Verifying LLM-extracted text with token alignment

A. Sina Booeshaghi*[1] and Aaron Streets[1,2,3,4]

[1]Department of Bioengineering, University of California Berkeley, Berkeley, USA
[2]Center for Computational Biology, University of California, Berkeley, USA
[3]Biophysics Graduate Group, University of California, Berkeley, USA
[4]Chan Zuckerberg Biohub – San Francisco, USA

Abstract

Large language models excel at text extraction, but they sometimes hallucinate. A simple way to avoid hallucinations is to remove any extracted text that does not appear in the original source. This is easy when the extracted text is contiguous (findable with exact string matching), but much harder when it is discontiguous. Techniques for finding discontiguous phrases depend heavily on how the text is split—i.e., how it is tokenized. In this study, we show that splitting text along subword boundaries, with LLM-specific tokenizers, and aligning extracted text with ordered alignment algorithms, improves alignment by about 50% compared to word-level tokenization. To demonstrate this, we introduce the Berkeley Ordered Alignment of Text (`BOAT`) dataset, a modification of the Stanford Question Answering Dataset (`SQuAD`) that includes non-contiguous phrases, and `BIO-BOAT` a biomedical variant built from 51 bioRxiv preprints. We show that text-alignment methods form a partially ordered set, and that ordered alignment is the most practical choice for verifying LLM-extracted text. We implement this approach in `taln`, which enumerates ordinal subword alignments.

1 Introduction

Text extraction is a basic task for large language models (LLMs). A user asks the model to extract information from a passage, and the model is expected to return text that actually appears in it. Reliable text extraction is enabling. It allows users from many domains to turn unstructured text—such as a scientific article, historical newspaper, health record, or financial report—into structured data [Gupta et al., 2022, Dunn et al., 2022, Agrawal et al., 2022, Kartchner et al., 2023, Huang et al., 2023, Dagdelen et al., 2024, Zhang et al., 2024, Polak and Morgan, 2024, Chataut et al., 2024, Perot et al., 2024, Dagli et al., 2024, Le Guellec et al., 2024, Liu et al., 2024, Castro et al., 2024, Gartlehner et al., 2024, Pai et al., 2024, Ntinopoulos et al., 2025, Tang et al., 2024, Geevarghese et al., 2024, Schilling-Wilhelmi et al., 2025, Azar et al., 2025, Schwitter, 2025, Qin et al., 2025, Yamagishi et al., 2025, Srivastava et al., 2025, Jiang et al., 2025, Fei et al., 2025, Zhou et al., 2025].

In medicine and biology, accurate text extraction is especially important. For instance, organ procurement organizations rely on lab values and vital signs—[1]often buried inside unstructured clinical notes—to judge whether a potential donor's organs are suitable [Adam et al., 2024]. Errors introduced during extraction can potentially misinform clinical decisions, and ultimately harm patients.

Fortunately, text extraction can be verified: we can align the extracted text back to its source (Figure 1). For uninterrupted ("contiguous") phrases, verification is straightforward. An exact match confirms its existence. For non-contiguous phrases, verification is harder; alignment must handle intervening elements. Parenthetical notes, superscripts, symbols, qualifiers, or punctuation may interrupt a logically unified (but practically discontinuous) phrase. For example, the phrase "`natural killer cells`" in Figure 1 cannot be contiguously aligned because the intervening parenthetical expression "`(NK)`" breaks it. Even though the phrase exists, it cannot be found by naive alignment. This situation is particularly acute in biology and medicine where non-contiguous phrases and non-standard nomenclature are highly abundant [Lever et al., 2020, Shinohara et al., 2024].

One way to verify the existence of non-contiguous phrases is with LLM-as-a-judge [Gu et al., 2025]. In this setting, language models are given a passage and extracted piece of text and are prompted to determine its existence. While sometimes effective, LLM-as-a-judge is mismatched to the task. Text is a concrete feature of a passage whereas LLM judgments are semantic and probabilistic distributions over a vocabulary (that includes text from the passage, as well as other text). Therefore, LLMs may approve paraphrases, or hallucinate, making them unsuitable for verifying the existence of text.

Practical issues with LLMs compound this problem. LLM inference is slow, expensive, and non-deterministic.

*Corresponding author: sinab@berkeley.edu

[1]The em dashes reflect our stylistic preference and should not be interpreted as evidence of AI-assisted text generation.

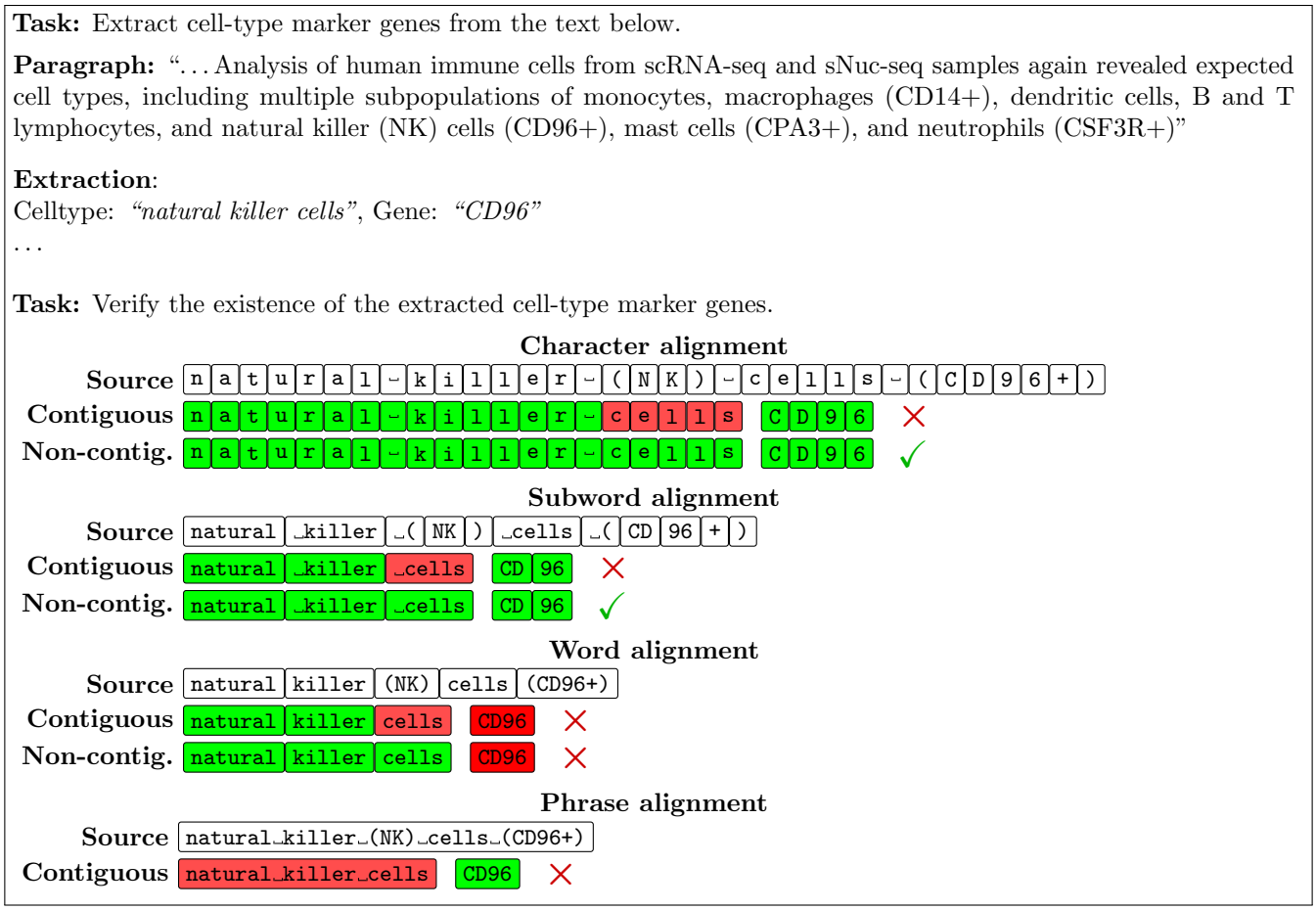

Figure 1: An example of common text-extraction and verification tasks using text from [Hildreth et al., 2021]. LLMs correctly extract marker genes from a sentence, but classical verification tools—that align phrases or words—fail (red). The example cell-type and marker gene are found (green) only when aligned with non-contiguous alignment under subword or character tokenization. Compared to character tokenization, subword tokenization produces fewer tokens resulting in a smaller memory usage.

For example, using GPT5.1 ($1.25 per million input tokens, $10 per million output tokens pricing as of late 2025), verifying a single 10-character string in a 1,000-character passage costs approximately $0.001 and takes 1-2 seconds. Across 100,000 verifications, this amounts to tens to hundreds of dollars and tens of hours of compute—resources spent replicating what deterministic string matching could achieve instantly at negligible cost.

But as we saw, naive methods are insufficient. More complex methods—that split the phrase and align parts of it—offer an improvement. By splitting the text into units (also known as tokens) and aligning each independently, rather than matching consecutively occurring characters, these methods can align non-contiguous text. Because tokens only need to appear in the same order, non-contiguous matches become possible. Removing the contiguity constraint allows alignment to skip irrelevant parts of a sentence.

Tokens can, in principle, be any unit of text—from individual characters to full words or even multi-word expressions—but token size introduces a fundamental tradeoff [Whittington et al., 2024]. Small tokens (e.g. characters) offer maximal flexibility. Every character can be aligned independently, allowing methods to recover highly fragmented phrases (Figure 1). But this flexibility is computationally expensive. If a phrase has k characters and the source has n characters (all the same), aligning at the character level can produce up to $\binom{n}{k} \approx \frac{n^k}{k!}$ possible non-contiguous matches. For example, aligning a 10-character target (e.g. `AAAAAAAAAA` to a 1,000 character (`A...A`) source produces roughly 10^{23} possible alignments, compared to only $n-k+1 = 991$ matches for simple contiguous search. Conversely, very large tokens (e.g. whole sentences) reduce the search space by merging punctuation, symbols, acronyms, or alphanumeric identifiers in ways that may hide LLM-extracted phrases. Thus, token size directly determines what alignments are findable and computationally tractable.

A common tokenization method is word-level: text is split on the spaces between characters and tokens are

Table 1: The Berkeley Ordered Alignment of Text (BOAT) dataset consists of multiple human-annotated text extractions derived from the SQuAD dataset. Each example includes a target text and a corresponding source text, where the target can be identified by direct string matching in the source. BOAT has two parts: a contiguous one, where the full target appears in the source, and a non-contiguous one, where part of the target is ablated and the target cannot be found contiguously. We report the average character (Char), whitespace-delimited word (WS), and subword token (Tok) lengths of both source and target texts. WS corresponds to word-level, and Tok to subword tokenization. Note that a small fraction of targets that are present in the source cannot be found with non-contiguous subword tokenization due to incorrectly set tokenization boundaries present in the SQuAD dataset. We also report the same statistics for the BIO-BOAT dataset.

BOAT	N	$N_{0\text{-align}}$ (%)	μ_{align}	Text	Chars	WS	Tok
Contiguous	35,080	159 (0.4)	8	source	764	121	158
				target	38	6	8
Non-contiguous	35,080	198 (0.6)	7	source	764	121	158
				target	33	5	7
Contiguous (BIO)	5,316	171 (3.2)	94	source	607	91	132
				target	97	14	20
Non-contiguous (BIO)	5,316	584 (11.0)	89	source	607	91	132
				target	90	13	18

whole words. But word-level tokenization breaks down in scientific text. As shown in Figure 1, if an LLM extracts the gene name "CD96" from the phrase "natural killer (NK) cells (CD96+)", word-level alignment fails because that tokenization merges the gene name with parentheses, and a plus sign. This situation is quite common. In biological text, gene names often contain characters like parentheses, plus signs, and mixed alphanumeric formatting that are inconsistently attached to genes, causing word-level methods to miss alignments.

In this work, we study how tokenization and alignment strategy affect the recovery of missing alignments resulting from non-contiguous phrases. We introduce two benchmark datasets, formalize text alignment to motivate a practical alignment method, and evaluate the impact of tokenization with this method. Based on these benchmarks, we release an open-source tool, called taln, for accurate non-contiguous text alignment.

2 Results

We constructed two benchmark datasets: BOAT, derived from SQuAD [Rajpurkar et al., 2016, 2018], and BIO-BOAT, a biomedical extension built from 51 bioRxiv preprints (Table 1). Using these datasets, and a formalization of alignment methods, we show that ordered alignment is most practical for aligning non-contiguous phrases and that tokenization greatly impacts alignment accuracy. Furthermore, subword tokenization [Sennrich et al., 2016] recovers substantially more alignments than word-level tokenization while remaining computationally practical.

2.1 Dataset design

BOAT consists of 35,080 source–target pairs derived from SQuAD, where each target aligns contiguously in its source passage. Sources average 764 characters (121 whitespace tokens; 158 subword tokens), and targets average 38 characters (Table 1). To evaluate non-contiguous alignment, we constructed a variant in which one interior token is removed from each target, ensuring that the resulting phrase can no longer be found by exact contiguous alignment.

To test whether alignment behavior generalizes beyond general-domain text, we constructed BIO-BOAT, containing 5,316 source–target pairs extracted from 51 bioRxiv preprints, with both contiguous and non-contiguous variants (Methods) [van Vliet et al., 2022, Cobo-López et al., 2023, Meyer et al., 2023, Zimyanin et al., 2023, Singer et al., 2024, Shen et al., 2024, Frolov et al., 2024, Ryan et al., 2024, Pfalzgraf et al., 2024, Onishi et al., 2024, Wondim et al., 2024, Crnkovic et al., 2024, Kallenborn et al., 2024, Peroutka-Bigus et al., 2024, Azim et al., 2024, Bhalla et al., 2024, Jones et al., 2024, Fei et al., 2024, Combredet and Brunet, 2024, Simmonds et al., 2025, Adhinarta et al., 2025, Elangovan et al., 2025, Tang et al., 2025, Maya-Miles et al., 2025, Samame-Caramutti et al., 2025, Nguyen et al., 2025, Terstege and Epp, 2025, Chen et al., 2025, Li et al., 2025, Kattner et al., 2025, Shruti and Vijay Prakash, 2025, Velázquez et al., 2025, Rull, 2025, Dölker et al., 2025, Keikhosravi et al., 2025, Agazzi et al., 2025, Choltus et al., 2025, Maestro et al., 2025, Widney et al., 2025, Smith et al., 2025, Marín, 2025, Dadonaite et al., 2025, Green et al., 2025, Sülzen et al., 2025, Pearce et al., 2025, Biller et al., 2025, Laub et al., 2025, Johnson et al., 2025, Na et al., 2025, Tozzi, 2025]. Biomedical text introduces denser punctu-

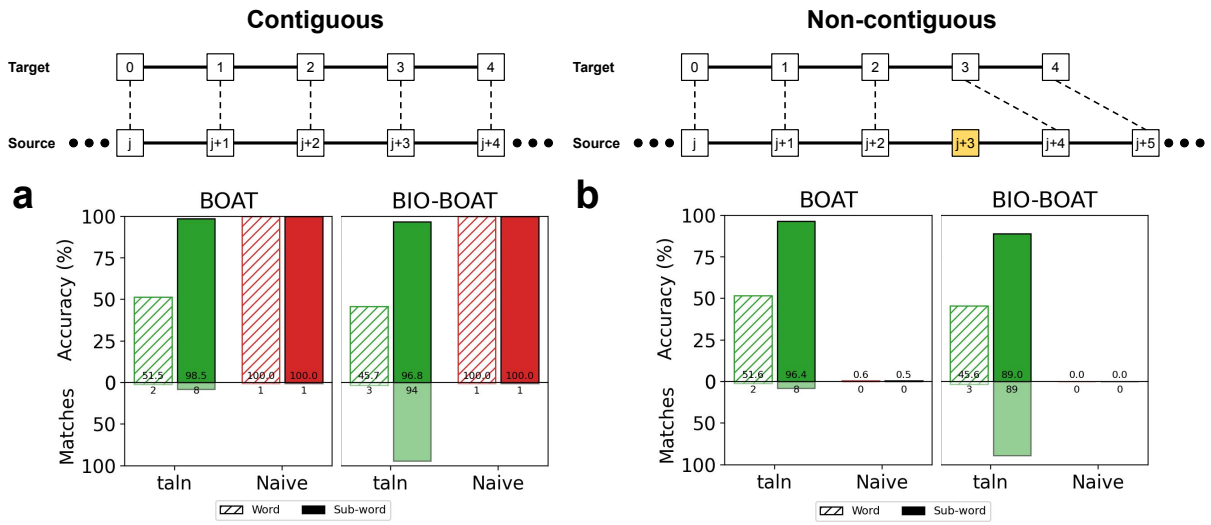

Figure 2: Contiguous and non-contiguous alignment. a. Contiguous alignment accuracy for all samples in the `BOAT` (left) and `BIO-BOAT` (right) datasets for `taln` (green) vs. Naive alignment (red) tokenized at the word (striped) and subword (solid) level as well as the average number of alignment matches for each method (below). **b.** Non-contiguous alignment accuracy and enumeration for `taln` and Naive alignment on the non-contiguous `BOAT` (left) and `BIO-BOAT` (right) datasets.

ation, symbols, and alphanumeric identifiers, making it a stringent test for alignment methods and tokenizers (which are often built from general-domain texts.)

2.2 Alignment algorithms

Many alignment methods exist for verifying extracted text, from exact string matching to algorithms that allow gaps or reorderings. To compare them, we formalized alignment as a map from target token positions to source token positions (Supplementary Note).

Under this formalization, alignment methods form a hierarchy defined by the constraints they impose on this map. Contiguous alignment requires adjacent tokens to match. Ordered alignment drops adjacency but retains order. Permutation alignment drops order but requires a one-to-one correspondence. General rearrangement imposes no constraints. Each relaxation admits more valid alignments, but the number of possible alignments grows rapidly, from linear (contiguous) to factorial (permutation) to exponential (rearrangement).

Ordered alignment sits at a practical point in this hierarchy: it recovers non-contiguous phrases while keeping the number of alignments manageable. In practice, relaxing constraints beyond ordered alignment produces spurious alignments, making permutation and rearrangement impractical. We therefore focus our study on ordered alignment and implement it in a tool called `taln`, short for **T**oken **Al**igner. `taln` is a Python library and command-line tool that returns all order-preserving alignments between a target and source. `taln` is tokenizer-agnostic and supports various tokenization strategies.

2.3 Text Tokenization

Having selected an alignment algorithm, we next asked how tokenization affects alignment accuracy. We compared naive phrase matching, word-level tokenization (whitespace), and subword tokenization using the `cl100k_base` tokenizer (Methods). We benchmarked `taln`, longest common subsequence (LCS), and Python's `difflib.SequenceMatcher` on both contiguous and non-contiguous variants of `BOAT` and `BIO-BOAT` (Figure 2).

2.3.1 Contiguous text

On contiguous phrases from `BOAT`, naive alignment recovered all targets, as expected. In contrast, under word-level tokenization, nearly half of targets were missed by all ordered-alignment methods despite being fully contiguous (47–58%, depending on the tool; Figure 2a and Supplementary Figure S1a, striped).

These failures arose from inconsistencies introduced by whitespace tokenization, where punctuation or quotation marks are inconsistently attached to neighboring words, preventing correct token-level alignment. For example, punctuation like quotation marks and commas are often omitted from the beginning or end of `SQuAD` answers (e.g. `a lawless place'' or ``a place under no jurisdiction`) even when present in the passage (`...of ``a lawless place'' or ``a place under no jurisdiction''...`). When split by whitespace, the missing adjacent quotes prevent correct alignment, causing otherwise contiguous phrases to be missed.

To address these missing alignments, we repeated the evaluation using subword tokenization (Methods). Subword tokenization largely resolves this issue. By

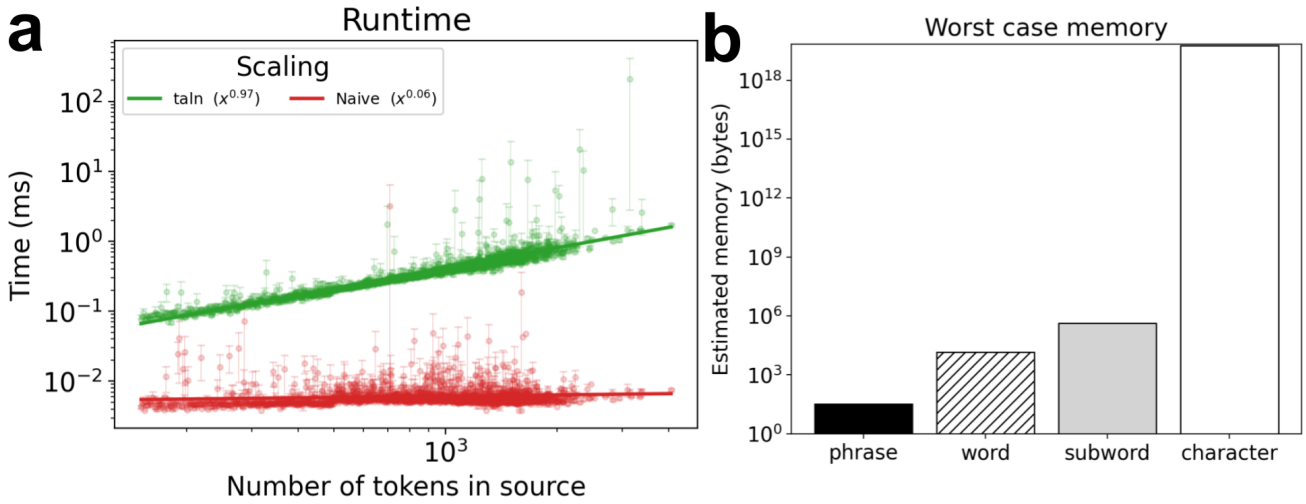

Figure 3: Runtime & Memory: **a.** Runtime of `taln` vs. naive phrase alignment as a function of source length. The inset power law shows the runtime scaling as a function of the number of tokens in the source. **b.** Estimated worst-case memory requirements for phrase (black), word (striped), subword (gray), and character (white) tokenization.

splitting words into smaller fragments, subword tokenization separates punctuation and symbols from meaningful text—e.g., (`of ``a lawless`) splits into (`of, ``, a, law, less`). As a result, alignment accuracy increased to 98–99% across all methods (Figure 2a and Supplementary Figure S1a, solid), demonstrating that tokenization—not the alignment algorithm alone—also contributes to alignment accuracy.

2.3.2 Non-contiguous text

To evaluate alignment performance on non-contiguous phrases, we used the non-contiguous `BOAT` dataset. In this setting, naive alignment failed almost entirely (0.6%). Under word-level tokenization, ordered alignment recovered only about half of targets (51–52%; Figure 2b and Supplementary Figure S1c, striped).

Failures arose from the same issue observed in the contiguous setting: punctuation, symbols, or alphanumeric patterns are inconsistently segmented at the word level, making ordered alignment difficult once phrases are tokenized.

Subword tokenization dramatically improved performance. With subword tokens, ordered alignment recovered over 96% of non-contiguous targets across all tools, including `taln`, LCS, and `difflib` (Figure 2b and Supplementary Figure S1c, solid). This shows that allowing gaps is necessary but not sufficient; alignment of non-contiguous phrases depends on the granularity of subwords defined by the tokenization strategy.

2.3.3 Generalization to biomedical text

To test whether these results generalize to biomedical text, we repeated the evaluation on `BIO-BOAT` (Figure 2a,b, right; Supplementary Figure S1b,d).

The same alignment patterns hold. For contiguous phrases, word-level tokenization missed roughly half of targets, while subword tokenization recovered over 97%. For non-contiguous phrases, word-level tokenization recovered only 45–46% of targets, whereas subword tokenization improved accuracy to approximately 89% (Figure 2).

The slightly reduced alignment on `BIO-BOAT` likely reflects the complexity of biomedical vocabulary, where gene names and chemical identifiers are often split inconsistently and therefore cannot be reliably captured by a tokenizer trained on general-domain text.

2.4 Alignment enumeration

Unlike LCS and `difflib`, `taln` enumerates all valid alignments rather than returning a single match. This matters when extracted phrases appear multiple times in a passage—common in scientific text where repeated entities (gene names, conditions, acronyms) may indicate importance [Swayamdipta et al., 2018, Zhang et al., 2023, Moon et al., 2023, Wang et al., 2025].

On `BOAT`, `taln` returned an average of 8 alignments per contiguous target and 7 per non-contiguous target (Table 1), compared to a single alignment produced by standard tools.

On `BIO-BOAT`, `taln` returned an average of 95 alignments per contiguous target and 89 per non-contiguous target. This increase likley reflects finer token resolution caused by a mismatch between biomedical text and general-domain tokenizers. For example, phrases such as "`secrete the anti-angiogenic isoforms VEGF-A165b`" are split into multiple tokens; the gene "`VEGF-A165b`" alone is decomposed into five

tokens, which combinatorially increases the number of potential alignments when using `taln`.

2.5 Runtime & Memory

Finally, we evaluated runtime and worst-case memory requirements (Figure 3). While naive alignment is fastest, all ordered alignment methods operate in the millisecond range for typical passage lengths (∼1,000 tokens). `taln`, LCS, and `difflib` scale approximately linearly with source length, remaining practical for standard short-passage alignments (Supplementary Figure S3).

Memory usage, however, varies sharply with the granularity of tokenization. Character-level tokenization incurs orders-of-magnitude higher memory costs, making it impractical despite its flexibility. Subword tokenization offers a practical compromise, retaining most of the alignment flexibility of character-level methods while remaining computationally feasible.

3 Discussion

LLMs are powerful tools for data mining. They can ingest massive amounts of text and rapidly synthesize, extract, and return specific pieces of information. But unlike other data-mining tasks, such as summarization, text extraction has a strict constraint on the generated text: the output must already exist in the input. Humans can often verify that this constraint is not violated, but doing so at scale becomes prohibitive. This is where text-alignment algorithms can assist human evaluators by providing automated, verifiable checks. Our results show that this verification constraint can be satisfied deterministically using token-based alignment, without requiring semantic judgment by an LLM.

Alignment algorithms are well suited to this verification task, but traditional text alignment algorithms and tokenization strategies struggle with scientific text—despite their long history and development in disparate fields [Hirschberg, 1975, Knuth et al., 1977, Melsted et al., 2021, Sennrich et al., 2016, Sullivan et al., 2024]. Extracted phrases are often interrupted by symbols, qualifiers, or parenthetical statements. Exact phrase matching therefore misses many valid alignments. Alignment methods that allow gaps or skips between aligned segments offer a practical solution, but they introduce a new question: what constitutes a meaningful "part" of a phrase?

This work shows that subword tokenization—used by many modern LLMs for training and inference—provides a natural answer. When combined with ordered alignment methods such as `taln`, it enables reliable alignment even when their contiguity is unknown. Subword tokenization improves accuracy in both contiguous and non-contiguous settings, and `taln`, by enumerating all valid alignments, recovers matches that standard tools miss, while remaining fast and deterministic compared to LLM-based verification approaches.

However, enumerating many valid alignments introduces a new challenge: selecting the most meaningful one. As recall improves, specificity becomes the limiting factor. Techniques for selecting among numerous valid alignments are a promising direction to improve the specificity of `taln`, for example using approaches based on token embedding radius (or related selection criteria, Supplementary Note.) While not discussed here, methods for verifying the existence of LLM-extracted phrases with semantic (instead of exact) matching are also a promising direction.

The selection challenge is amplified by domain-specific differences. The lower accuracy, and greater alignments, on `BIO-BOAT` compared to `BOAT` likely reflects more complex tokenization boundaries in biomedical text—gene names, chemical formulas, and mixed alphanumeric identifiers are common and often split inconsistently even by subword tokenizers, suggesting the need for domain-specific tokenizers to further improve alignment specificity.

Together, these results demonstrate that alignment algorithms and tokenization matter for verifying LLM-extracted text, and how methods from computational biology and linguistics can be combined to develop useful verification tools. Practically speaking, we envision non-contiguous alignment algorithms used with subword tokenization in text-extraction workflows where refined text alignment can be performed: passages can be extracted contiguously from large documents, and relevant—possibly non-contiguous—text can then be aligned to those passages, augmenting human ability to parse documents and ground results at scale.

A Impact Statement

Large language models are increasingly used to extract structured information from text in scientific research, medicine, and other data-intensive fields. This use is attractive because models can process volumes of text far beyond human capacity. However, extraction is only useful if the returned text actually appears in the source—i.e. it is not hallucinated. In high-stakes settings such as clinical decision support or scientific data curation, failures of verification can propagate errors and undermine trust.

This work addresses verification by improving how extracted text is aligned back to its source. We show that token-based alignment is most effective when it uses subword tokenization, which matches the representation used by LLMs, rather than whitespace- or character-based tokenization. On our benchmarks, this choice recovers about 50% more alignments than word-level while being more memory efficient than character level. By providing deterministic and transparent alignment algorithms, this work supports more trustworthy use of LLM-based text extraction in scientific and clinical workflows, and reduces reliance on probabilistic or fragile verification methods.

B Methods

B.1 `BOAT` Building

We derived `BOAT` from the Stanford Question Answering Dataset (`SQuAD`) v2.0. Each `SQuAD` example provides a passage (source), an answer string (target), and an answer start index (`answer_start`) locating the target as a contiguous character sequence in the source.

We first flattened `SQuAD` into records with fields `source`, `target`, and `idx_start` (plus metadata such as title and question identifiers). We retained records where naive contiguous string match occurs: `source[idx_start:idx_start+len(target)]` = `target`. We then deduplicate on `(source, target, idx_start)`.

To make character indices consistent with subword tokenization, we applied a "shift" normalization: if the character immediately preceding `idx_start` is a space, we prepend a space to the target and decrement `idx_start` by one. This ensures that leading whitespace, when present in the source, is part of the target phrase; a feature that is often seen in substring tokenization vocabularies. Note this does not affect whitespace tokenization since the space is prepended to the string.

Because identical `(source, target)` pairs can occur with multiple valid start indices (e.g., repeated mentions in a passage), we grouped by `(source, target)` and stored `idx_start` as a set of start positions. In downstream analyses we restricted to multi-word targets to focus on phrases where ablating an interior token introduces a meaningful non-contiguity.

We constructed two `BOAT` variants: Contiguous `BOAT` and Non-contiguous `BOAT`. Contiguous `BOAT` consists of the original `(source, target)` pairs and their `idx_start` sets. Non-contiguous `BOAT` was constructed by removing, for each `(source, target)` pair, one interior token of the target to create a non-contiguous variant. In the benchmarking notebook this ablation removes one interior whitespace-delimited word. In the dataset construction notebook we also generate a subword-token ablation (removing one interior subword token).

B.1.1 `BIO-BOAT` Building

We downloaded 51 papers from the bioRxiv AWS bucket `s3://biorxiv-src-monthly/Current_content/`, dated January 2025 and licensed under CC BY 4.0. We converted the manuscript XML to Markdown and ran `taln extract` to generate source–target pairs, where each target is a contiguous span of the source. We then generated non-contiguous `BIO-BOAT` in the same manner as `BOAT` described above.

B.2 Text Tokenization

We evaluated alignment under two tokenization schemes: word (whitespace) tokens and subword tokens. Word tokens are maximal non-whitespace runs found by regex matching. Each token retains its character start and end offsets in the source string. Subword tokens are produced by the OpenAI `cl100k_base` tokenizer (via `tiktoken`). We decode tokens with offsets to recover character start and end positions for each token.

Before tokenization, source and target text are normalized by applying Unicode normalization and ASCII transliteration (`unidecode`), mapping several common punctuation/symbol variants to ASCII, and collapsing whitespace and newlines. Because token offsets are defined on the normalized string, we exclude examples where normalization changes string length.

B.3 Alignment Algorithms

Given a source sequence of tokens and a target sequence of tokens, we evaluate both contiguous and non-contiguous alignment methods: naive phrase matching, longest common subsequence (LCS), `difflib`, and `taln`.

Naive matching checks whether `target` occurs as an exact substring of `source` and returns all matches. LCS computes an order-preserving alignment using dynamic programming and returns the matched token positions (whitespace or subword). `difflib` uses Python's `difflib.SequenceMatcher` over tokens (whitespace or subword) to identify and return matching tokens.

`taln` builds an index of token n-grams from the source (with $k = 1$ in this work), aligns each target token to all occurrences of that token in the source, and then enumerates all order-preserving alignments. Each alignment is returned as a list of token objects with character offsets; multiple valid alignments are enumerated explicitly.

When benchmarking alignment, we report two complementary metrics: reconstruction accuracy (Figure S1) and localization accuracy (Figure S2). Reconstruction asks whether a method can recover the target string, independent of where it came from in the source. Concretely, reconstruction succeeds if at least one returned alignment reconstructs the target exactly (after applying the same normalization used for tokenization). Localization asks whether a method recovers the ground-truth target (positionally). For each fully reconstructed alignment, we take its bounding position ($\texttt{start_idx}, \texttt{end_idx}$). Localization succeeds if any reconstruction matches the positions of the target (s, $s + |\texttt{target}|$) for some $s \in \texttt{idx_start}$.

For non-contiguous alignments, localization accuracy is defined by the bounding character indices: an alignment is correct if the left and right character boundaries ($\texttt{start_idx}, \texttt{end_idx}$) of at least one fully reconstructed alignment match the expected target.

B.4 Runtime Results

We measure per-sample runtime for each alignment method by wall-clock timing around the alignment call on each example, reporting elapsed time in milliseconds. Runtime and accuracy are summarized as a function of source length (measured in tokens) to illustrate scaling behavior with input size.

C Data and Code Availability

The `SQuAD` datasets can be found here https://rajpurkar.github.io/SQuAD-explorer/. The `taln` tool and the `BOAT` and `BIO-BOAT` datasets can be found here: https://github.com/sbooeshaghi/taln.

D Acknowledgments

We thank Lior Pachter and Gennady Gorin for helpful feedback on the presentation of the main text as well as for comments on the Supplementary Note. We thank Nithya Appannagaari for her work—on verifying the existence of extracted marker genes with A.S.B.—that partly inspired the creation of `taln`. We thank Hanna Krasowski, Lauren Malek, and Sewon Min for helpful discussions and Hanna Krasowski for contributing improvements to `taln`. We also thank the Howard Hughes Medical Institute for supporting A.S.B. through the Hanna H. Gray Fellows program.

E Author Contributions

A.S.B. conceived the project, developed the benchmarks, implemented the `taln` software and analysis code, and wrote the manuscript. A.S. reviewed the manuscript.

F Disclosures

All em dashes were placed, in the text, by the author's own hands. Generative language models were used by the authors to check for grammatical errors in the text, fix bugs in code, and debug LaTeX issues.

G Supplementary Figures

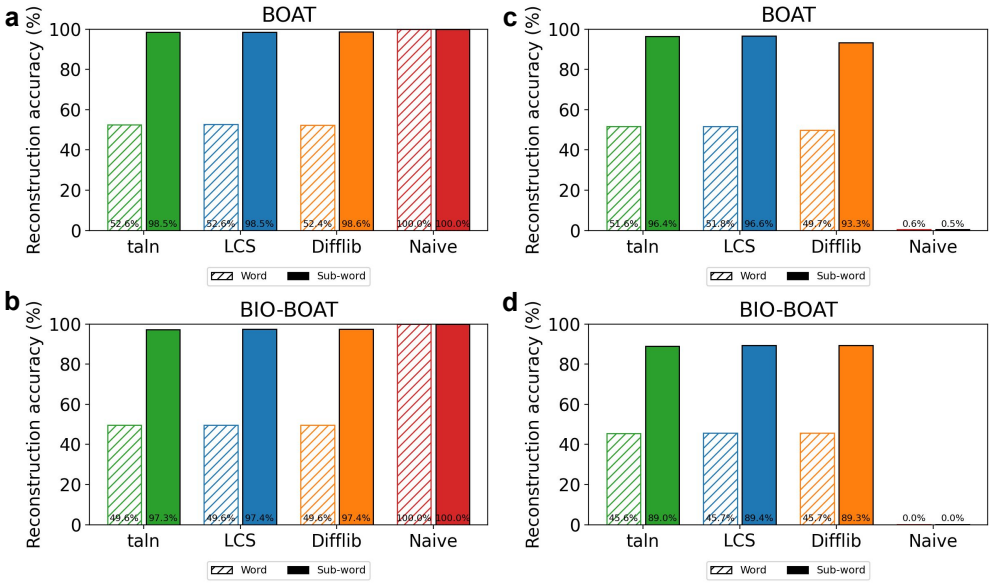

Figure S1: Reconstruction accuracy for contiguous and non-contiguous alignment. Contiguous alignment reconstruction accuracy for all samples in the `BOAT` (a) and `BIO-BOAT` (b) dataset for `taln` (green), LCS (blue), Difflib (orange), and Naive alignment (red) tokenized at the word (striped) and subword (solid). Non-contiguous alignment reconstruction accuracy all tools on the non-contiguous `BOAT` (c) and `BIO-BOAT` dataset (d).

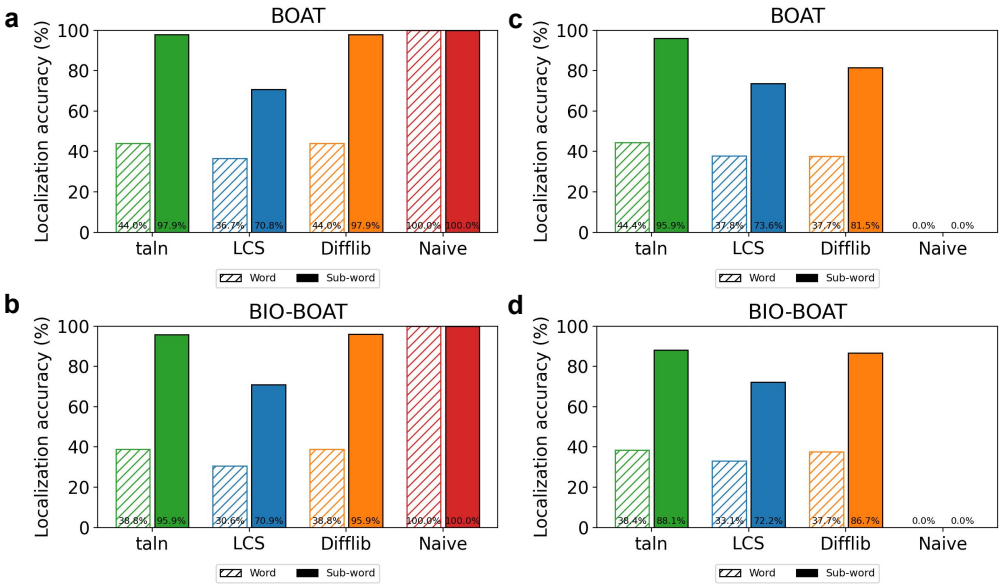

Figure S2: Localization accuracy for contiguous and non-contiguous alignment. Contiguous alignment localization accuracy for all samples in the `BOAT` (a) and `BIO-BOAT` (b) dataset for `taln` (green), LCS (blue), Difflib (orange), and Naive alignment (red) tokenized at the word (striped) and subword (solid). Non-contiguous alignment localization accuracy all tools on the non-contiguous `BOAT` (c) and `BIO-BOAT` dataset (d).

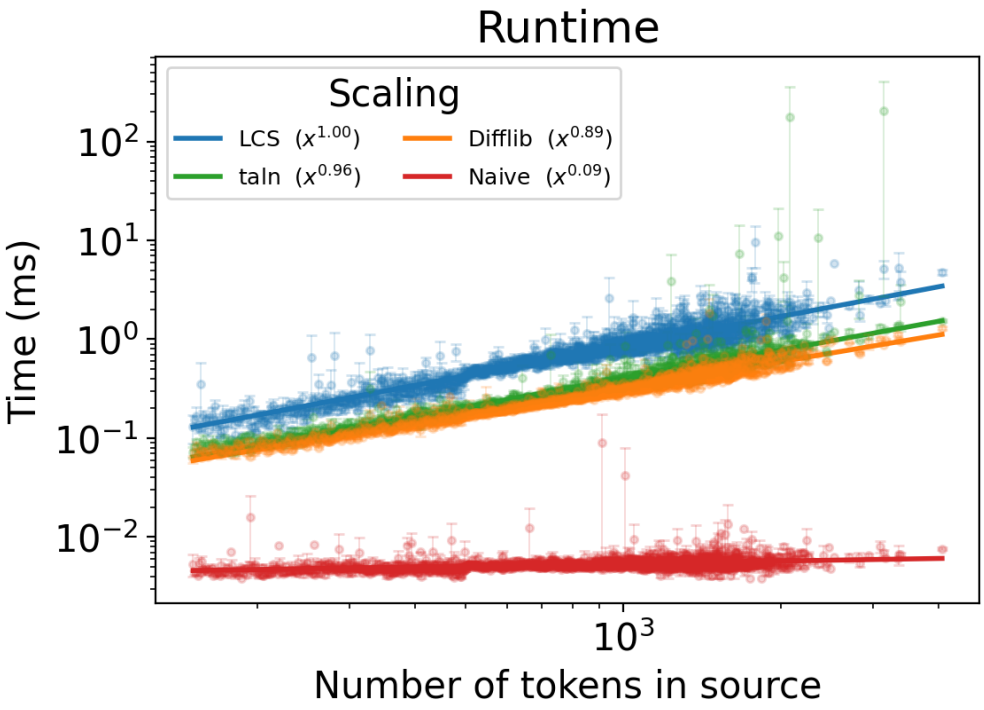

Figure S3: Runtime of various text alignment tools. Log–log plot of runtime (ms) versus number of tokens in the source for four alignment methods. Points show individual source-target pairs; lines show power-law fits with exponents indicated in the legend. LCS and `taln` scale roughly linearly, Difflib is slightly sublinear, and the naive baseline is nearly constant.

References

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1), 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00784-w. URL http://dx.doi.org/10.1038/s41524-022-00784-w.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, 2022. URL https://arxiv.org/abs/2212.05238.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022. URL https://arxiv.org/abs/2205.12689.

David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.bionlp-1.37. URL http://dx.doi.org/10.18653/v1/2023.bionlp-1.37.

Allen H. Huang, Hui Wang, and Yi Yang. ¡scp¿finbert¡/scp¿: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2):806–841, 2023. ISSN 1911-3846. doi: 10.1111/1911-3846.12832. URL http://dx.doi.org/10.1111/1911-3846.12832.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1), 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45563-x. URL http://dx.doi.org/10.1038/s41467-024-45563-x.

Wei Zhang, Qinggong Wang, Xiangtai Kong, Jiacheng Xiong, Shengkun Ni, Duanhua Cao, Buying Niu, Mingan Chen, Yameng Li, Runze Zhang, Yitian Wang, Lehan Zhang, Xutong Li, Zhaoping Xiong, Qian Shi, Ziming Huang, Zunyun Fu, and Mingyue Zheng. Fine-tuning large language models for chemical text mining. *Chemical Science*, 15(27): 10600–10611, 2024. ISSN 2041-6539. doi: 10.1039/d4sc00924j. URL http://dx.doi.org/10.1039/D4SC00924J.

Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1), 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45914-8. URL http://dx.doi.org/10.1038/s41467-024-45914-8.

Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024. URL https://arxiv.org/abs/2404.02330.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. Lmdx: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15140–15168. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.899. URL http://dx.doi.org/10.18653/v1/2024.findings-acl.899.

Mert Marcel Dagli, Yohannes Ghenbot, Hasan S. Ahmad, Daksh Chauhan, Ryan Turlip, Patrick Wang, William C. Welch, Ali K. Ozturk, and Jang W Yoon. Development and validation of a novel ai framework using nlp with llm integration for relevant clinical data extraction through automated chart review. *Scientific Reports*, 14(1), 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-77535-y. URL http://dx.doi.org/10.1038/s41598-024-77535-y.

Bastien Le Guellec, Alexandre Lefèvre, Charlotte Geay, Lucas Shorten, Cyril Bruge, Lotfi Hacein-Bey, Philippe Amouyel, Jean-Pierre Pruvo, Gregory Kuchcinski, and Aghiles Hamroun. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiology: Artificial Intelligence*, 6(4), 2024. ISSN 2638-6100. doi: 10.1148/ryai.230364. URL http://dx.doi.org/10.1148/ryai.230364.

Xiangfeng Liu, Jianting Sun, Aidi Lei, and Junzhi Zhu. Research and applications of large language models for converting unstructured data into structured data. In *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, pages 305–308. IEEE, 2024. doi: 10.1109/cbase64041.2024.10824634. URL http://dx.doi.org/10.1109/CBASE64041.2024.10824634.

Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, 82:102742, 2024. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2024.102742. URL http://dx.doi.org/10.1016/j.ecoinf.2024.102742.

Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, Amanda Konet, and Robert Chew. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*, 15(4): 576–589, 2024. ISSN 1759-2887. doi: 10.1002/jrsm.1710. URL http://dx.doi.org/10.1002/jrsm.1710.

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. A survey on open information extraction from rule-based model to large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.560. URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.560.

Vasileios Ntinopoulos, Hector Rodriguez Cetina Biefer, Igor Tudorache, Nestoras Papadopoulos, Dragan Odavic, Petar Risteski, Achim Haeussler, and Omer Dzemali. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health & Care Informatics*, 32(1):e101139, 2025. ISSN 2632-1009. doi: 10.1136/bmjhci-2024-101139. URL http://dx.doi.org/10.1136/bmjhci-2024-101139.

Yiyi Tang, Ziyan Xiao, Xue Li, Qiwen Fang, Qingpeng Zhang, Daniel Yee Tak Fong, Francisco Tsz Tsun Lai, Celine Sze Ling Chui, Esther Wai Yin Chan, and Ian Chi Kei Wong. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: A comparative study of gpt-3.5 and gpt-4. *medRxiv*, 2024. doi: 10.1101/2024.03.20.24304572. URL http://dx.doi.org/10.1101/2024.03.20.24304572.

Ruben Geevarghese, Carlie Sigel, John Cadley, Subrata Chatterjee, Pulkit Jain, Alex Hollingsworth, Avijit Chatterjee, Nathaniel Swinburne, Khawaja Hasan Bilal, and Brett Marinelli. Extraction and classification of structured data from unstructured hepatobiliary pathology reports using large language models: a feasibility study compared with rules-based natural language processing. *Journal of Clinical Pathology*, 78(2):135–138, 2024. ISSN 1472-4146. doi: 10.1136/jcp-2024-209669. URL http://dx.doi.org/10.1136/jcp-2024-209669.

Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 54(3):1125–1150, 2025. ISSN 1460-4744. doi: 10.1039/d4cs00913d. URL http://dx.doi.org/10.1039/D4CS00913D.

William S. Azar, Dylan M. Junkin, Charles Hesswani, Christopher R. Koller, Sahil H. Parikh, Kyle C. Schuppe, Nicholas Williams, Daniel Nethala, Neil Mendhiratta, Alexander P. Kenigsberg, Baris Turkbey, Maria J. Merino, George Zaki, Janelle Cortner, Sandeep Gurram, and Peter A. Pinto. Llm-mediated data extraction from patient records after radical prostatectomy. *NEJM AI*, 2(6), 2025. ISSN 2836-9386. doi: 10.1056/aics2400943. URL http://dx.doi.org/10.1056/AIcs2400943.

Nicole Schwitter. Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences. *Methodological Innovations*, 18(1):61–65, 2025. ISSN 2059-7991. doi: 10.1177/20597991251313876. URL http://dx.doi.org/10.1177/20597991251313876.

Mingyuan Qin, Lei Feng, Jing Lu, Ziyan Sun, Zhengyu Yu, and Lianyi Han. Zerotunebio ner: A three-stage framework for zero-shot and zero-tuning biomedical entity extraction using large language models and prompt engineering. *Computer Methods and Programs in Biomedicine*, 272:109070, 2025. ISSN 0169-2607. doi: 10.1016/j.cmpb.2025.109070. URL http://dx.doi.org/10.1016/j.cmpb.2025.109070.

Yosuke Yamagishi, Yuta Nakamura, Shouhei Hanaoka, and Osamu Abe. Large language model approach for zero-shot information extraction and clustering of japanese radiology reports: Algorithm development and validation. *JMIR Cancer*, 11:e57275–e57275, 2025. ISSN 2369-1999. doi: 10.2196/57275. URL http://dx.doi.org/10.2196/57275.

Roomani Srivastava, Lipika Bhat, Suraj Prasad, Sarvesh Deshpande, Barnali Das, and Kshitij Jadhav. Medpromptextract (medical data extraction tool): Anonymization and high-fidelity automated data extraction using natural language processing and prompt engineering. *The Journal of Applied Laboratory Medicine*, 10(4):793–805, 2025. ISSN 2475-7241. doi: 10.1093/jalm/jfaf034. URL http://dx.doi.org/10.1093/jalm/jfaf034.

Xue Jiang, Weiren Wang, Shaohan Tian, Hao Wang, Turab Lookman, and Yanjing Su. Applications of natural language processing and large language models in materials discovery. *npj Computational Materials*, 11(1), 2025. ISSN 2057-3960. doi: 10.1038/s41524-025-01554-0. URL http://dx.doi.org/10.1038/s41524-025-01554-0.

Yunqiao Fei, Jingchao Fan, and Guomin Zhou. Extracting fruit disease knowledge from research papers based on large language models and prompt engineering. *Applied Sciences*, 15(2):628, 2025. ISSN 2076-3417. doi: 10.3390/app15020628. URL http://dx.doi.org/10.3390/app15020628.

Chenyue Zhou, Gürkan Solmaz, Flavio Cirillo, Kiril Gashteovski, and Jonathan Fürst. Textmine: Data, evaluation framework and ontology-guided llm pipeline for humanitarian mine action, 2025. URL https://arxiv.org/abs/2509.15098.

Hammaad Adam, Junjing Lin, Jianchang Lin, Hillary Keenan, Ashia Wilson, and Marzyeh Ghassemi. Clinical information extraction with large language models: A case study on organ procurement. *AMIA Annual Symposium Proceedings*, 2024:115–123, 2024. ISSN 1942-597X. URL https://pubmed.ncbi.nlm.nih.gov/40417525/. Epub ahead of print. Published online 2025 May 22.

Jake Lever, Russ Altman, and Jin-Dong Kim. Extending textae for annotation of non-contiguous entities. *Genomics & Informatics*, 18(2):e15, 2020. ISSN 2234-0742. doi: 10.5808/gi.2020.18.2.e15. URL http://dx.doi.org/10.5808/GI.2020.18.2.e15.

Emiko Shinohara, Kiminori Shimamoto, and Yoshimasa Kawazoe. Uncovering condition information loss in medical text extraction: The challenge of non-contiguous spans. *Next Research*, 1(2):100044, 2024. ISSN 3050-4759. doi: 10.1016/j.nexres.2024.100044. URL http://dx.doi.org/10.1016/j.nexres.2024.100044.

Andrew D. Hildreth, Feiyang Ma, Yung Yu Wong, Ryan Sun, Matteo Pellegrini, and Timothy E. O'Sullivan. Single-cell sequencing of human white adipose tissue identifies new cell states in health and obesity. *Nature Immunology*, 22(5):639–653, 2021. ISSN 1529-2916. doi: 10.1038/s41590-021-00922-4. URL http://dx.doi.org/10.1038/s41590-021-00922-4.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Philip Whittington, Gregor Bachmann, and Tiago Pimentel. Tokenisation is np-complete, 2024. URL https://arxiv.org/abs/2412.15210.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. URL https://arxiv.org/abs/1806.03822.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016. URL https://arxiv.org/abs/1508.07909.

Marijn van Vliet, Oona Rinkinen, Takao Shimizu, Anni-Mari Niskanen, Barry Devereux, and Riitta Salmelin. Convolutional networks can model the functional modulation of the MEG responses associated with feed-forward processes during visual word recognition. *bioRxiv*, 2022. doi: 10.1101/2022.02.08.479654. URL http://dx.doi.org/10.1101/2022.02.08.479654.

Sergio Cobo-López, Lucas J. Carbajal, Matthew Witt, Forest L. Rohwer, and Antoni Luque. Emergence of tipping points and transient dynamics in finite observations. *bioRxiv*, 2023. doi: 10.1101/2023.12.27.573307. URL http://dx.doi.org/10.1101/2023.12.27.573307.

Daria Meyer, Winfried Goettsch, Jannes Spannenberg, Bettina Stieber, Sebastian Krautwurst, Martin Hölzer, Christian Brandt, Jörg Linde, Christian Höner zu Siederdissen, Akash Srivastava, Milena Zarkovic, Damian Wollny, and Manja Marz. Unlocking the Full Potential of Nanopore Sequencing: Tips, Tricks, and Advanced Data Analysis Techniques. *bioRxiv*, 2023. doi: 10.1101/2023.12.06.570356. URL http://dx.doi.org/10.1101/2023.12.06.570356.

Vitaly Zimyanin, Magdalena Magaj, Nadia Ingabire Manzi, Che-Hang Yu, Theresa Gibney, Yu-Zen Chen, Mustafa Basaran, Xavier Horton, Karsten Siller, Ariel Pani, Daniel Needleman, Daniel J. Dickinson, and Stefanie Redemann. Chromokinesin Klp-19 regulates microtubule overlap and dynamics during anaphase in C. elegans. *bioRxiv*, 2023. doi: 10.1101/2023.10.26.564275. URL http://dx.doi.org/10.1101/2023.10.26.564275.

Ruth A. Singer, Veronika Rajchin, Kwanghoon Park, Nathaniel Heintz, and Robert B. Darnell. Opto-CLIP reveals dynamic FMRP regulation of mRNAs upon CA1 neuronal activation. *bioRxiv*, 2024. doi: 10.1101/2024.08.13.607210. URL http://dx.doi.org/10.1101/2024.08.13.607210.

Mingzhe Shen, Daniel Kortzak, Simon Ambrozak, Shubham Bhatnagar, Ian Buchanan, Ruibin Liu, and Jana Shen. KaMLs for Predicting Protein pKa Values and Ionization States: Are Trees All You Need? *bioRxiv*, 2024. doi: 10.1101/2024.11.09.622800. URL http://dx.doi.org/10.1101/2024.11.09.622800.

Dmitry Frolov, Ilya Ibraev, and Igor Sedov. Quantum chemical properties of chlorinated polycyclic aromatic hydrocarbons for delta machine learning. *bioRxiv*, 2024. doi: 10.1101/2024.12.19.628874. URL http://dx.doi.org/10.1101/2024.12.19.628874.

Annie K. Ryan, Brooke I. Asemota, Tyler Heisler-Taylor, Claire Mello, Luis Rodriguez, William E. Sponsel, Julie Racine, Tonia S. Rex, Randolph D. Glickman, and Matthew A. Reilly. Torsion-Induced Traumatic Optic Neuropathy (TITON): A Physiologically Relevant Animal Model of Traumatic Optic Neuropathy. *bioRxiv*, 2024. doi: 10.1101/2024.10.08.617086. URL http://dx.doi.org/10.1101/2024.10.08.617086.

Hans E. Pfalzgraf, Aditya G. Rao, Kakali Sen, Hannah R. Adams, Marcus Edwards, You Lu, Chin Yong, Sofia Jaho, Takehiko Tosha, Hiroshi Sugimoto, Sam Horrell, James Beilsten-Edmands, Robin L. Owen, Colin R. Andrew, Jonathan A. R. Worrall, Ivo Tews, Adrian J. Mulholland, Michael A. Hough, and Thomas W. Keal. Double crossed? Structural and computational studies of an unusual crosslinked heme in Methylococcus capsulatus cytochrome P460. *bioRxiv*, 2024. doi: 10.1101/2024.12.21.628999. URL http://dx.doi.org/10.1101/2024.12.21.628999.

Noriyo Onishi, Nicodemo Mazzaferro, Špela Kunstelj, Daisy A. Alvarado, Anna M. Muller, and Frank X. Vázquez. Molecular Dynamics Study of α-Synuclein Domain Deletion Mutant Monomers. *bioRxiv*, 2024. doi: 10.1101/2024.03.23.586267. URL http://dx.doi.org/10.1101/2024.03.23.586267.

Yirga Kebede Wondim, Ayalew Wondie Melese, and Workiyie Worie Assefa. Mapping of Shore Area Wetlands in Lake Tana Biosphere Reserve, Ethiopia Using Sentinel-1A SAR and Multi-Source Data. *bioRxiv*, 2024. doi: 10.1101/2024.12.30.630805. URL http://dx.doi.org/10.1101/2024.12.30.630805.

Slaven Crnkovic, Helene Thekkekara Puthenparampil, Shirin Mulch, Valentina Biasin, Nemanja Radic, Jochen Wilhelm, Marek Bartkuhn, Ehsan Bonyadi Rad, Alicja Wawrzen, Ingrid Matzer, Ankita Mitra, Ryan D. Leib, Bence Miklos Nagy, Anita Sahu-Osen, Francesco Valzano, Natalie Bordag, Matthias Evermann, Konrad Hoetzenecker, Andrea Olschewski, Senka Ljubojevic-Holzer, Malgorzata Wygrecka, Kurt Stenmark, Leigh M Marsh, Vinicio de Jesus Perez, and Grazyna Kwapiszewska. Adventitial fibroblasts direct smooth muscle cell-state transition in pulmonary vascular disease. *bioRxiv*, 2024. doi: 10.1101/2024.05.15.594343. URL http://dx.doi.org/10.1101/2024.05.15.594343.

Felix Kallenborn, Alejandro Chacon, Christian Hundt, Hassan Sirelkhatim, Kieran Didi, Sooyoung Cha, Christian Dallago, Milot Mirdita, Bertil Schmidt, and Martin Steinegger. GPU-accelerated homology search with MMseqs2. *bioRxiv*, 2024. doi: 10.1101/2024.11.13.623350. URL http://dx.doi.org/10.1101/2024.11.13.623350.

Nathan Peroutka-Bigus, Elizabeth J. Brooks, Michelle E. Ramey, Hope D'Erasmo, Jacqueline P. Ernest, Allison A. Bauman, Lisa K Woolhiser, Radojka M. Savic, Anne J. Lenaerts, Bree B. Aldridge, Jansy P. Sarathy, and Gregory T. Robertson. Contribution of front-line, standard of care drugs to bactericidal responses, resistance emergence, and cure in murine models of easy- or hard-to-treat tuberculosis disease. *bioRxiv*, 2024. doi: 10.1101/2024.12.18.629278. URL http://dx.doi.org/10.1101/2024.12.18.629278.

Tasfia Azim, Dru Myerscough, and Jonathan J. Silberg. Synonymous mutations in AAV Rep enhance genome packaging in a library selection. *bioRxiv*, 2024. doi: 10.1101/2024.10.08.617207. URL http://dx.doi.org/10.1101/2024.10.08.617207.

Sohni Ria Bhalla, Mussarat Wahid, Jason O Amartey, Amira Zawia, Federica Riu, Yizhuo Gao, Jyoti Agrawal, Amy P Lynch, Maria JC Machado, Tom Hawtrey, Ryosuke Kikuchi, Kathryn R Green, Lydia Teboul, Claire Allen, Zoe Blackley, Keerthana Rajaji, Daisy Marsden, Jennifer Batson, Steven J Harper, Sebastian Oltean, Winfried Amoaku, Andrew V Benest, Jonathan C Morris, Bruce Braithwaite, and David O Bates. The splicing factor kinase SRPK1 is a therapeutic target for Peripheral Vascular Disease. *bioRxiv*, 2024. doi: 10.1101/2024.04.17.589996. URL http://dx.doi.org/10.1101/2024.04.17.589996.

Anthony E. Jones, Amy Rios, Neira Ibrahimovic, Carolina Chavez, Nicholas A. Bayley, Andréa B. Ball, Wei Yuan Hsieh, Alessandro Sammarco, Amber R. Bianchi, Angel A. Cortez, Thomas G. Graeber, Alexander Hoffmann,

Steven J. Bensinger, and Ajit S. Divakaruni. The metabolic cofactor Coenzyme A enhances alternative macrophage activation via MyD88-linked signaling. *bioRxiv*, 2024. doi: 10.1101/2024.03.28.587096. URL http://dx.doi.org/10.1101/2024.03.28.587096.

Haiyi Fei, Xiaowen Lu, Zhan Shi, Xiu Liu, Cuiyu Yang, Xiaohong Zhu, Yuhan Lin, Ziqun Jiang, Jianmin Wang, Dong Huang, Liu Liu, Songying Zhang, and Lingling Jiang. Deciphering the Preeclampsia-Specific Immune Microenvironment: The Role of Pro-inflammatory Macrophages. *bioRxiv*, 2024. doi: 10.1101/2024.06.05.597577. URL http://dx.doi.org/10.1101/2024.06.05.597577.

Chantal Combredet and Thibaut Brunet. A fast and robust gene knockout method for Salpingoeca rosetta informs the genetics of choanoflagellate multicellular development. *bioRxiv*, 2024. doi: 10.1101/2024.07.13.603360. URL http://dx.doi.org/10.1101/2024.07.13.603360.

Peter Simmonds, Anamarija Butković, Joe Grove, Richard Mayne, Jonathon C. O. Mifsud, Martin Beer, Jens Bukh, J. Felix Drexler, Amit Kapoor, Volker Lohmann, Donald B. Smith, Jack T. Stapleton, Nikos Vasilakis, and Jens H. Kuhn. Integrated analysis of protein sequence and structure redefines viral diversity and the taxonomy of the Flaviviridae. *bioRxiv*, 2025. doi: 10.1101/2025.01.17.632993. URL http://dx.doi.org/10.1101/2025.01.17.632993.

Jason Adhinarta, Jizheng Dong, Tianxiao He, Junxiang Huang, Daniel Sprague, Jia Wan, Hyun Jee Lee, Zikai Yu, Peng Liu, Hang Lu, Eviatar Yemini, Saul Kato, Erdem Varol, and Donglai Wei. WormID-Bench: A Benchmark for Whole-Brain Activity Extraction in C. elegans. *bioRxiv*, 2025. doi: 10.1101/2025.01.06.631621. URL http://dx.doi.org/10.1101/2025.01.06.631621.

Karan Elangovan, Tamsin A. Spelman, Gilles Dupouy, Gaurav Singh, Marie-Edith Chaboute, and Henrik Jonsson. Machine learning segmentation tool trained on synthetic data for tracking cytoskeleton polymerisation and depolymerisation. *bioRxiv*, 2025. doi: 10.1101/2025.01.04.631322. URL http://dx.doi.org/10.1101/2025.01.04.631322.

Gechlang Tang, Alex V. Carr, Crystal Perez, Katherine Ramos Sarmiento, Lisa Levy, Johanna W. Lampe, Christian Diener, and Sean M. Gibbons. Metagenomic estimation of absolute bacterial biomass in the mammalian gut through host-derived read normalization. *bioRxiv*, 2025. doi: 10.1101/2025.01.07.631807. URL http://dx.doi.org/10.1101/2025.01.07.631807.

Douglas Maya-Miles, José García-Martínez, Ildefonso Cases, Rocío Pasión, Jesús de la Cruz, José E. Pérez-Ortín, María de la Cruz Muñoz-Centeno, and Sebastián Chávez. Regulation of transcription elongation anticipates alternative gene expression strategies across the cell cycle. *bioRxiv*, 2025. doi: 10.1101/2025.01.05.631386. URL http://dx.doi.org/10.1101/2025.01.05.631386.

Edert Abel Samame-Caramutti, Luis Salcedo-Valdez, and Silvia Suárez-Cunza. In vitro antioxidant and anti-inflammatory activities, total phenolic and flavonoid contents of Anadenanthera colubrina from Northern Peru. *bioRxiv*, 2025. doi: 10.1101/2025.01.24.633472. URL http://dx.doi.org/10.1101/2025.01.24.633472.

Hung Manh Nguyen, Beery Yaakov, Pedro Beca-Carretero, Gabriele Procaccini, Guannan Wang, Maheshi Dassanayake, Gidon Winters, and Simon Barak. Transcriptome responses of two Halophila stipulacea seagrass populations from pristine and impacted habitats, to single and combined thermal and excess nutrient stressors, reveal local adaptive features and core stress-response genes. *bioRxiv*, 2025. doi: 10.1101/2025.01.21.634158. URL http://dx.doi.org/10.1101/2025.01.21.634158.

Dylan J. Terstege and Jonathan R. Epp. Cognitive enrichment preserves retrosplenial parvalbumin density and cognitive function in female 5xFAD mice. *bioRxiv*, 2025. doi: 10.1101/2025.01.15.633249. URL http://dx.doi.org/10.1101/2025.01.15.633249.

Lumiao Chen, Lingao Dai, Shenghao Nie, Yue Zhao, Fanyuan Li, Xiujiao Deng, Chunlin Liu, Moran Li, Zewei Luo, Yuqing Liu, Weiwei Ge, Xiaofei Liu, Qian Yi, Yongkui Li, Pan Pan, and Xiaoping Wang. VZV gE Promotes STING Degradation via Midnolin-Proteasome Pathway to Inhibit cGAS-STING Signaling. *bioRxiv*, 2025. doi: 10.1101/2025.01.13.632693. URL http://dx.doi.org/10.1101/2025.01.13.632693.

Jie Li, Baojian Li, Xinbo Shi, Yuangui Yang, and Zhongxing Song. Optimization of Extraction Process by Response Surface Method-ology of Total Flavonoids from Ziziphus Jujube Flesh and its Sedative-Hypnotic Effects. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.632005. URL http://dx.doi.org/10.1101/2025.01.08.632005.

Evan A Kattner, Terrence R Stanford, and Emilio Salinas. Contributions of distinct attention mechanisms to saccadic choices in a gamified, dynamic environment. *bioRxiv*, 2025. doi: 10.1101/2025.01.25.634882. URL http://dx.doi.org/10.1101/2025.01.25.634882.

I S Shruti and S Vijay Prakash. Fitting multiple bacterial growth data using continued fraction of straight lines. *bioRxiv*, 2025. doi: 10.1101/2025.01.27.634991. URL http://dx.doi.org/10.1101/2025.01.27.634991.

Carmen Velázquez, Alejandro Arce-Rodríguez, Jon Altuna-Alvarez, Jessica Rojas-Palomino, Andony Flores-Ceron, Citlaly Cando-Narvaez, Adrián Ruiz, Javier De la Peña Noya, Amaia González-Magaña, María Queralt-Martín, Antonio Alcaraz, David Albesa-Jové, and Patricia Bernal. Tke5 is a novel Pseudomonas putida toxin that depolarises membranes killing plant pathogens. *bioRxiv*, 2025. doi: 10.1101/2025.01.11.632536. URL http://dx.doi.org/10.1101/2025.01.11.632536.

Valentí Rull. Where were the Caribbean mangroves during the Last Glacial Maximum? A microtopographical approach. *bioRxiv*, 2025. doi: 10.1101/2025.01.13.632686. URL http://dx.doi.org/10.1101/2025.01.13.632686.

Eva-Maria Dölker, Yasemin Cabuk, Tino Kühn, and Jens Haueisen. Comparison of TENS electrodes and textile electrodes for electrocutaneous warning. *bioRxiv*, 2025. doi: 10.1101/2025.01.15.633237. URL http://dx.doi.org/10.1101/2025.01.15.633237.

Adib Keikhosravi, Krishnendu Guin, Gianluca Pegoraro, and Tom Misteli. Simulation and quantitative analysis of spatial centromere distribution patterns. *bioRxiv*, 2025. doi: 10.1101/2025.01.22.634320. URL http://dx.doi.org/10.1101/2025.01.22.634320.

Germain Agazzi, Camille Carpentier, Olivia Bleeckx, and Frédérik De Laender. Network theory predicts ecosystem robustness across environmental conditions. *bioRxiv*, 2025. doi: 10.1101/2025.01.29.635264. URL http://dx.doi.org/10.1101/2025.01.29.635264.

Hélèna Choltus, Julien Prados, Niccolo Bianchi, Nelli Heikkila, Wolfram Karenovics, Veronique Serre Beinier, Nina Leitner, Benoit Bédat, Fedor Bezrukov, Beryl Mazel-Sanchez, Virginie Prendki, Alexander Lobrinus, Christiane Eberhardt, Simone Becattini, and Mirco Schmolke. Ex vivo modelling of lung tissue resident antimicrobial responses. *bioRxiv*, 2025. doi: 10.1101/2025.01.27.634989. URL http://dx.doi.org/10.1101/2025.01.27.634989.

David Maestro, Ana Palanca, Helena Soto, I Llarena, Alisa Nicole DeGrave, Gabriela Guedes, Guilherme Henrique de Oliveira, André Luiz Coelho Conceição, Verónica Mieites, Jose M. Icardo, Carlos Sanchez-Cano, Olga M. Conde, Susanne Lutz, Aitziber L. Cortajarena, and Ana V. Villar. Cardiac fibrosis inhibitor CTPR390 prevents structural and morphological changes in collagen and fibroblasts of engineered human connective tissue. *bioRxiv*, 2025. doi: 10.1101/2025.01.15.633162. URL http://dx.doi.org/10.1101/2025.01.15.633162.

Karl A. Widney, Lauren C. Phillips, Leo M. Rusch, and Shelley D. Copley. A cheater founds the winning lineages during evolution of a novel metabolic pathway. *bioRxiv*, 2025. doi: 10.1101/2025.01.26.634942. URL http://dx.doi.org/10.1101/2025.01.26.634942.

Stephen M. Smith, Karla L. Miller, and Thomas E. Nichols. Characterising ongoing brain aging and baseline effects from cross-sectional data. *bioRxiv*, 2025. doi: 10.1101/2025.01.21.634044. URL http://dx.doi.org/10.1101/2025.01.21.634044.

Alan Marín. The first complete mitogenome of the endangered spotted eagle ray, Aetobatus ocellatus, reinforces its classification within the family Aetobatidae. *bioRxiv*, 2025. doi: 10.1101/2025.01.13.632509. URL http://dx.doi.org/10.1101/2025.01.13.632509.

Bernadeta Dadonaite, Allison R. Burrell, Jenni Logue, Helen Y. Chu, Daniel C. Payne, David B. Haslam, Mary A. Staat, and Jesse D. Bloom. SARS-CoV-2 neutralizing antibody specificities differ dramatically between recently infected infants and immune-imprinted individuals. *bioRxiv*, 2025. doi: 10.1101/2025.01.17.633612. URL http://dx.doi.org/10.1101/2025.01.17.633612.

R Green, H Richards, D Ozbilek, F Tyrrell, V Barton, SC Lovell, DR Gifford, M Lagator, AJ McBain, R Krašovec, and CG Knight. Evolutionary potential of the Escherichia coli antimutator Δ`nudJ` is reduced via altered mutational spectrum. *bioRxiv*, 2025. doi: 10.1101/2025.01.28.635215. URL http://dx.doi.org/10.1101/2025.01.28.635215.

Hagen Sülzen, Pavla Fajtova, Anthony J O'Donoghue, Evzen Boura, and Jan Silhan. Structural insights into Salinosporamide A mediated inhibition of the human 20S proteasome. *bioRxiv*, 2025. doi: 10.1101/2025.01.28.635221. URL http://dx.doi.org/10.1101/2025.01.28.635221.

Abigail Pearce, Poonam Kumari, Claudia M Sisk, Matthew Harris, Ho Yan Yeung, Sabrina Winfield, Kathleen M Caron, and Graham Ladds. Receptor Activity-Modifying Protein 3 enhances GLP-1-mediated Insulin Secretion. *bioRxiv*, 2025. doi: 10.1101/2025.01.24.634724. URL http://dx.doi.org/10.1101/2025.01.24.634724.

Anna M Biller, Johannes Zauner, Christian Cajochen, Marisa A Gerle, Vineetha Kalavally, Anas Mohamed, Lukas Rottländer, Ming-Yi Seah, Oliver Stefani, and Manuel Spitschan. Physiologically-relevant light exposure and light behaviour in Switzerland and Malaysia. *bioRxiv*, 2025. doi: 10.1101/2025.01.07.631760. URL http://dx.doi.org/10.1101/2025.01.07.631760.

David Laub, Aaron Ho, Jeff Jaureguy, Adam Klie, Rany M. Salem, Graham McVicker, and Hannah Carter. GenVarLoader: An accelerated dataloader for applying deep learning to personalized genomics. *bioRxiv*, 2025. doi: 10.1101/2025.01.15.633240. URL http://dx.doi.org/10.1101/2025.01.15.633240.

Hamish Johnson, James Longden, Gary Cameron, Gordon D. Waiter, Fergal M. Waldron, Jenna M. Gregory, and Holly Spence. Machine learning identifies routine blood tests as accurate predictive measures of pollution-dependent poor cognitive function. *bioRxiv*, 2025. doi: 10.1101/2025.01.10.632396. URL http://dx.doi.org/10.1101/2025.01.10.632396.

Jongbum Na, Joongyu Heo, Minchan Jeong, Sangho Ji, Young Ho Ko, Alaleh Shafiei, Nilufer Baldir, Hasan DeMirci, Wookyung Yu, and Jin Hae Kim. The N-terminal order-disorder transition is a critical determinant for a metamorphosis of IscU. *bioRxiv*, 2025. doi: 10.1101/2025.01.09.632086. URL http://dx.doi.org/10.1101/2025.01.09.632086.

Arturo Tozzi. An extended Stokes' theorem for spiral paths: applications to rotational flows in Trachelospermum jasminoides stems and flowers. *bioRxiv*, 2025. doi: 10.1101/2025.01.15.633128. URL http://dx.doi.org/10.1101/2025.01.15.633128.

Swabha Swayamdipta, Ankur P. Parikh, and Tom Kwiatkowski. Multi-mention learning for reading comprehension with neural cascades. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyRnez-RW.

Chen Zhang, Jiuheng Lin, Xiao Liu, Yuxuan Lai, Yansong Feng, and Dongyan Zhao. How many answers should i give? an empirical study of multi-answer reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023. URL https://arxiv.org/abs/2306.00435.

Sungrim Moon, Huan He, Heling Jia, Hongfang Liu, and Jungwei Wilfred Fan. Extractive clinical question-answering with multianswer and multifocus questions: Data set development and evaluation study. *JMIR AI*, 2:e41818, 2023. ISSN 2817-1705. doi: 10.2196/41818. URL http://dx.doi.org/10.2196/41818.

Jui-I Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. MESAQA: A dataset for multi-span contextual and evidence-grounded question answering. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10891–10901, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.724/.

D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343, 1975. ISSN 1557-7317. doi: 10.1145/360825.360861. URL http://dx.doi.org/10.1145/360825.360861.

Donald E. Knuth, James H. Morris Jr., and Vaughan R. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977. doi: 10.1137/0206024. URL https://doi.org/10.1137/0206024.

Páll Melsted, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell rna-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00870-2. URL http://dx.doi.org/10.1038/s41587-021-00870-2.

Delaney K. Sullivan, Kyung Hoi Min, Kristján Eldjárn Hjörleifsson, Laura Luebbert, Guillaume Holley, Lambda Moses, Johan Gustafsson, Nicolas L. Bray, Harold Pimentel, A. Sina Booeshaghi, Páll Melsted, and Lior Pachter. kallisto, bustools and kb-python for quantifying bulk, single-cell and single-nucleus rna-seq. *Nature Protocols*, 20(3):587–607, 2024. ISSN 1750-2799. doi: 10.1038/s41596-024-01057-0. URL http://dx.doi.org/10.1038/s41596-024-01057-0.