, a2g2s, siglec-8, a2g2

Figure 10. Training loss curve for GPT-4o model fine-tuning.

Training loss versus optimization step during distillation of GPT-4o to the compact 4o-mini model on the glycan-biomarker PubMed abstract corpus.



Tables

Table 1. Classification of glycan notations and nomenclatures.

Category	Examples
Monosaccharide	GlcNAc (N-acetyl-glucosamine), Gal (galactose), Man (mannose), Fuc (fucose), Neu5Ac (sialic acid)
Composition	Hex5HexNAc4Fuc1, NeuAc(2)Hex(3)HexNAc(2)
IUPAC	Galβ1-4GlcNAc, Neu5Acα2-6Galβ1-4GlcNAc, Fucα1-2Galβ1-3GlcNAc
LINUCS	Gal(b1-3)GalNAc(a1-Ser/Thr), Man(a1-6)[Man(a1-3)]Man
GlycoCT	RES 1b:b-dgal-HEX-1:5 2s:n-acetyl ...
WURCS	WURCS=1.0/1,0/[12122h|1,5|2*NCC/3=O]
SNFG Symbol	Gal (●), GlcNAc (■), Fuc (▲)
GlyTouCan ID	G00053MO, G00123AB
N-linked, Code	FA2, FA2G2, A3F, G1Fa
N-linked, Textual	Core-fucosylated bi-antennary complex N-glycan, asialylated tri-antennary complex-type
Generic Class	N-linked, O-linked, glycosphingolipid, O-oligosaccharide
Antigen, Serological	Sialyl-Lewis X, T antigen, Tn antigen, Sd^a antigen, TF antigen, Type-I blood group H antigen
Glycolipid Abbreviation	GM3, GM1a, GD1b, Gb3

Antigen Abbreviation	sLe^x, sLe^a, TF, Le^y
Description of Process	sialylated, core-fucosylated, branched
Description of Property	Acidic glycans, free glycans

Table 2. Representative terms for each abstract cluster (k = 3).

Number of Clusters = 3		
Cluster	**Abstracts**	**Top 10 tokens**
0	3,173	muc1, antigen, surface, tumor, antibodies, antibody, acid, 1, receptor, activity
1	1,220	serum, biomarkers, disease, igg, biomarker, psa, hcc, samples, p, levels
2	202	hiv-1, cd4, gp120, env, virus, envelope, neutralizing, neutralization, hiv, antibodies

Table 3. Representative terms for each abstract cluster (k = 6).

Number of Clusters (k) = 6		
Cluster	**Abstracts**	**Top 10 tokens**
0	1,789	tumor, antigen, levels, tn, breast, surface, antibodies, mice, increased, normal
1	1,394	serum, hcc, psa, biomarkers, prostate, biomarker, samples, liver, mass, n-glycan
2	704	antigen, surface, receptor, domain, beta, glycoproteins, acid, alpha, structure, lewis
3	269	hiv-1, gp120, cd4, env, virus, envelope, neutralizing, neutralization, hiv, antibodies
4	245	igg, fc, disease, glyca, gamma, iga, igan, sub, p, iga1
5	194	muc1, mucin, breast, repeat, tandem, peptide, core, o-glycosylation, glycosylated, tumor

Table 4. Summary of extracted glycan entities.

Synonymous or closely related glycan terms were grouped into broad structural classes. Specific names with a frequency of 1, if not binned into the broader terms under "Representative Name", were excluded from the table. Note that some motifs can be structurally related but functionally distinct.

Representative Name	Synonyms / Variants	Aggregated Count
N-linked Glycans	N-glycan, N-glycans, N-glycosylation, N-linked glycosylation, N-linked oligosaccharides, N-linked carbohydrates, "N-glycan branching," "N-glycosylation site", etc.	~566
O-linked Glycans	O-glycosylation, O-glycan, O-glycans, O-linked glycans, O-linked glycosylation, O-linked oligosaccharides	~118
Tn Antigen	Tn antigen, Tn, Tn-antigen, Tn (GalNAc)	~147
Generic Sialic Acid / Sialylation	sialic acid, sialylation, sialylated glycans, sialylated N-glycans, sialic acids, sialic acid residues	~161
Sialyl-Tn Antigen	Sialyl-Tn, sialyl-Tn antigen, Sialyl-Tn, Sialyl-Tn antigen, sialosyl-Tn antigen, sialyl-Tn (STn)	~57
Fucosylation	fucosylation, core fucosylation, fucosylated glycans/glycan, core-fucosylated N-glycans	~95
Lewis Antigens	Lewis X (or Lewis(x), Lewis x, LewisX, Le(x), Lewis X antigen), plus Lewis Y (LewisY, Lewis(y), Le(y))	~92
MUC1-related Glycans	MUC1, MUC1 mucin, MUC1 glycan, MUC1 glycopeptide, MUC1 glycoforms	~37
GalNAc Residues	GalNAc, N-acetylgalactosamine, N-acetylgalactosamine (GalNAc)	~32

GlcNAc-related Glycans	N-acetylglucosamine, GlcNAc, "bisecting GlcNAc" (or bisecting N-acetylglucosamine)	~41
Mannose-rich Glycans	mannose, high mannose, high-mannose glycans, high-mannose N-glycans, high-mannose-type glycans, Man	~46
GPI Anchors	glycosylphosphatidylinositol (GPI), glycosylphosphatidylinositol anchor, glycosylphosphatidylinositol glycan	~18
SSEA Antigens	SSEA-4, SSEA-3	~21
Generic "Glycan" Terms	glycan, glycans, glycan structures, glycan structure (i.e. non-specific "glycan" labels)	~43

References

1. Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. *Nat Rev Mol Cell Biol*. Jun 22 2012;13(7):448-62. doi:10.1038/nrm3383

2. Moremen KW, Haltiwanger RS. Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat Chem Biol*. Sep 2019;15(9):853-864. doi:10.1038/s41589-019-0350-2

3. National Research Council (US) Committee on Research Opportunities in Biology. Molecular Structure and Function. *Opportunities in Biology*. National Academies Press (US); 1989:chap 3. https://www.ncbi.nlm.nih.gov/books/NBK217812/

4. Colley KJ, Varki A, Haltiwanger RS, Kinoshita T. Cellular Organization of Glycosylation. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of Glycobiology [Internet]*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2022:chap 4.

5. Helenius A, Aebi M. Intracellular functions of N-linked glycans. *Science*. Mar 23 2001;291(5512):2364-9. doi:10.1126/science.291.5512.2364

6. Gagneux P, Hennet T, Varki A. Biological Functions of Glycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of Glycobiology [Internet]*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2022:chap 7.

7. Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer*. Sep 2015;15(9):540-55. doi:10.1038/nrc3982

8. Wang M, Zhu J, Lubman DM, Gao C. Aberrant glycosylation and cancer biomarker discovery: a promising and thorny journey. *Clin Chem Lab Med*. Mar 26 2019;57(4):407-416. doi:10.1515/cclm-2018-0379

9. Magalhaes A, Duarte HO, Reis CA. The role of O-glycosylation in human disease. *Mol Aspects Med*. Jun 2021;79:100964. doi:10.1016/j.mam.2021.100964

10. Lin Y, Lubman DM. The role of N-glycosylation in cancer. *Acta Pharm Sin B*. Mar 2024;14(3):1098-1110. doi:10.1016/j.apsb.2023.10.014

11. Gendler SJ, Lancaster CA, Taylor-Papadimitriou J, et al. Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J Biol Chem*. Sep 5 1990;265(25):15286-93.

12. Carneiro F, David L, Sobrinho-Simoes M. Sialyl-Tn expression in gastric carcinoma. *Eur J Cancer*. 1994;30A(9):1398-9. doi:10.1016/0959-8049(94)90194-5

13. Novak J, Barratt J, Julian BA, Renfrow MB. Aberrant Glycosylation of the IgA1 Molecule in IgA Nephropathy. *Semin Nephrol*. Sep 2018;38(5):461-476. doi:10.1016/j.semnephrol.2018.05.016

14. Kudelka MR, Stowell SR, Cummings RD, Neish AS. Intestinal epithelial glycosylation in homeostasis and gut microbiota interactions in IBD. *Nat Rev Gastroenterol Hepatol*. Oct 2020;17(10):597-617. doi:10.1038/s41575-020-0331-7

15. Krishnamurthy VR, Sardar MY, Ying Y, et al. Glycopeptide analogues of PSGL-1 inhibit P-selectin in vitro and in vivo. *Nat Commun*. Mar 31 2015;6:6387. doi:10.1038/ncomms7387

16. Yang J, Xu R, Wang C, Qiu J, Ren B, You L. Early screening and diagnosis strategies of pancreatic cancer: a comprehensive review. *Cancer Commun (Lond)*. Dec 2021;41(12):1257-1274. doi:10.1002/cac2.12204

17. Guo HB, Lee I, Kamar M, Akiyama SK, Pierce M. Aberrant N-glycosylation of beta1 integrin causes reduced alpha5beta1 integrin clustering and stimulates cell migration. *Cancer Res*. Dec 1 2002;62(23):6837-45.

18. Aoyagi Y, Isemura M, Suzuki Y, et al. Fucosylated alpha-fetoprotein as marker of early hepatocellular carcinoma. *Lancet*. Dec 14 1985;2(8468):1353-4. doi:10.1016/s0140-6736(85)92643-1

19. Oh MJ, Lee SH, Kim U, An HJ. In-depth investigation of altered glycosylation in human haptoglobin associated cancer by mass spectrometry. *Mass Spectrom Rev*. Mar 2023;42(2):496-518. doi:10.1002/mas.21707

20. Llop E, Ferrer-Batalle M, Barrabes S, et al. Improvement of Prostate Cancer Diagnosis by Detecting PSA Glycosylation-Specific Changes. *Theranostics*. 2016;6(8):1190-204. doi:10.7150/thno.15226

21. Huang YL, Liang CY, Labitzky V, et al. Site-specific N-glycosylation of integrin alpha2 mediates collagen-dependent cell survival. *iScience*. Oct 22 2021;24(10):103168. doi:10.1016/j.isci.2021.103168

22. Freeze HH. Understanding human glycosylation disorders: biochemistry leads the charge. *J Biol Chem*. Mar 8 2013;288(10):6936-45. doi:10.1074/jbc.R112.429274

23. York WS, Mazumder R, Ranzinger R, et al. GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology*. Jan 28 2020;30(2):72-73. doi:10.1093/glycob/cwz080

24. McNaught AD. Nomenclature of carbohydrates (recommendations 1996). *Adv Carbohydr Chem Biochem*. 1997;52(2-Carb-38. Use of symbols for defining oligosaccharide structures):43-177.

25. Bohne-Lang A, Lang E, Forster T, von der Lieth CW. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res*. Nov 1 2001;336(1):1-11. doi:10.1016/s0008-6215(01)00230-0

26. Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr Res*. Aug 11 2008;343(12):2162-71. doi:10.1016/j.carres.2008.03.011

27. Matsubara M, Aoki-Kinoshita KF, Aoki NP, Yamada I, Narimatsu H. WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. *J Chem Inf Model*. Apr 24 2017;57(4):632-637. doi:10.1021/acs.jcim.6b00650

28. Neelamegham S, Aoki-Kinoshita K, Bolton E, et al. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology*. Aug 20 2019;29(9):620-624. doi:10.1093/glycob/cwz045

29. Fujita A, Aoki NP, Shinmachi D, et al. The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res*. Jan 8 2021;49(D1):D1529-D1533. doi:10.1093/nar/gkaa947

30. Ranzinger R, Kochut KJ, Miller JA, Eavenson M, Lutteke T, York WS. GLYDE-II: The GLYcan data exchange format. *Perspect Sci (Neth)*. Jan 2017;11:24-30. doi:10.1016/j.pisc.2016.05.013

31. Ranzinger R, York WS. GlycomeDB. *Methods Mol Biol*. 2015;1273:109-24. doi:10.1007/978-1-4939-2343-4_8

32. Kim JD, Wang Y. PubAnnotation: a persistent and sharable corpus and annotation repository. presented at: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; 2012; Montreal, Canada.

33. Kim JD, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB. Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. *Bioinformatics*. Nov 1 2019;35(21):4372-4380. doi:10.1093/bioinformatics/btz227

34. Thomes L, Burkholz R, Bojar D. Glycowork: A Python package for glycan data science and machine learning. *Glycobiology*. Nov 18 2021;31(10):1240-1244. doi:10.1093/glycob/cwab067

35. Hsu YY, Kao HY. Curatable Named-Entity Recognition Using Semantic Relations. *IEEE/ACM Trans Comput Biol Bioinform*. Jul-Aug 2015;12(4):785-92. doi:10.1109/TCBB.2014.2366770

36. Gerner M, Sarafraz F, Bergman CM, Nenadic G. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*. Aug 15 2012;28(16):2154-61. doi:10.1093/bioinformatics/bts332

37. Gill JK, Chetty M, Lim S, Hallinan J. Large language model based framework for automated extraction of genetic interactions from unstructured data. *PLoS One*. 2024;19(5):e0303231. doi:10.1371/journal.pone.0303231

38. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Feb 15 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

39. Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing [preprint]. *arXiv*. September 16, 2021 2021;doi:10.1145/3458754

40. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners [preprint]. *arXiv*. July 22, 2020 2020;doi:10.48550/arXiv.2005.14165

41. Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv*. Mon, 27 Feb 2023 2023;doi:10.48550/arXiv.2302.13971

42. OpenAI. OpenAI Documentation: Structured Outputs. OpenAI Platform. Accessed 1 April, 2025, 2025. https://platform.openai.com/docs/guides/structured-outputs

43. Shah SV. Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records. *JAMA Netw Open*. Aug 1 2024;7(8):e2425953. doi:10.1001/jamanetworkopen.2024.25953

44. Dean L. Blood Groups and Red Cell Antigens [Internet]. National Center for Biotechnology Information (US); 2005:chap 5, The ABO blood group.

45. Yamamoto F, Clausen H, White T, Marken J, Hakomori S. Molecular genetic basis of the histo-blood group ABO system. *Nature*. May 17 1990;345(6272):229-33. doi:10.1038/345229a0

46. Lee T, Teng TZJ, Shelat VG. Carbohydrate antigen 19-9 - tumor marker: Past, present, and future. *World J Gastrointest Surg*. Dec 27 2020;12(12):468-490. doi:10.4240/wjgs.v12.i12.468

47. Hanisch FG, Uhlenbruck G, Dienst C. Structure of tumor-associated carbohydrate antigen Ca 19-9 on human seminal-plasma glycoproteins from healthy donors. *Eur J Biochem*. Nov 2 1984;144(3):467-73. doi:10.1111/j.1432-1033.1984.tb08489.x

48. Varki A, Kornfeld S. Historical Background and Overview. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of Glycobiology [Internet]*. 4th ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2022:chap 1.

49. Varki A, Cummings RD, Aebi M, et al. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology*. Dec 2015;25(12):1323-4. doi:10.1093/glycob/cwv091

50. Tanaka K, Aoki-Kinoshita KF, Kotera M, et al. WURCS: the Web3 unique representation of carbohydrate structures. *J Chem Inf Model*. Jun 23 2014;54(6):1558-66. doi:10.1021/ci400571e

51. Aoki-Kinoshita K, Agravat S, Aoki NP, et al. GlyTouCan 1.0--The international glycan structure repository. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1237-42. doi:10.1093/nar/gkv1041

52. Campbell MP, Packer NH. UniCarbKB: New database features for integrating glycan structure abundance, compositional glycoproteomics data, and disease associations. *Biochim Biophys Acta*. Aug 2016;1860(8):1669-75. doi:10.1016/j.bbagen.2016.02.016

53. Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. Jan 4 2016;44(D1):D1214-9. doi:10.1093/nar/gkv1031

54. UniProt C. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res*. Jan 6 2025;53(D1):D609-D617. doi:10.1093/nar/gkae1010

55. Burley SK, Bhatt R, Bhikadiya C, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res*. Jan 6 2025;53(D1):D564-D574. doi:10.1093/nar/gkae1091

56. Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res*. Jan 7 2022;50(D1):D571-D577. doi:10.1093/nar/gkab1045

57. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res*. Jan 6 2025;53(D1):D672-D677. doi:10.1093/nar/gkae909

58. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. Jan 8 2019;47(D1):D1038-D1043. doi:10.1093/nar/gky1151

59. Lee S, Ono T, Masaaki S, et al. Updates implemented in version 4 of the GlyCosmos Glycoscience Portal. *Anal Bioanal Chem*. Feb 2025;417(5):907-919. doi:10.1007/s00216-024-05692-0

60. Said IT, Shamsuddin AM, Sherief MA, Taleb SG, Aref WF, Kumar D. Comparison of different techniques for detection of Gal-GalNAc, an early marker of colonic neoplasia. *Histol Histopathol*. Apr 1999;14(2):351-7. doi:10.14670/HH-14.351

61. Hall N, Ong EG, Ade-Ajayi N, et al. T cryptantigen activation is associated with advanced necrotizing enterocolitis. *J Pediatr Surg*. May 2002;37(5):791-3. doi:10.1053/jpsu.2002.32289

62. Grant HW, Hadley GP, Adhikari M, Fernandes-Costa F. T-cryptantigen (TCA) activation in surgical neonates: a hidden problem. *Pediatr Surg Int*. Dec 1998;14(3):204-7. doi:10.1007/s003830050488

63. Grunwald-Gruber C, Thader A, Maresch D, Dalik T, Altmann F. Determination of true ratios of different N-glycan structures in electrospray ionization mass spectrometry. *Anal Bioanal Chem*. Apr 2017;409(10):2519-2530. doi:10.1007/s00216-017-0235-8

64. Neustein A, Imambi SS, Rodrigues M, Teixeira A, Ferreira L. Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature. In: Neustein A, ed. *Text Mining of Web-Based Medical Content*. 1 ed. De Gruyter, Inc.; 2014:chap 1.

65. Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*. Dec 27 2019;20(1):735. doi:10.1186/s12859-019-3321-4

66. Seal RL, Braschi B, Gray K, et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res*. Jan 6 2023;51(D1):D1003-D1009. doi:10.1093/nar/gkac888

67. Schriml LM, Munro JB, Schor M, et al. The Human Disease Ontology 2022 update. *Nucleic Acids Res*. Jan 7 2022;50(D1):D1255-D1261. doi:10.1093/nar/gkab1063

68. Rosário-Ferreira N, Marques-Pereira C, Pires M, et al. The Treasury Chest of Text Mining: Piling Available Resources for Powerful Biomedical Text Mining. *BioChem*. 2021;1(2):60-80. doi:https://doi.org/10.3390/biochem1020007

69. Martinez K, Agirre J, Akune Y, et al. Functional implications of glycans and their curation: insights from the workshop held at the 16th Annual International Biocuration Conference in Padua, Italy. *Database (Oxford)*. Aug 13 2024;2024doi:10.1093/database/baae073

70. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics*. 2003;19 Suppl 1:i180-2. doi:10.1093/bioinformatics/btg1023

71. Islamaj R, Wei CH, Lai PT, et al. The biomedical relationship corpus of the BioRED track at the BioCreative VIII challenge and workshop. *Database (Oxford)*. Aug 9 2024;2024doi:10.1093/database/baae071

72. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *Ieee T Signal Proces*. Nov 1997;45(11):2673-2681. doi:Doi 10.1109/78.650093

73. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [preprint]. *arXiv*. 9 Aug 2015;doi:https://doi.org/10.48550/arXiv.1508.01991

74. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [preprint]. *arXiv*. 11 Oct 2018;doi:10.48550/arXiv.1810.04805

75. National Center for Biotechnology Information (NCBI) / National Library of Medicine. PubMed User Guide. Updated February 25, 2025. Accessed March 6, 2025. https://pubmed.ncbi.nlm.nih.gov/help/

76. Gibney G, Baxevanis AD. Searching NCBI databases using Entrez. *Curr Protoc Bioinformatics*. Jun 2011;Chapter 1:1 3 1-1 3 25. doi:10.1002/0471250953.bi0103s34

77. Schuler GD, Epstein JA, Ohkawa H, Kans JA. Entrez: molecular biology database and retrieval system. *Methods Enzymol*. 1996;266:141-62. doi:10.1016/s0076-6879(96)66012-1

78. Sayers E. The E-utilities In-Depth: Parameters, Syntax and More. *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US); 2010. https://www.ncbi.nlm.nih.gov/books/NBK25499/

79. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*. Sep-Oct 2011;18(5):601-6. doi:10.1136/amiajnl-2011-000163

80. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*. May 1 2018;34(9):1547-1554. doi:10.1093/bioinformatics/btx815

81. Yang C, Deng J, Chen X, An Y. SPBERE: Boosting span-based pipeline biomedical entity and relation extraction via entity information. *J Biomed Inform*. Sep 2023;145:104456. doi:10.1016/j.jbi.2023.104456

82. Dogan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform*. Feb 2014;47:1-10. doi:10.1016/j.jbi.2013.12.006

83. Yang J, Liu C, Deng W, et al. Enhancing Phenotype Recognition in Clinical Notes Using Large Language Models: PhenoBCBERT and PhenoGPT. *ArXiv*. Nov 9 2023;

84. Li Z, Yang Z, Xiang Y, Luo L, Sun Y, Lin H. Exploiting sequence labeling framework to extract document-level relations from biomedical texts. *BMC Bioinformatics*. Mar 27 2020;21(1):125. doi:10.1186/s12859-020-3457-2

85. Le HQ, Can DC, Collier N. Exploiting document graphs for inter sentence relation extraction. *J Biomed Semantics*. Jun 3 2022;13(1):15. doi:10.1186/s13326-022-00267-3

86. Li Y, Hui L, Zou L, et al. Relation Extraction in Biomedical Texts Based on Multi-Head Attention Model With Syntactic Dependency Feature: Modeling Study. *JMIR Med Inform*. Oct 20 2022;10(10):e41136. doi:10.2196/41136

87. Weber L, Sanger M, Garda S, Barth F, Alt C, Leser U. Chemical-protein relation extraction with ensembles of carefully tuned pretrained language models. *Database (Oxford)*. Nov 18 2022;2022doi:10.1093/database/baac098

88. Miranda-Escalada A, Mehryary F, Luoma J, et al. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical-protein relations. *Database (Oxford)*. Nov 28 2023;2023doi:10.1093/database/baad080

89. OpenAI. OpenAI: Introducing GPT-4.1 in the API. Accessed April 14, 2025. https://openai.com/index/gpt-4-1/

90. Pokrass M. OpenAI: Introducing Structured Outputs in the API. Accessed August 6, 2024. https://openai.com/index/introducing-structured-outputs-in-the-api/

91. OpenAI. OpenAI platform: Retrieval. Accessed March 31, 2025. https://platform.openai.com/docs/guides/retrieval

92. OpenAI. OpenAI Core Concepts: Function calling. Accessed March 31, 2025. https://platform.openai.com/docs/guides/function-calling?api-mode=chat

93. OpenAI. OpenAI Core Concepts: Images and vision. Accessed March 31, 2025. https://platform.openai.com/docs/guides/images-vision?api-mode=chat

94. OpenAI. OpenAI Built-In Tools: File search. Accessed March 31, 2025. https://platform.openai.com/docs/guides/images-vision?api-mode=chat

95. Jason Wei; Xuezhi Wang; Dale Schuurmans; Maarten Bosma; Brian Ichter; Fei Xia ECQLDZ. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*. 2022;doi:10.48550/arXiv.2201.11903

96. OpenAI. OpenAI Platform: Fine-tuning. 2025. https://platform.openai.com/docs/guides/fine-tuning/create-a-fine-tuned-model

97. Mehta AY, Cummings RD. GlycoGlyph: a glycan visualizing, drawing and naming application. *Bioinformatics*. Jun 1 2020;36(11):3613-3614. doi:10.1093/bioinformatics/btaa190

Appendices

List of Supplementary Information

SI 1. OpenAI API Request Prompt Text. 81
SI 2. OpenAI API Request JSON Schema 82

SI 1. OpenAI API Request Prompt Text.

```
Objective: Extract relationships between glycans and diseases from a given PubMed abstract. Return one JSON object that follows the JSON schema.
Instructions:
I. Extract Entities:
(If an entity is not mentioned/missing, return an empty string.)
Use the "PMID", "Title", "Abstract", and "Keywords" rows of the abstract text. Identify 13 entities:
    1. "glycan_entity"
    2. "glycan_entity_change"
    3. "glycoprotein_entity_name"
    4. "glycoprotein_gene_symbol"
    5. "glycoenzyme_entity_name"
    6. "glycoenzyme_gene_symbol"
    7. "disease_name"
    8. "medical_intervention"
    9. "biomarker_role"
    10. "organism"
    11. "specimen_type"
    12. "pmid"
    13. "sentence_in_abstract"

II. Map Additional Data to *_mapped fields:
Use the "Chemical names", and "MeSH names" rows of the abstract text to supplement the extracted entities (reference fields) from part I.
- Multiple values in "Keywords", "Chemical names", and "MeSH names" are provided as "|"-separated strings.
Try to fill in the four additional 'mapped' fields:
  1. "glycoprotein_mapped", for any "Chemical names" values mapping to "glycoprotein_entity_name".
  2. "glycoprotein_mapped", for any "Chemical names" values mapping to "glycoenzyme_entity_name".
  3. "disease_mapped", for any "MeSH names" values mapping to "disease_name".
  4. "specimen_mapped", for any "MeSH names" values mapping to "specimen_type".
##Note: When there are multiple entities in the mapped field, return them as "|"-separated objects while maintaining their order as in the reference field
- use an empty string "" for each unmapped value.

III. Handling Multiple Entities:
- When there are multiple (i.e., "|"-separated) values in "glycan_entity", "glycan_entity_change", or "disease_name", split them into separate objects in the "glycan_biomarkers" array.
- Pipe-separated strings are allowed for all other fields.
- Prioritize precision over sensitivity. When multiple glycan terms are associated with multiple diseases, report each association as a separate "glycan_biomarkers" object, accurately.
```

SI 2. OpenAI API Request JSON Schema

```
{
  "type": "object",
  "properties": {
    "glycan_biomarkers": {
      "type": "array",
      "description": "One glycan_biomarkers entity consists of one change in one glycan_entity associated with one associated_disease",
      "items": {
        "type": "object",
        "properties": {
          "assessed_biomarker_entity": {
            "type": "object",
            "properties": {
              "glycan_entity": {
                "type": "string",
                "description": "glycan/complex carbohydrate names/motifs/epitopes/notations, or structures(e.g. branching, bisecting, core-fucosylated). Exclude enzymes, glycoproteins, non-glycans and NS peaks (e.g. GP1)."
              },
              "glycan_entity_change": {
                "type": "string",
                "description": "Change associated with the glycan_entity",
                "enum": [
                  "increased level",
                  "decreased level",
                  "increased ratio",
                  "decreased ratio",
                  "altered level",
                  "unaltered level",
                  "acquisition of structure",
                  "loss of structure",
                  "undetermined"
                ]
              },
              "glycoprotein_entity_name": {
                "type": "string",
                "description": "Protein(s) expressessing the glycan_entity. Do not fill glycoenzyme entities. Include ONLY protein names in this field."
              },
              "glycoprotein_gene_name": {
                "type": "string",
                "description": "Gene symbol of glycoprotein_entity_name if explicitly mentioned."
              },
              "glycoprotein_mapped": {
                "type": "string",
                "description": "Copy Chemical_names value here if any of them maps to the glycoprotein_entity_name value."
              },
              "glycoenzyme_entity_name": {
                "type": "string",
                "description": "Glycosyltransferases/glycosylhydrolases that directly caused the glycan_entity_change. Exclude enzymes used in experimental pipelines (e.g. PNGase F)."
              },
              "glycoenzyme_gene_name": {
                "type": "string",
                "description": "Gene symbol of glycoprotein_entity_name if explicitly mentioned"
              },
              "glycoenzyme_mapped": {
                "type": "string",
                "description": "Copy Chemical_names value here if any of them maps to the glycoenzyme_entity_name value."
              }
```

```
        }
    },
    "required": [
        "glycan_entity",
        "glycan_entity_change",
        "glycoprotein_entity_name",
        "glycoprotein_gene_name",
        "glycoprotein_mapped",
        "glycoenzyme_entity_name",
        "glycoenzyme_gene_name",
        "glycoenzyme_mapped"
    ],
    "additionalProperties": false
},
"associated_disease": {
    "type": "object",
    "properties": {
        "disease_name": {
            "type": "string",
            "description": "Disease(s) or condition(s) directly causing the change in glycan_entity. Exclude fill cells/tissues/organs."
        },
        "disease_mapped": {
            "type": "string",
            "description": "Copy MeSH_names value here if any of them maps to the disease_name value."
        },
        "medical_intervention": {
            "type": "string",
            "description": "Mentioned therapeutic or surgical measures applied to treat the disease if association with glycan change is apparent."
        }
    },
    "required": [
        "disease_name",
        "disease_mapped",
        "medical_intervention"
    ],
    "additionalProperties": false
},
"biomarker_role": {
    "type": "array",
    "description": "Classified biomarker roles using definitions from the FDA-NIH BEST Resource",
    "items": {
        "type": "string",
        "enum": [
            "diagnostic",
            "monitoring",
            "predictive",
            "prognostic",
            "response",
            "risk",
            "safety"
        ]
    }
},
"sample": {
    "type": "object",
    "properties": {
```

```
                        "organism": {
                            "type": "string",
                            "description": "Organism of which the specimen was collected from"
                        },
                        "specimen_type": {
                            "type": "string",
                            "description": "Tissue/cell type/cell line/organ where the glycan_entity is being observed"
                        },
                        "specimen_mapped": {
                            "type": "string",
                            "description": "Copy MeSH_names value here if any of them maps to the specimen_type value."
                        }
                    },
                    "required": [
                        "organism",
                        "specimen_type",
                        "specimen_mapped"
                    ],
                    "additionalProperties": false
                },
                "evidence": {
                    "type": "object",
                    "properties": {
                        "pmid": {
                            "type": "string",
                            "description": "PMID. Mandatory field"
                        },
                        "sentence_in_abstract": {
                            "type": "string",
                            "description": "A complete sentence from the abstract text mentioning the glycan entity, verbatim"
                        }
                    },
                    "required": [
                        "pmid",
                        "sentence_in_abstract"
                    ],
                    "additionalProperties": false
                }
            },
            "required": [
                "assessed_biomarker_entity",
                "associated_disease",
                "biomarker_role",
                "sample",
                "evidence"
            ],
            "additionalProperties": false
        }
    }
},
"required": [
    "glycan_biomarkers"
],
"additionalProperties": false
}
```