MIDAS_SYNUR at MEDIQA-SYNUR 2026: A Prompting Study for Clinical Observation Extraction from Nurse Dictation Transcriptions

Swetha Krishna Sriram, Akshitaa Sahoo
MIDAS Lab, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
New Delhi, India
kswetha036@gmail.com, akshitaa23068@iiitd.ac.in

Abstract

This paper describes MIDAS_SYNUR, a system developed for the MEDIQA-SYNUR task at ClinicalNLP 2026 on observation extraction from nurse dictations. The primary system adopts a single-prompt, field-rich few-shot strategy using GPT-5.2, jointly generating all schema fields in one structured output. Few-shot demonstrations are curated and grouped by value type, with five examples per type, promoting consistency across heterogeneous value distributions while leveraging global context to resolve cross-field dependencies. To analyze design trade-offs, this holistic strategy is compared against a field-wise decomposed prompting baseline, where each schema field is extracted independently using explicit positive and NULL demonstrations to improve absence detection and reduce cross-field interference. Zero-shot variants of both approaches are also evaluated to isolate the contribution of in-context examples. The results highlight inference-time prompting as a simple, reproducible, and competitive baseline for large-scale clinical observation extraction from conversational nurse dictations. **Keywords:** Prompt Engineering, Few-Shot Prompting, In-Context Learning, Structured Output

1. Introduction

Nurse dictation transcripts capture clinically important observations and care details, but they are expressed in conversational language and often lack explicit structure. Automatically extracting and normalizing these observations can reduce documentation burden and improve the completeness of patient records, yet it remains challenging due to variability in phrasing, implicit mentions, and frequent negations.

MEDIQA-SYNUR @ ClinicalNLP 2026 addresses this setting by requiring systems to populate a large set of predefined fields under a strict output schema. A practical challenge is deciding how to prompt large language models to cover many fields reliably: extracting fields independently can simplify NULL handling and reduce cross-field interference, while extracting all fields jointly can leverage global context and improve coherence.

In this work, we present MIDAS_SYNUR and focus on a comparative study of prompting strategies for observation extraction from nurse dictations. We evaluate two approaches using GPT-5.2: (i) *field-wise decomposed prompting*, where each field is extracted with a dedicated prompt and explicit positive/NULL demonstrations, and (ii) *single-prompt holistic extraction*, where all fields are predicted jointly in one structured prompt with value-type-grouped few-shot examples. We also include zero-shot variants to quantify the contribution of in-context examples, and discuss the resulting trade-offs in reliability, interference, and overall extraction quality.

2. Related Work

Clinical information extraction (IE) from unstructured text has been a long-standing research area in clinical natural language processing, driven by the need to transform free-text documentation into structured representations usable for downstream applications such as decision support, cohort identification, and workflow automation (Wang et al., 2018). Early clinical IE systems were predominantly rule-based, relying on hand-crafted patterns, lexicons, and domain ontologies to extract clinical concepts and events from electronic health records. Widely used systems such as MedLEE, MetaMap, and cTAKES demonstrated strong precision in constrained settings but suffered from limited scalability and portability across institutions due to variation in clinical language and documentation practices (Y. et al., 2012).

Subsequent work shifted toward statistical and machine learning approaches, including Conditional Random Fields and other feature-based models, which improved adaptability over purely rule-based systems. With the advent of deep learning, neural sequence models such as BiLSTMs and attention-based architectures further improved performance by learning distributed representations directly from data. Transformer-based models marked a significant advance, with domain-specific encoders such as BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and BlueBERT (Peng et al., 2019) achieving strong results across a range of clinical NLP tasks (Turchin et al., 2023), including entity recognition and relation extraction. Pre-training on biomedical or clinical corpora allows these models to better understand medical terminology, abbreviations, and multi-word expres-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 244–250

sions compared to general-purpose BERT. However, these approaches typically require substantial annotated data and computational resources for fine-tuning, which can limit their practicality in resource-constrained settings.

More recently, large language models (LLMs) have been explored for clinical information extraction using zero-shot or few-shot prompting. Prior work has shown that general-purpose LLMs can perform competitively on structured extraction tasks when guided by carefully designed prompts, sometimes without any domain-specific fine-tuning. Within the MEDIQA shared task series, several systems have demonstrated the effectiveness of prompt-based approaches (Mehta, 2025; Karim and Uzuner, 2025) for extracting structured medical orders from conversational transcripts, highlighting the role of schema constraints, in-context examples, and output validation in improving reliability.

Our work builds on this line of research by focusing on schema-constrained extraction for MEDIQA-SYNUR, emphasizing strict format compliance and evidence-backed extraction across a large number of heterogeneous clinical fields. Rather than pursuing domain-specific fine-tuning, we explore how prompt engineering and schema-aware inference with a general-purpose LLM can support scalable and reliable structured extraction from nurse dictation transcripts.

3. Task Description

The MEDIQA-SYNUR shared task (Michalopoulos et al., 2026) focuses on the extraction and normalization of structured clinical observations from nurse dictation transcripts. These transcripts are derived from spoken nurse documentation and contain detailed patient observations expressed in conversational and often unstructured language. The objective of the task is to automatically capture clinically salient information from nurse-patient conversations, normalize extracted observations, and map them to a large ontology of clinical concepts.

Given a nurse dictation transcript $D = \{u_i\}_{i=1}^{N}$, where each u_i represents an utterance segment within the dictation, systems are required to populate a predefined set of clinical fields specified by the official MEDIQA-SYNUR output schema. Each field corresponds to a particular clinical concept or observation and may require extracting values such as textual descriptions, categorical labels, numerical measurements, or normalized clinical indicators. Multiple fields may be populated from a single transcript, and not all fields are guaranteed to be present in every instance.

In addition to identifying relevant clinical information, systems must distinguish between clinically meaningful content and incidental or non-actionable language that commonly appears in spoken dictations. Submissions must also adhere to the provided schema by using the correct field identifiers and value types, and by handling missing information consistently (e.g., outputting NULL when a field is not supported by the transcript). Overall, the task emphasizes robust clinical language understanding together with reliable, structured prediction at scale.

4. Dataset Description

The MEDIQA-SYNUR dataset supports structured extraction of clinical observations from nurse dictation transcripts using a predefined schema. The task targets nursing documentation and aims to reduce documentation burden by automatically capturing clinically salient information from conversational nurse dictations. The dataset follows the schema design and annotation framework introduced in the SYNUR dataset (Corbeil et al., 2025).

The output schema defines 193 target clinical observation fields. Each field is specified by a unique identifier, descriptive name, and value type. Supported value types include `STRING`, `NUMERIC`, `SINGLE_SELECT`, and `MULTI_SELECT`. Categorical fields dominate the schema, with 130 `SINGLE_SELECT` and 12 `MULTI_SELECT` fields, and an average of approximately three enumerated values per categorical field.

The dataset is released with 122 training instances, 101 development instances, and 199 test instances. Across splits, nurse dictation transcripts are of comparable length, averaging approximately 180–185 words per instance. Despite the large schema size, individual transcripts populate only a small subset of fields: on average, 12–14 observations per instance, corresponding to less than 8% of the full schema. Fields not supported by the transcript are omitted from the output, resulting in sparse, variable-length structured annotations. Figure 1 shows the cumulative distribution of populated schema fields per instance across training, development, and test splits. The cumulative distribution of populated fields per instance is closely aligned across splits, indicating comparable structural density.

For evaluation, the test split contains only nurse dictation transcripts without gold annotations. Systems are required to generate complete, schema-compliant JSON outputs for each input instance, with gold labels withheld to enable blind evaluation.

5. Methodology

We frame MEDIQA-SYNUR as a schema-constrained extraction problem in which a large

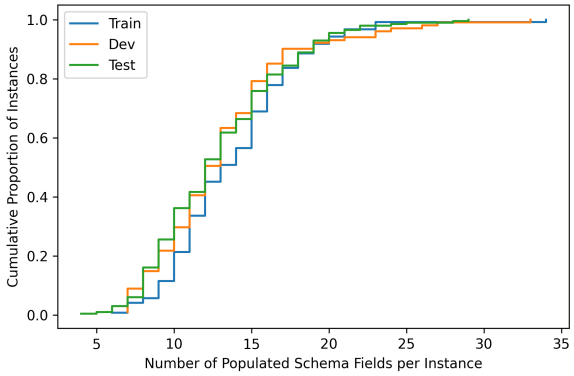

Figure 1: Cumulative distribution of the number of populated schema fields per instance across training, development, and test splits.

language model maps a nurse dictation transcript to a set of structured clinical observations defined by a fixed schema. We do not perform task-specific fine-tuning. Instead, we rely on prompt design, in-context examples, and post-generation validation to control extraction behavior, with particular emphasis on null handling and schema compliance.

All experiments are run using the OpenAI API with deterministic decoding (temperature = 0). The submitted system uses GPT-5.2, while additional models are evaluated for comparative analysis. The maximum output length is set to 16,384 tokens to avoid truncation. No constrained decoding or grammar-based enforcement is applied; instead, output validity is enforced through explicit prompt instructions and post-processing.

5.1. Prompt Design

We design structured prompt templates to control few shot sample usage, schema grounding, null handling, and output formatting. The prompts are organized into modular blocks that remain consistent across extraction strategies. A unified template is shown in Figure 2. Our prompt templates consist of the following blocks.

Block A: Core Instructions and Rules. This block contains the shared instructions used across all prompt variants. The model is instructed to extract information strictly from the provided transcript and not to infer, assume, or introduce information that is not explicitly stated. If a field is not supported by clear evidence in the transcript, it must be returned as `NULL` in the field-wise setting or omitted entirely in the single-prompt setting. The output must contain only the required fields in the specified format, with no additional explanation. The instructions allow limited clinical interpretation when mapping common clinical expressions to schema fields (e.g., device mentions), but guessing missing values is explicitly prohibited. When the evidence is unclear or indirect, the model is instructed not to extract the field. These rules aim to reduce unsupported extractions and ensure consistent handling of sparse or incomplete information.

For a subset of schema fields that are prone to systematic ambiguity, we include short, field-specific notes directly in the prompt. These notes are triggered by the field name and provide clarifications that distinguish closely related clinical concepts (e.g., bladder scan volume versus urinary output, suprapubic region versus abdomen) or require explicit mention of a specific assessment or checklist. This mechanism is used sparingly and only for fields where preliminary analysis showed consistent false positives without such clarification. The complete set of notes is shown in Figure 3. Notes were manually constructed based on error inspection on the development set.

Block B: Schema Exposure. The schema definition is injected directly into the prompt. For each field, the following metadata is provided: field identifier, field name, value type (`NUMERIC`, `SINGLE_SELECT`, `MULTI_SELECT`, `STRING`), and allowed enumeration values when applicable. This constrains the output space through explicit instruction rather than decoding constraints.

Block C: Demonstrations (Few-Shot Examples). Few-shot variants include minimal demonstrations. In both settings, demonstrations are drawn exclusively from the training set. For the field-wise setting, examples are selected as the first positive instance and first NULL instance found for each field when iterating over training examples in order. For the single-prompt setting, up to five positive examples per value type are collected by iterating through schema fields in order and taking the first matching training instance; once the quota for a type is reached, further examples of that type are skipped. In both cases, the same fixed set of demonstrations is used for all test instances, i.e., examples are not selected dynamically per input. Demonstrations are explicitly marked as format guidance and are not to be used as evidence for the target instance.

Block D: Output Specification and Target Instance. The final block contains the target transcript and the required output format. The expected structure differs across strategies, as described below.

```
BLOCK A: CORE RULES
------------------
- Use ONLY the TARGET TRANSCRIPT as
evidence.
- Do NOT infer, guess, or paraphrase.
- Do NOT use examples as evidence for
answers.
- If not explicitly mentioned:
  * Field-wise: output NULL
  * Single-prompt: omit field
- ...
- <May include specific notes for
certain fields>

BLOCK B: SCHEMA EXPOSURE
-----------------------
For each field:
- Field ID
- Field name
- Value type
- Allowed values (if any)

Type constraint example:
NUMERIC:
- Extract digits only.
- No units or text.

BLOCK C: DEMONSTRATIONS
----------------------
(Few-shot variants only)

Positive example:
Transcript: "..."
Output: <value>

NULL example:
Transcript: "..."
Output: NULL

BLOCK D: OUTPUT
----------------------
TARGET TRANSCRIPT:
"<text>"

OUTPUT FORMAT:
Field-wise: ANSWER: <value | NULL>
Single-prompt: valid JSON
```

Figure 2: Illustrative prompt template with block-wise structure.

5.1.1. Field-Wise Decomposed Extraction

In the field-wise strategy, each schema field is extracted independently using a separate model call. Block B contains only the metadata for the target field. The required output is a single value or `NULL`. Type-specific constraints (e.g., numeric-only output for `NUMERIC` fields, strict enumeration matching for categorical fields) are defined in the schema exposure block. This strategy isolates extraction decisions but requires one model call per schema field, adding up to `~190` calls per transcript.

5.1.2. Single-Prompt Holistic Extraction

In the single-prompt strategy, the full schema is provided in Block B. The model is instructed to return a valid JSON containing only fields explicitly supported by the transcript. Unsupported fields must not appear in the output. Unlike the field-wise formulation, this approach allows cross-field contextual reasoning within a single generation while reducing inference cost to one model call per transcript.

5.1.3. Zero-Shot Variants

To evaluate the contribution of demonstrations, we remove Block C while keeping Blocks A, B, and D unchanged. This isolates the effect of in-context examples without altering rule enforcement or schema exposure.

6. Results and Discussion

Our submitted system initially achieved an F1 of 0.773 on the MEDIQA-SYNUR leaderboard, ranking 7th. The top-ranked system obtained an F1 of 0.814. The submitted configuration uses GPT-5.2 in a single-prompt holistic extraction setting with type-wise few-shot demonstrations.

Post-competition, we identified inconsistencies in predicted field IDs caused by leading zeros (e.g., `02` instead of `2`). These arose because the model occasionally reproduced field IDs with zero-padding, which caused misalignments during evaluation since the official schema uses unpadded integers. We applied a one-time post-processing step that strips leading zeros from all predicted field IDs before evaluation. This normalization improved our submitted F1 from 0.773 to 0.82. All results reported below use this corrected output. No other modifications to predictions were made.

6.1. Overall Performance

Table 1 presents a comparative analysis of prompting strategies explored during development, including single-prompt and field-wise extraction settings under both few-shot and zero-shot conditions for GPT-5.1 and GPT-5.2. All reported results are computed on the official test set, for which gold annotations were released after the competition concluded.

We find that GPT-5.2 outperforms GPT-5.1 across all prompting variants, with F1 gains of $\sim$0.25–0.30. Recall exceeds precision for GPT-5.1

```
BLOCK A: FIELD-SPECIFIC NOTES
------------------------------
Applied sparingly to fields with
systematic false positives:

[Voiding function]
NOTE: Extract a value ONLY if the
term "Voiding function" is explicitly
used in the target transcript.

[Speech content]
NOTE: Refers to the nature of the
patient's communication or speech.
Extract ONLY if such nature of
communication/speech is explicitly
mentioned.

[Broset violence checklist fields]
NOTE: A specific checklist that must
be noted/performed by the nurse.
Extract (Yes/No) ONLY if the
transcript EXPLICITLY mentions the
field name, i.e., both the checklist
and the attribute together.

[Bladder scan volume]
NOTE: Bladder scan volume is different
from urinary output/volume. Look for
explicit mentions of Bladder scan
volume only.

[Fall risk identification]
NOTE: Refers to a patient's risk of
falls or conclusions from a fall risk
assessment. Extract the level of risk
ONLY if explicitly mentioned.

[Patient safety]
NOTE: Answer Yes ONLY if measures for
patient safety are explicitly stated
as taken. Answer No ONLY if explicitly
stated as not taken/not required.
Answer NULL if not explicitly mentioned.

[Patient identification]
NOTE: Refers to age and gender only.
Age must be an exact number (e.g.,
"X years old"); ignore descriptors
like "elderly" or "young". Extract
ONLY if both age and gender appear
together (e.g., "X-year-old female").

[Abdomen fields]
NOTE: Suprapubic region is NOT the
same as abdomen.

These notes were manually constructed
from development set error analysis.
```

Figure 3: Field-specific clarification notes included in Block A for fields prone to systematic ambiguity.

in all settings, suggesting a bias toward extracting potentially relevant fields even when evidence is weak. In addition to task-specific factors, the performance gap between GPT-5.1 and GPT-5.2 is consistent with generally reported improvements in the latter model. On this task, GPT-5.2 achieves higher precision and recall across all prompting configurations, suggesting more reliable extraction overall. GPT-5.1 underperforms on both dimensions, with particularly low precision, indicating a tendency to predict fields even when transcript evidence is weak. These gains likely reflect broader improvements in instruction-following and structured output generation in GPT-5.2, though the precise reasons remain difficult to isolate from prompt-level evaluation alone.

Few-shot prompting is seen to consistently improve precision and recall. Field-wise prompting offers modest gains in recall, with an advantage that likely stems from clearer task scoping. However, the margin is small, indicating that well-structured unified prompts can achieve comparable performance at a much lower inference cost.

Model	Approach	P	R	F1
GPT-5.1	Single-prompt (few-shot)	0.47	0.65	0.54
GPT-5.1	Field-wise (few-shot)	0.43	0.68	0.53
GPT-5.1	Single-prompt (zero-shot)	0.41	0.64	0.50
GPT-5.1	Field-wise (zero-shot)	0.41	0.66	0.51
GPT-5.2	Single-prompt (few-shot)	**0.82**	**0.81**	**0.82**
GPT-5.2	Field-wise (few-shot)	0.82	0.83	**0.83**
GPT-5.2	Single-prompt (zero-shot)	0.77	0.75	0.76
GPT-5.2	Field-wise (zero-shot)	0.77	0.75	0.76

Table 1: MEDIQA-SYNUR results (precision, recall, F1).

6.2. Type-Level Performance

Tables 2 and 3 present type-level analysis for GPT-5.1 and GPT-5.2 in the single-prompt holistic extraction setting.

GPT-5.2 shows substantial gains across numeric and single-select fields, driven by improved unit parsing, schema adherence, and reduced false positives. Multi-select fields improve moderately, while string fields remain challenging due to sparsity and ambiguous textual cues. Detailed analysis of string fields reveals that many failures are

Field Type	Precision	Recall	F1
MULTI_SELECT	0.680	0.708	0.693
NUMERIC	0.473	0.798	0.594
SINGLE_SELECT	0.466	0.641	0.539
STRING	0.020	0.051	0.029

Table 2: Type-level performance for GPT-5.1 (single-prompt few-shot).

Field Type	Precision	Recall	F1
NUMERIC	0.919	0.965	0.941
MULTI_SELECT	0.939	0.817	0.874
SINGLE_SELECT	0.846	0.832	0.839
STRING	0.021	0.031	0.025

Table 3: Type-level performance for GPT-5.2 (single-prompt few-shot).

caused by exact-match evaluation, case or whitespace differences, minor formatting variations (e.g., “75 mL” vs “75mL”), and predictions that include extra descriptive text. Moreover, the model exhibits a tendency to overgenerate text, producing verbose outputs that partially match gold values but are penalized under strict matching. These insights suggest that STRING field evaluation may benefit from partial-match or semantic similarity metrics to better capture meaningful predictions.

6.3. Field-Level Analysis

Field-level performance was analyzed for GPT-5.2 and compared to GPT-5.1 for the single-prompt holistic extraction setting. We highlight trends using a bar plot of the top 5 fields with the largest F1 improvements. On average, GPT-5.2 improved field-wise F1 scores by 0.271 over GPT-5.1.

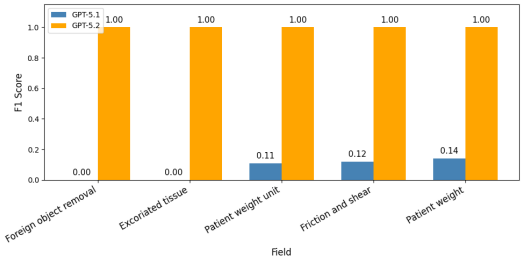

Figure 4: Top 5 fields showing the largest F1 score improvement from GPT-5.1 to GPT-5.2.

6.4. Error Analysis

We identify four main failure modes from examining test set predictions.

STRING field verbosity. The most consistent failure mode across both models is on STRING fields (F1: 0.025–0.029) with severe degradation in performance. The model frequently generates verbose outputs that partially match gold values but are penalized under strict evaluation. These errors suggest that STRING field evaluation would benefit from semantic similarity scoring rather than string equality.

False positives from cross-field confusion. A recurring error in both field-wise and single-prompt settings is the prediction of semantically related but distinct fields. For instance, urinary output values are occasionally extracted for the Bladder scan volume field, and abdomen-related observations are predicted for suprapubic fields. These confusions motivated the field-specific clarification notes described in Section 5.1. Their persistence in the results suggests that even explicit disambiguation notes do not fully resolve ambiguity in cases where clinical phrasing is inherently overlapping.

NULL prediction errors. GPT-5.1 shows a strong bias toward predicting fields even when transcript evidence is weak, reflected in its consistently low precision across all settings. GPT-5.2 largely corrects this, though occasional false positives remain, particularly for SINGLE_SELECT fields where a plausible enumerated value exists even without direct transcript support.

Schema compliance failures. A small number of predictions contain formatting inconsistencies such as zero-padded field IDs, extra whitespace, or unit strings appended to NUMERIC values (e.g., `75 bpm` instead of `75`). Padding inconsistencies are addressed by post-processing but the other issues remain.

7. Conclusion

In this work, we presented **MIDAS_SYNUR**, a prompting-based system for schema-constrained extraction of clinical observations from nurse dictation transcripts. Our experiments show that **GPT-5.2 with few-shot examples** achieves the strongest performance, consistently outperforming GPT-5.1 across all evaluation metrics.

We find that the single-prompt holistic extraction approach is preferable to field-wise decomposed extraction from a deployment standpoint. While field-wise few-shot extraction achieves a marginally higher F1 (0.83 vs. 0.82), the single-prompt approach requires only one inference call per transcript compared to approximately 190 for field-wise extraction, resulting in substantially reduced computation time and lower token usage. Given this cost difference, the single-prompt approach represents a more practical choice when the marginal F1 gap does not justify the added inference overhead.

Beyond overall performance, we provide a detailed analysis of predictions at both the type and field level. `STRING` fields remain the primary weakness, with errors largely due to sparse mentions, minor formatting inconsistencies, and strict exact-match evaluation criteria. These insights suggest potential avenues for future improvement, such as semantic similarity scoring or relaxed matching strategies for textual fields.

Overall, our study demonstrates that carefully designed prompt strategies, combined with few-shot demonstrations, provide a simple, reproducible, and effective framework for large-scale structured extraction from conversational clinical documentation.

8. Ethics Statement

This work benchmarks large language models (GPT-5.1 and GPT-5.2) in a controlled evaluation setting. As these models are pretrained on large-scale data, they may reflect and amplify societal biases related to language, culture, or demographics. Accordingly, performance differences should not be interpreted as purely objective measures of capability.We note that these models are proprietary and subject to change, which may limit full reproducibility. We mitigate this by clearly documenting prompts and evaluation settings. Finally, this work is intended for research purposes only. The evaluated models may produce incorrect or misleading outputs, and should not be deployed in high-stakes settings without appropriate safeguards and human oversight.

9. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jerome Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, Nathan Bodenstab, Thomas Lin, Wenxuan Teng, Francois Beaulieu, and Paul Vozila. 2025. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Suzhou (China). Association for Computational Linguistics.

A. H. M. Rezaul Karim and Özlem Uzuner. 2025. Masonnlp at mediqa-oe 2025: Assessing large language models for structured medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 57–67. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Parth Mehta. 2025. Pnlp at mediqa-oe 2025: A zero-shot prompting strategy with gemini for medical order extraction. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 75–83. Association for Computational Linguistics.

George Michalopoulos, Jean-Philippe Corbeil, Cari Bader, Nate Bodenstab, and Asma Ben Abacha. 2026. Overview of the mediqa-synur 2026 shared task on observation extraction from nurse dictations. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Alexander Turchin, Stanislav Masharsky, and Marinka Zitnik. 2023. Comparison of bert implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36:101139.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Wu Y., Denny J.C., Rosenbloom S.T., Miller R.A., Giuse D.A., and Xu H. 2012. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA Annual Symposium Proceedings*, pages 997–1003. PMID: 23304386.