Journal of KIIT. Vol. 23, No. 10, pp. 23-32, Oct. 31, 2025. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2025.23.10.23

# 가우시안 확률과 LLM을 이용한 비정형 의료데이터 입력 시스템

구서연*, 최권택**

# Unstructured Medical Data Entry System using Gaussian Probabilities and Large Language Models

Seoyon Koo*, Kwon-Taeg Choi**


## 요 약

최근 AI 및 IT 기술의 발전에 따라 의료 분야에서도 데이터 입력 및 처리와 같은 업무 영역에 이러한 기술들을 도입하려는 시도가 이루어지고 있다. 그러나 기존 시스템은 문맥이 없는 비정형 데이터나 실시간 자연어 입력을 효과적으로 처리하지 못하는 한계가 있다. 본 연구는 이러한 한계를 해결하기 위해, LLM 기반 프롬프트 엔지니어링에 수치형 필드의 가우시안 확률 정보를 결합한 구조화 방법론을 제안한다. 실험에서는 COVID, Diabetes 2개 데이터 세트와 5개의 LLM 모델로 혼합 입력, 불완전한 입력 등 다양한 환경에서의 실험을 통하여 제안 방법의 강건성을 확인하였다. 특히, COVID, Diabetes 데이터 세트에서 기존 방식 대비 제안 방식의 성능이 각각 평균 26.11%와 49.06% 개선되었다. 이를 통해, 본 연구는 기존 연구들과는 달리 문맥 없는 비정형 의료 입력 데이터를 구조화된 데이터로 변환하는 방향성을 제시한다.

## Abstract

With the recent advancements in AI and IT technologies, there have been increasing efforts to integrate these technologies into the medical domain, particularly for tasks such as data entry and processing. However, current systems often struggle with real-time or context-deficient unstructured inputs. To address these limitations, we propose a novel structuring methodology that combines LLM-based prompt engineering with Gaussian probability modeling for numerical fields. Experiments on COVID and Diabetes datasets with five LLM models under various conditions confirmed the robustness of the proposed approach, yielding performance gains of 26.11% and 49.06% over conventional methods. This study offers a new direction for converting context-poor, unstructured medical inputs into structured data.



Keywords

large language models, Gaussian probability distribution, unstructured data, field mapping, prompt engineering



* 강남대학교 인공지능융합공학부 학사과정
- ORCID: https://orcid.org/0009-0006-8210-9916
** 강남대학교 ICT융합공학부 교수(교신저자)
- ORCID: http://orcid.org/0000-0001-5331-321X



· Received: Jul. 21, 2025, Revised: Aug. 20, 2025, Accepted: Aug. 23, 2025


· Corresponding Author: Kwon-Taeg Choi
Dept. of ICT Convergence Engineering, Kangnam University, Yongin, Korea
Tel.: +82-31-289-3660, Email: kwontaeg.choi@kangnam.ac.kr

# Ⅰ. 서 론

최근 AI 및 IT 기술의 도입을 통한 의료 데이터 처리 시스템의 발전은 의료진의 업무 부담을 줄이며 효율성을 높이고 있다[1][2]. 특히 신경망 기반의 자연어 처리 모델은 임상 기록, 전자 건강 기록(EHR, Electronic Health Records) 등의 비정형 의료 데이터가 증가함에 따라, 이를 구조화하여 의미 있는 정보를 추출하는 데 활용된다[3]. 기계학습 기반의 오류 탐지 기술은 의료진이 입력한 데이터의 비정상 패턴을 자동으로 식별하며, 입력 오류에 대한 사전 경고 시스템도 개발되고 있다[4]. 광학 문자 인식(OCR, Optical Character Recognition) 기술은 스캔한 문서나 수기 기록을 텍스트로 전환하여, 기존의 종이 기반 자료를 디지털화하는 데 중요한 역할을 한다[5]. 이러한 기술들은 정형 의료데이터 입력을 지원하거나 문맥이 있는 비정형 의료데이터의 정보 추출 등 구조화에 높은 수준의 성과를 보인다.

그러나 이러한 기존 시스템은, 문맥이 없는 비정형 데이터 처리, 실시간 자연어 데이터 매핑 한계 등으로 문제를 겪었다. 의료 상담과 같이 입력 대상자와 대화를 병행해야 하는 경우, 자연어로 입력하는 것이 효율적이다[6]. 특히 의료데이터는 그 특성상 응급 상황이나 현장 진료, 병실 간 이동 등 정적인 환경이 아닌 동적·모바일 환경에서 입력이 이뤄지는 경우가 많다. 이러한 환경에서는 수기 작성이나 키보드 사용이 어렵고 문맥을 지키기 힘든 경우가 많으며, 입력 오류가 빈번하게 발생한다. 응급 상황에서는 음성 입력과 자동 구조화 기능이 필수적이며 데이터가 실시간으로 반영되어야 하나, 기존 시스템만으로는 이러한 특성의 의료데이터를 효과적으로 처리하지 못하고 있다[7].

또한, 대규모 언어 모델은 일반적으로 공개된 인터넷 자료나 도서 등에서 사전 학습되며, 이는 다양한 주제의 문맥 이해에는 유리하다. 그러나, LLM은 사전 학습하지 않은 내용에 대해 실제와 다른 내용을 사실처럼 생성하는 환각 현상을 보이는 문제점이 있어[8], 특정 분야—특히 고신뢰성과 정밀성이 요구되는 의료 분야—에서는 이러한 한계를 인식하고 의료 분야에 특화된 정밀한 데이터 구조화를 위한 새로운 접근법이 필요하다[9].

본 연구는 대규모 언어 모델에 통계적 분포 정보를 결합하여 모바일 환경에서 입력한 음성 및 텍스트 의료데이터를 구조화하는 방법론을 제안한다. 특히 실수 데이터의 가우시안 정규 분포에 따라 필드 적합도를 수치화하여 적절한 필드로의 매핑을 유도한다. 이를 통해 단위 없는 수치 정보 등 불완전하게 제공되는 입력 데이터, 또는 혼합되어 있거나 맥락 정보가 부족한 비정형 입력 데이터 등에서도 올바른 필드에 자동 매핑함으로써, 기존 연구들과는 달리 문맥 없는 비정형 의료 입력 데이터를 구조화된 데이터로 변환하는 방향성을 제시한다.

본 논문의 구조는 다음과 같이 구성되어 있다. 2장에서는 연구에 이용된 배경지식과 관련 연구들을 설명하고, 3장에서는 시스템 개요 및 비정형 의료데이터를 효과적으로 매핑하기 위한 제안 방법의 프롬프트 구성을 설명한다. 4장에서는 부분 입력, 혼합 입력, 가우시안 분포가 겹치는 입력 등 다양한 시나리오를 설정하고 이에 대한 제안 방식의 일반화 성능을 검증하고, 기존 방식에 제안 방식의 각 구성을 개별 또는 조합 형태로 추가하여 제안 방식의 구성 요소별 기여도를 분석하였다. 마지막으로 5장에서 결론과 향후 연구 방향을 제시하였다.

# Ⅱ. 관련 연구

## 2.1 의료데이터 입력에 대한 관련 연구

의료데이터는 종종 진료 중 기록되는 문장 형태의 음성, 메모 등 비정형 형태로 존재하며, 이는 전통적인 데이터베이스나 구조화 모델로는 쉽게 처리하기 어렵다[10]. 전자 건강 기록은 의료 분야의 필수적인 자원으로 자리 잡았으나, 전자 건강 기록 전체 데이터의 약 80%를 차지하는 비정형 정보는 주석 부족 등으로 인해 2차 활용에 어려움을 겪고 있다[11][12]. 특히 상담 기록, 응급 대응 상황, 현장 진료 등에서는 자연어 입력이 빠르게 이루어져야 하며, 이 과정에서 키보드 사용 제약이나 시간 부족이 데이터 품질 저하로 이어질 수 있다[13].

이러한 상황에서 LLM은 긴 문장의 PDF 보고서나 의료 진단서와 같이 문맥이 명확하게 제공되는 비정형 문서에서는 뛰어난 정보 추출 성능을 보였

으나, 문맥이 없는 비정형 데이터 처리에서는 지금까지도 어려움을 겪고 있다[14][15].

### 2.2 의료 분야의 프롬프트 엔지니어링

프롬프트 엔지니어링은 모델을 재학습하지 않고 LLM의 사전 학습 지식을 프롬프트 설계만으로 최대한 끌어내어 활용하는 방식으로, 도메인 지식 없이도 복잡한 태스크를 해결할 수 있는 핵심 기술이다. 의료 분야에서는 정보 추출의 정확성과 신뢰성이 특히 중요하기 때문에 다양한 프롬프트 기법이 활발히 연구되고 있다[16][17]. 구조화된 프롬프트는 비정형 데이터로부터 정보를 추출하는 파이프라인을 최적화하여 성능을 높일 수 있다[18].

## III. 제안 방법

### 3.1 시스템 개요

본 연구는 LLM 기반 비정형 의료데이터 입력 시스템을 제안한다. 사용자는 자신에게 해당하는 필드를 선택해, 전체 혹은 일부의 필드에 대한 데이터를 텍스트 또는 음성으로 입력한다. 이 상황 정보는 Flask 서버로 전송되고, 서버에서는 LLM을 활용한 프롬프트 엔지니어링을 통해 입력된 데이터를 분석하고 구조화된 정보로 변환하여 JSON 구조체를 생성한다. 최종적으로 이 결과를 UI에 매핑한다. 예를 들어 그림 1에서처럼, 7개의 필드 중, 사용자가 모바일 환경에서 “Buckingham 62.0 121.0”와 같이 3개 필드에 해당하는 상황 정보를 혼합하여 입력하면 그림 1의 왼쪽 상단 모바일 입력 화면에서 해당 필드 “location”, “height”, “weight”에 각각 자동으로 매핑되어 출력된다. 매핑되지 않은 필드는 빈칸으로 비워지고, 매핑된 필드는 빨간색으로 강조된다.

이를 통해 사용자 중심의 효율적이고 직관적인 의료데이터 입력 시스템을 구현함으로써, 의료데이터 관리의 정확성과 효율성을 높일 수 있다.

### 3.2 프롬프트 설계 전략

실제 의료데이터 입력 환경에서는 문자열과 수치형이 혼합된 다양한 필드 중 일부만을 사용자가 선택적으로 입력하는 경우가 많으며, 이에 따라 정보가 부분적으로 불충분한 상태로 제공된다. 특히 의료 현장에서는 입력 속도와 효율성이 중요하여, 문맥이 부족한 입력이 자주 발생한다. 기존 LLM은 이러한 입력에 대해서는 적절한 필드 추론에 한계를 보인다. 따라서, LLM이 정확한 추론을 수행하기 위해 필드에 대한 명확한 정보 제공이 필수적이다.

그림 2는 LLM이 사용자 입력과 필드를 올바르게 매핑한 JSON 형식의 출력을 생성할 수 있도록 필요한 메타데이터와 확률 기반 통계 정보를 구성하여 입력으로 제공하는 과정을 나타낸 의사 코드이다. 제안하는 방식의 프롬프트는 비정형 의료데이터를 효과적으로 올바른 필드에 매핑하기 위해 다음과 같은 4가지 구성으로 나누어져 있다.

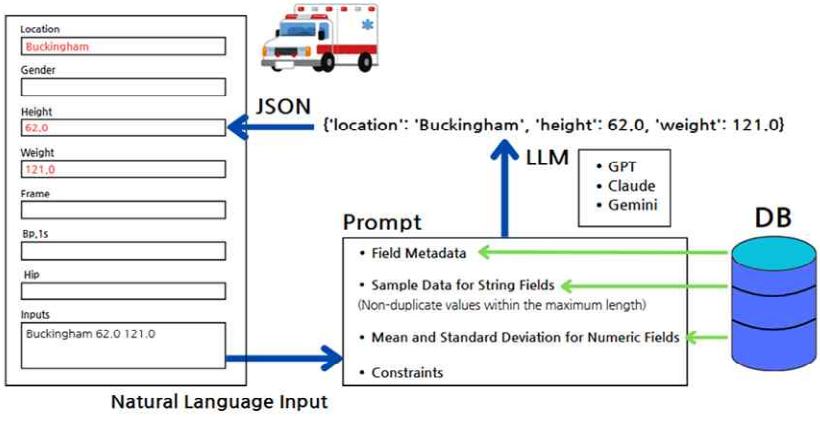


그림 1. 시스템 개요
Fig. 1. System overview

```
Algorithm
INPUT: - 1. field_names: list[string]    2. user_inputs: list[any]

PROCEDURE:
1. metadata, numeric_fields ← [], []
2. FOR field IN field_names:                          # GENERATE metadata
      type ← get_field_type(field)
      desc ← get_desc_from_db(field)
      data ← string(unique(get_values_from_db(field))[0:max_length]) IF type == 'string' ELSE ""
      metadata.append({name: field, desc: desc, data: data})
      IF type == 'numeric': numeric_fields.append(field)
3. stats ← get_stats_for_fields(numeric_fields)          # GENERATE probability
   probability_matrix ← [
      [get_probability(input, stats[field]) FOR field IN numeric_fields]  FOR input IN user_inputs
   ]
   probability ← CREATE_DATAFRAME(data = probability_matrix, index = user_inputs, columns = numeric_fields)
4. tuple_list ← [                                    # GENERATE sort_by_probability
    (numeric_fields[col_index], user_inputs[row_index], probability_matrix[row_index][col_index])
    FOR row_index IN range(len(user_inputs))
    FOR col_index IN range(len(numeric_fields))
   ]
sort_by_probability ← SORT(tuple_list, key = third_element, order = descending)

5. prompt
   - "Request field mapping for the given content"
   - "1) Metadata of Input Fields {metadata - name, desc}"
   - "2) Sample Data {metadata - data}"
   - "3) Probability Computation & Sort Information for Numerical Data {probability, sort_by_probability}"
   - "4) Constraints {constraints}"
   - "user input {user_inputs}"
   - "Request JSON output format {one-shot}"

6. Send the prompt to the LLM and parse the response
```

그림 2. 프롬프트 설계 의사 코드
Fig. 2. Prompt design pseudo code

### 3.2.1 입력 필드의 메타데이터

제안 프롬프트는 단순한 필드명 나열을 넘어 각 필드의 의미를 명시적으로 제안함으로써, LLM이 필드 간 개념적 차이를 이해할 수 있도록 한다. 의미 기반의 매핑은 잘못된 필드 연결을 줄이고, 특히 필드명이 유사하거나 직관성이 떨어지는 모호한 비정형 데이터의 구조화 정확도를 개선할 수 있다.

### 3.2.2 범주형 데이터 처리

사용자가 모바일 환경에서 특정 필드에 해당하는 데이터를 입력할 것을 전제로, LLM에 해당 필드들의 정보를 사전에 제공함으로써 필드 간 의미 차이를 이해할 수 있도록 한다. 이를 통해 입력값이 주어졌을 때, LLM이 더 정밀하게 해당 값을 적절한 필드에 매핑할 수 있도록 유도할 수 있다.

문자열 필드에는 실제로 사용되는 값들의 예시가 함께 제공되며, 이는 문자열 입력 데이터와의 표현적 유사도를 판단하는 기준으로 작용한다. 이를 통해 LLM은 사전 학습되지 않은 도메인에서도 입력값이 어떤 필드에 적절한지를 안정적으로 추론할 수 있다. 샘플 데이터는 특히 범주형 데이터의 잘못된 분류를 줄이는 데 효과적이다[19].

### 3.2.3 수치형 데이터에 대한 확률 계산

그림 3에 제시된 바와 같이, “bp.1s”와 “hip”과 확률분포가 서로 중첩되는 입력이 동시에 들어오면 LLM 모델이 판단하기 어려운 케이스로 분류할 수 있으며, 반대로 확률분포 중첩 정도가 낮은 입력은 상대적으로 분류가 용이한 케이스로 구분할 수 있다. 본 논문에서는 분포가 중첩된 경우와 중첩되지 않은 경우를 확률 기반 모델링을 통해 정량적으로 평가하였으며, 그에 따른 모델 성능의 변화와 주요 요인들이 미치는 영향을 분석했다. 수치형 입력값의 필드 매핑은 필드별로 계산한 평균($\mu$)과 표준편차($\sigma$)를 기반으로 확률 밀도 함수를 적용하여 이루어진다. 입력값이 각 필드 분포 내에서 가질 수 있는 상대적 가능성을 정량화하고, 이를 정규화하여 전체 필드에 대한 확률 분포를 계산한다. 그 결과, 수치형 입력은 확률이 가장 높은 필드부터 우선 매핑되며, 이는 데이터 기반의 타당한 필드 대응을 가능하게 한다. 본 방식은 가우시안 분포 기반의 필드 정보를 제공함으로써, LLM 내에서 필드 간 의미 공간을 보다 정확하게 모델링할 수 있도록 한다[20].

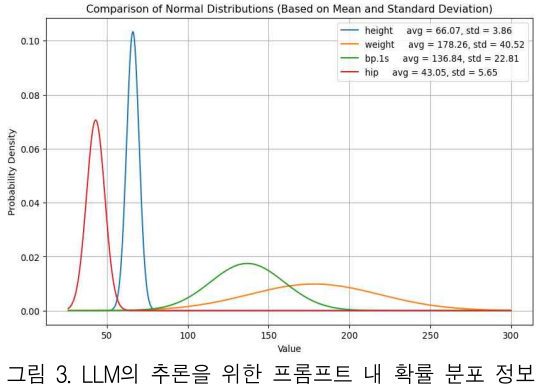


그림 3. LLM의 추론을 위한 프롬프트 내 확률 분포 정보
Fig. 3. Gaussian distribution included in the prompt for LLM-based inference

### 3.2.4 성능 향상을 위한 제약 조건

LLM은 유사한 의미의 필드가 존재할 경우, 존재하지 않는 입력 데이터를 생성하거나 하나의 입력값을 여러 필드에 중복으로 대응하는 문제가 발생한다. 이를 방지하기 위해, LLM에 적절한 제약 조건을 주어서 올바르게 매핑하도록 유도해야 한다. 따라서 본 연구는 제약 조건을 통해 LLM이 새로운 값을 생성하고 기존값을 변형하는 행위를 금지하고,

입력값을 필드와 1:1로 대응시킨다. 이를 통해, LLM이 존재하지 않는 입력 데이터를 생성하여 필드에 매핑하는 등의 환각을 발생시키는 것을 방지하고, 데이터 구조화 파이프라인을 최적화한다.

## Ⅳ. 실 험

### 4.1 실험 데이터 세트

본 연구에서는 비정형 의료데이터 입력 시스템의 성능을 평가하기 위해 Kaggle의 Covide 19 India Individual Patient Dataset와 Diabetes 공개 의료데이터 세트를 활용하여 실험하였다[21][22]. 각각 전염성 질환(COVID-19)과 만성질환(당뇨병)이라는 상이한 도메인을 다루며, 다양한 변수와 형태의 데이터를 포함하고 있어 실험 목적에 적합하다.

본 연구에서 사용한 COVID 데이터 세트의 일부 필드는 일관되지 않은 문장 구조와 유사 의미 데이터 등으로 구성되어 있어, LLM 기반 입력 시스템이 문맥이 불완전하거나 모호한 텍스트로부터 필드 정보를 얼마나 의미 있게 추출하고 자동 매핑할 수 있는지를 평가하는 데 적합한 사례이다. 두 번째로 사용된 Diabetes 데이터 세트는 수치형 변수와 문자열 변수로 구성된 혼합 의료데이터 세트이다. 따라서, 문자열 중심의 감염병 데이터와 비교하여 맥락 정보 없이 단일 수치 혹은 단어만으로 입력되는 데이터에 대한 시스템의 대응력을 평가할 수 있다.

### 4.2 실험 개요

본 연구에서는 제안한 입력 매핑 기법의 성능과 일반화 가능성을 검증하기 위해 총 3가지 테스트를 수행하였다. 첫 번째로, OpenAI의 GPT-4o 모델에서 두 데이터 세트에 대한 기존 방식(BASE)과 제안 방식(PROPOSED)의 성능을 비교하였다. 두 번째에서는 OpenAI의 GPT-4o 모델로 제안 방식의 프롬프트 구성 요소 단일, 조합별로 모델 성능에 미치는 영향을 분석하였다. 구성 요소 단일 실험은 표 2에 나타나 있으며, 구성 요소 조합 실험은 표 3과 표 4에 나타나 있다. 이를 통해 구성 요소들 사이의 상호작용과 시스템의 구조적 강점을 종합적으로 확인하고자 하였다. 마지막으로, GPT-4o, Claude-3.5-Sonnet, Gemini-2.5-Pro 등을 포함한 5개의 LLM 모델을 통해 문자열 정보와 수치형 정보가 함께 들어오는 혼잡 입력 상황에 대한 제안 방식의 일반화 성능을 평가하였다.

본 연구에서는, 입력 데이터 필드 매핑 성능을 정량적으로 평가하기 위해 세 가지 기준을 설정하였다. 첫째, 입력 데이터 기준 정확도는 입력으로 제공된 데이터 각각이 올바른 정답 필드로 대응되는지 평가한다. 둘째, 전체 필드를 기준으로 정답 매핑 비율을 측정한다. 이 두 정확도를 구분하여 측정하는 이유는, 사용자가 전체 필드 중 일부만 부분적으로 입력하는 실제 환경에서는 두 평가지표의 성능 해석이 상이하기 때문이다. 마지막으로, 행 기준 정확도는 모든 필드를 정확하게 예측한 경우에만 정답으로 간주하는 완전 일치 기준이다. 이후 실험 결과표에 제시되는 정확도는 위 순서(입력 기준, 전체 기준, 행 기준)에 따라 정렬되어 있다.

### 4.3 가우시안 확률과 LLM을 결합한 비정형 의료데이터 필드 매핑

COVID 및 Diabetes 두 가지 공개 의료데이터 세트를 기반으로, LLM 기반 입력 매핑에서 기존 방식과 제안 방식의 성능을 비교하였다. COVID는 9개의 문자열 필드로 구성되어 있어 순수 비정형 텍스트 기반 입력 조건을 시뮬레이션하며, Diabetes는 문자열과 수치형이 혼합된 입력과 더불어 수치 필드 간 가우시안 분포가 중첩된 상황을 반영한다. 실험에서는 필드의 입력 순서를 임의로 섞어 문맥 정보 없이 입력이 주어지는 현실적인 조건을 구성하였다. 또한 일부 필드만 입력되는 부분 입력 조건과, 유사한 의미나 가우시안 분포를 가지는 입력이 포함된 상황 등 다양한 시나리오를 설정으로써 다각도에서 평가하였다.

실제 환경에서는 사용자가 전체 필드를 입력하는 이상적인 상황보다, 자신에게 필요한 특정 필드만 부분적으로 입력하는 부분 입력 상황이 더 자주 발생할 수 있다. 또한, LLM이 유사 의미 필드에 대한 입력도 구분할 능력이 있는지 확인하는 것이 필요하다. 따라서, COVID 데이터 세트는 표 1에서 볼

수 있듯이 9개의 문자열 필드를 기준으로, 전체 필드를 모두 입력한 경우(9/9), 선택된 5개 필드만을 부분적으로 입력한 경우(9/5), 유사 의미 필드 3개만을 입력한 경우(9/3)로 실험을 구성하였다. Diabetes는 표 1과 같이, 실수형과 문자열이 혼합된 7개 필드를 대상으로 전체 필드를 모두 입력한 경우(7/7)와 가우시안 분포가 겹치는 실수형('weight', 'bp.1s') 2개와 문자열 1개가 입력되는 혼합(7/3) 조건을 테스트하였다.

표 1. BASE / PROPOSED 성능 비교 (정확도 기준)
Table 1. BASE vs. PROPOSED performance comparison

| Dataset | COVID | | | Diabetes | |
|---|---|---|---|---|---|
| Input condition | 9/9 | 9/5 | 9/3 | 7/7 | 7/3 |
| Base | 0.747 | 0.638 | 0.679 | 0.649 | 0.353 |
| | 0.747 | 0.599 | 0.887 | 0.649 | 0.580 |
| | 0.387 | 0.005 | 0.484 | 0.160 | 0.015 |
| Proposed | 0.969 | 0.937 | 0.740 | 0.952 | 0.881 |
| | 0.969 | 0.934 | 0.896 | 0.952 | 0.940 |
| | 0.822 | 0.710 | 0.546 | 0.823 | 0.802 |

COVID와 Diabetes 두 가지 데이터 세트를 기반으로 기존 방식과 제안 방식의 성능을 비교한 결과, 기존 방식은 문자열 필드만 존재하는 이상적인 조건(9/9)에서도 입력 기준 정확도가 최대 74.77%에 머물렀으며, 자료형이 혼합된 조건(7/7)에서는 최대 64.92%, 자료형이 혼합되고 가우시안 분포가 겹치는 조건(7/3)에서는 35.35%까지 하락하는 등 불안정한 성능을 보였다. 반면, 본 연구의 제안 방식은 동일 조건(9/9)에서 최대 96.97%의 정확도를 기록하였고, 혼합 입력 조건(7/7)에서도 최대 95.22%, 기존 방식이 불안정한 성능을 보였던 자료형이 혼합되고 가우시안 분포가 겹친 조건(7/3)에서도 88.15%로 안정적인 정확도를 유지하였다. 특히 COVID 데이터 세트에서는 입력 기준 정확도가 평균 19% 이상, Diabetes 데이터 세트에서는 입력 기준 정확도가 평균 41% 증가하였다. 이는 LLM에 수치형 필드에 대한 가우시안 분포를 결합하는 것이 정밀 추론 성능 향상에 기여함을 시사한다. 또한, 제안 방식은 입력 정보의 문맥 결여, 혼합, 필드 간 중첩 등 다양한 조건에서 안정적인 성능을 유지하였으며, LLM 기반 입력 구조화 시스템의 실용 가능성을 보여 주었다.

다만, 제안 방식이 기존 방식과 대비하여 정확도 향상을 보였으나, 유사 의미 필드 입력인 '9/3' 조건에서의 전반적인 정확도는 낮은 수준에 머물렀다. 이는 유사 필드 3개인 지명 필드들에 "East Delhi (Mayur Vihar)"와 같은 비표준적 또는 비공식 지명 표현이 포함되어 있어, 성능 저하 요인으로 작용한 것으로 보인다. 특히 "Italians*"와 같이 행정 구역이 아닌 주석적 표현이 다수 존재함에 따라, 모델이 실제 지역 단위를 정확히 식별하는 데 혼란을 겪은 것으로 분석된다.

## 4.4 구성 요소별 기여도 분석

본 연구는 맥락 정보가 부족한 비정형 의료데이터를 LLM에 기반한 프롬프트 엔지니어링 기법과 가우시안 확률 분포를 결합하여 구조화하는 것을 목표로 한다. 이에 따라, 기존 방식에 다양한 제약과 정보를 추가한 제안 방식의 효과를 분석하였다. 기존 방식은 단순히 필드명만 제시한 상태에서 매핑을 요구했지만, 제안 방식은 기존에서 개선된 (1) 입력 필드의 메타데이터, (2) 문자열 필드의 샘플 데이터, (3) 수치형 필드의 가우시안 확률 정보, (4) 추가적인 제약 조건들로 구성되어 있다. 이를 통해 문자열은 표현 유사도, 수치는 확률 기반 적합도를 기준으로 정밀하고 일관된 필드 매핑을 수행한다.

이러한 추가 조건들이 실제로 의료데이터 필드 매핑의 성능 향상에 얼마나 핵심적으로 작용하는지를 확인하기 위하여, 기존 방식에 제안 방식의 각 조건을 표 2에서처럼 개별 형태로 추가한 것뿐만 아니라, 표 3과 표 4와 같이 조합 형태로 추가하여 실험하였다. 조합 실험은 다양한 구성 요소들의 조합이 가능하지만, 이들 간의 상호작용 여부와 상호작용 시 제안 방식에 미치는 효과를 고려하여 조합을 결정해야 한다. 실험 목적에 따라 구성 요소의 특성과 상호작용 가능성을 분석하여, 부정적 영향을 줄 수 있는 조합은 제외하고, 성능 향상에 기여할 가능성이 높은 주요 조합들을 선별하여 실험을 진행하였다. 이를 통해 단일 조건이 주는 영

향뿐 아니라, 조건 간의 상호작용 효과도 함께 평가할 수 있다.

본 연구는 올바르지 않은 필드에 매핑하는 것을 제외한, 존재하지 않는 입력 데이터를 생성하여 필드에 대응하는 것과 같은 치명적 오류를 환각으로 분류하였다. 그 결과, 기존 방식 평균 1.84%에서 제안 방식 0.25%로 환각을 감소시켰으며, 이는 프롬프트 설계를 통한 LLM 환각 억제 가능성을 시사한다.

표 2. LLM 기반 필드 매핑을 위한 프롬프트 개별 요소별 성능 기여도 분석
Table 2. Effectiveness of individual prompt components for structuring unstructured inputs with LLMs

| Dataset | COVID | | Diabetes | |
|---|---|---|---|---|
| Input condition | 9/9 | 9/5 | 7/7 | 7/3 |
| BASE | 0.747<br>0.747<br>0.387 | 0.638<br>0.599<br>0.005 | 0.649<br>0.649<br>0.161 | 0.353<br>0.580<br>0.015 |
| (1) Field information | 0.869<br>0.869<br>0.560 | 0.729<br>0.701<br>0.195 | 0.656<br>0.656<br>0.189 | 0.433<br>0.625<br>0.076 |
| (2) Handling of categorical data | 0.967<br>0.967<br>0.790 | 0.795<br>0.777<br>0.269 | 0.693<br>0.693<br>0.218 | 0.468<br>0.640<br>0.076 |
| (3) Probability data | - | - | 0.948<br>0.948<br>0.797 | 0.771<br>0.860<br>0.576 |
| (4) Constraints | 0.860<br>0.860<br>0.493 | 0.765<br>0.742<br>0.173 | 0.610<br>0.610<br>0.063 | 0.378<br>0.592<br>0.031 |

제안 방식의 구성 요소별 기여도를 분석한 결과, 필드 설명, 문자열 샘플 데이터, 제약 조건, 확률 정보가 COVID와 Diabetes 데이터 세트에서 모두 기존 방식의 성능을 개선하였다. 특히 개별 요소 실험에서는 문자열 필드의 샘플 데이터와 수치형 필드의 가우시안 확률 정보를 제공했을 때 가장 두드러진 성능 개선이 나타났다. 이를 통해, 문자열 필드의 샘플 데이터와 수치형 필드의 가우시안 확률 정보를 제공하는 제안 방식이, LLM의 사전 학습되지 않은 정보에 대한 입력 매핑 한계를 보완하고 입력값과 필드 간의 유사도를 효과적으로 추론함을 알 수 있다.

표 3. LLM 기반 필드 매핑을 위한 프롬프트 구성 요소 조합에 따른 성능 기여도 분석
Table 3. Contribution analysis of prompt component combinations for LLM-based field mapping

| Dataset | COVID | | Diabetes | |
|---|---|---|---|---|
| Input condition | 9/9 | 9/5 | 7/7 | 7/3 |
| BASE | 0.747<br>0.747<br>0.387 | 0.638<br>0.599<br>0.005 | 0.649<br>0.649<br>0.161 | 0.353<br>0.580<br>0.015 |
| (1) + (4) | 0.900<br>0.900<br>0.569 | 0.757<br>0.733<br>0.184 | 0.657<br>0.657<br>0.134 | 0.438<br>0.629<br>0.089 |
| (2) + (4) | 0.974<br>0.974<br>0.858 | 0.935<br>0.931<br>0.686 | 0.661<br>0.661<br>0.139 | 0.451<br>0.638<br>0.1131 |
| (3) + (4) | - | - | 0.941<br>0.941<br>0.773 | 0.829<br>0.898<br>0.678 |

조합 실험에서는 문자열 정보(의미 또는 샘플)와 제약 조건을 제공할 경우, 상호보완적으로 단독 구성보다 더 높은 성능을 보였다. 다만, Diabetes(7/3)와 같이 수치형 필드가 상대적으로 많은 혼합형 구조에서는 문자열 정보가 오히려 정확도 하락을 유발할 수 있음을 확인하였다.

표 4. 입력 유형별 최적의 프롬프트 구성 비교
Table 4. Optimal prompt configurations by input type

| Dataset | COVID | | Diabetes | |
|---|---|---|---|---|
| Input condition | 9/9 | 9/5 | 7/7 | 7/3 |
| String unification strategy | 0.977<br>0.977<br>0.865 | 0.946<br>0.944<br>0.751 | 0.672<br>0.672<br>0.147 | 0.479<br>0.652<br>0.126 |
| Proposed | 0.969<br>0.969<br>0.822 | 0.937<br>0.934<br>0.710 | 0.952<br>0.952<br>0.823 | 0.881<br>0.940<br>0.802 |

최종적으로 문자열 필드 중심의 상황에서는 확률 정보가 없이 구성된 문자열 통합 프롬프트가 가장 우수하였으며, 수치형과 문자열이 혼합된 현실적인 조건에서는 모든 요소를 포함한 제안 방식의 정확도가 가장 높고 안정적으로 유지되었다.

### 4.5 다양한 LLM에서 견고성 검증

LLM별 성능 비교 실험은 표 5와 같이 실제 입력 환경과 유사한 조건에서 OpenAI, Anthropic, Google DeepMind의 총 5가지의 모델을 통해 제안 방식의 일반화 능력을 평가하고자 설계하였다. 입력 데이터는 순서를 뒤섞은 문자열과 수치형 데이터가 혼합된 구조로 되어 있다. 실험은 두 가지 입력 조건으로 나뉘며, 7/7 평가 필드 설정은 모든 필드가 완전하게 입력되는 이상적 입력 조건, 7/3 평가는 사용자가 전체 필드 중 일부만 입력하는 부분 입력 조건으로 구성된다. 7/3 조건은 실제 환경에서의 불완전한 사용자 입력을 시뮬레이션하며, 특히 두 실수형 필드는 가우시안 분포가 겹치는 복잡한 형태로 구성되어 있어 모델의 추론 성능을 검증할 수 있다.

표 5. 다양한 LLM에서 견고성 검증
Table 5. Cross-model evaluation of prompt robustness

| Dataset | Diabetes | | | |
|---|---|---|---|---|
| Method | Base | | Proposed | |
| Inputs / Model | 7/3 | 7/7 | 7/3 | 7/7 |
| ChatGPT GPT-4o | 0.353<br>0.580<br>0.015 | 0.649<br>0.649<br>0.160 | 0.881<br>0.940<br>0.802 | 0.952<br>0.952<br>0.823 |
| Claude 3.5 Sonnet | 0.364<br>0.561<br>0.002 | 0.597<br>0.597<br>0.007 | 0.893<br>0.954<br>0.839 | 0.966<br>0.966<br>0.884 |
| Gemini 2.5 Pro | 0.296<br>0.510<br>0.000 | 0.628<br>0.628<br>0.065 | 0.894<br>0.954<br>0.842 | 0.954<br>0.954<br>0.839 |
| Gemini 2.5 flash | 0.276<br>0.499<br>0.027 | 0.649<br>0.649<br>0.150 | 0.894<br>0.954<br>0.842 | 0.954<br>0.954<br>0.839 |
| Gemini 2.0 flash | 0.296<br>0.523<br>0.000 | 0.523<br>0.523<br>0.013 | 0.826<br>0.925<br>0.739 | 0.950<br>0.950<br>0.826 |

실험 결과, 표 6에서 볼 수 있듯이 기존 방식은 입력값을 단순 필드명에 대응하는 구조이기 때문에 정확도의 변동 폭이 컸다. 특히 전체 필드 정확도와 입력 데이터 정확도가 최대 64.92%(GPT-4o), 최저 성능 27.65%(Gemini- 2.5-flash)로 각 모델별로 성능이 불안정하게 나타났다. 반면, 제안 프롬프트는 구조화된 제약 조건과 확률 분포를 조합한 최적화 매핑 전략을 통해 최대 성능 96.69%(Claude)와 최저 성능 82.63% (Gemini-2.0-flash)의 정확도를 기록하며 기존 대비 안정적인 높은 성능을 보였다.

표 6. 제안 방식의 성능 향상 폭 (행 정확도 기준)
Table 6. Improvement from BASE to PROPOSED prompts

| Inputs / Model | Diabetes 7/3 | Diabetes 7/7 |
|---|---|---|
| Claude 3.5 Sonnet | 0.002 → 0.839<br>**(+0.837)** | 0.007 → 0.884<br>**(+0.877)** |
| Gemini 2.5 Pro | 0.0 → 0.842<br>**(+0.842)** | 0.065 → 0.839<br>**(+0.774)** |
| ChatGPT GPT-4o | 0.015 → 0.802<br>**(+0.787)** | 0.16 → 0.823<br>**(+0.663)** |
| Gemini 2.5 flash | 0.027 → 0.842<br>**(+0.815)** | 0.15 → 0.839<br>**(+0.689)** |
| Gemini 2.0 flash | 0.0 → 0.739<br>**(+0.739)** | 0.013 → 0.826<br>**(+0.813)** |

특히 행 단위 정확도에서는 기존 방식 대비 매우 큰 개선 폭이 나타났으며, Gemini 2.0의 경우 행 정확도가 73.9% 향상되었다. 모든 모델에서 제안 방식은 압도적인 행 정확도 향상을 보였다. 이로써 제안 방식은 필드 유형에 상관없이 높은 정확도로 일반화가 가능함을 실증하였다.

## V. 결 론

의료 현장에서 발생하는 비정형 데이터, 특히 상담 기록, 응급 대응, 현장 진료와 같은 환경에서 생성되는 입력은 제한된 시간 내 핵심 내용만 입력해야 하는 중요한 특성이 있다. 본 연구는 이처럼 문맥이 부족한 입력 데이터를 구조화하기 위해, 가우시안 확률과 LLM을 결합한 비정형 의료데이터 입력 시스템을 제안하였다. 실험 결과, 제안 방식은 특히 혼합형 입력 환경에서 수치형 필드의 확률 기반 추론을 통해 일관된 높은 성능을 유지하였다. 이는 자연어 데이터 입력 시스템을 최적화하여 의료 데이터 품질을 향상하고, 의료진의 입력 부담을 완화하여 줄 것으로 기대된다.

필드에 매핑하는 확률 정보가 정량적으로 제공되지 않을 경우, LLM은 환각을 발생시킬 가능성이 증가한다. 이에 따라 본 연구는 원 가우시안 분포를 가정하여 문제를 해결했으나, 멀티모달 가우시안과 같은 복잡한 분포에 대응하기 위한 개선이 필요하다. 또한, 틀린 데이터가 입력으로 들어올 경우 일부 성능 저하가 발생하였다. 이에 따라, 향후 연구

에서는 분포 복잡도와 입력 노이즈 데이터 처리에 대한 알고리즘을 개선하고자 한다.

## References


[1] V. Ntinopoulos, H. R. C. Biefer, I. Tudorache, N. Papadopoulos, D. Odavic, P. Risteski, A. Haeussler, and O. Dzemali, "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation", BMJ Health Care Inform., Vol. 32, No. 1, pp. e101139, Jan. 2025. https://doi.org/10.1136/bmjhci-2024-101139.

[2] S. Pavuluri, R. Sangal, J. Sather, and R. A. Taylor, "Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics," BMJ Health Care Inform., Vol. 31, No. 1, pp. e101120, Aug. 2024. https://doi.org/10.1136/bmjhci-2024-101120.

[3] G. K. Thakur, A. Thakur, N. Khan, and H. Anush, "The role of natural language processing in medical data analysis and healthcare automation", Proc. Int. Conf. Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India, pp. 1-5, Apr. 2024. https://doi.org/10.1109/ICKECS61492.2024.10616749.

[4] T. V. Nguyen, et al., "Efficient automated error detection in medical data using deep-learning and label-clustering", Scientific Reports, Vol. 13, No. 19587, Nov. 2023. https://doi.org/10.1038/s41598-023-45946-y.

[5] S. Karthikeyan, A. G. S. de Herrera, F. Doctor, and A. Mirza, "An OCR post-correction approach using deep learning for processing medical reports", IEEE Trans. Circuits Syst. Video Technol., Vol. 32, No. 5, pp. 2574-2581, May 2022. https://doi.org/10.1109/TCSVT.2021.3087641.

[6] S. Koo and K. Choi, "Unstructured Medical Data Entry System Using Speech Recognition and Generative AI", Proc. Korean Soc. Comput. Inf. Conf., Cheonan, Korea, Jan. 2025.

[7] A. Cho, I. K. Min, S. Hong, H. S. Chung, H. S. Lee, and J. H. Kim, "Effect of Applying a Real-Time Medical Record Input Assistance System With Voice Artificial Intelligence on Triage Task Performance in the Emergency Department: Prospective Interventional Study", JMIR Med. Inform., Vol. 10, No. 8, pp. e39892, Aug. 2022. https://doi.org/10.2196/39892.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions", ACM Trans. Inf. Syst., Vol. 43, No. 2, pp. 1-55, Mar. 2025. https://doi.org/10.1145/3703155.

[9] A. Pal, L. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical domain hallucination test for large language models", Proc. Conf. Comput. Natural Lang. Learn. (CoNLL), Singapore, pp. 314-334, Dec. 2023. https://doi.org/10.18653/v1/2023.conll-1.21.

[10] S. R. Kadam, P. Chavan, D. S. Rani, M. Vats, R. Thapliyal, and R. D. Chaudhari, "Increasing the approach towards healthcare informatics and data analytics in medical science", Proc. Int. Conf. Healthcare Innovations, Software and Engineering Technologies (HISET), Karad, India, pp. 286-288, Jan. 2024. https://doi.org/10.1109/HISET61796.2024.00089.

[11] I. Li, et al., "Neural natural language processing for unstructured data in electronic health records: a review", Comput. Sci. Rev., Vol. 46, pp. 1-24, Nov. 2022. https://doi.org/10.1016/j.cosrev.2022.100511.

[12] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval, "Security and privacy in electronic health records: A systematic literature review", J. Biomed. Inform., Vol. 46, No. 3, pp. 541-562, Jun. 2013. https://doi.org/10.1016/j.jbi.2012.12.003.

[13] E. Sezgin, S. Hussain, S. Rust, and Y. Huang,

"Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data", JMIR Form. Res., Vol. 7, pp. e43014, Mar. 2023. https://doi.org/10.2196/43014.

[14] I. C. Wiest, D. Ferber, J. Zhu, M. Sohrabi, F. Ji, and J. L. Raisaro, "Privacy-preserving large language models for structured medical information retrieval", npj Digit. Med., Vol. 7, pp. 257, Sep. 2024. https://doi.org/10.1038/s41746-024-01233-2.

[15] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, "Fine-tuning BERT for joint entity and relation extraction in Chinese medical text", Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, pp. 892-897, Nov. 2019. https://doi.org/10.1109/BIBM47256.2019.8983370.

[16] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, "An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study", JMIR Med. Inform., Vol. 12, pp. e55318, Apr. 2024. https://doi.org/10.2196/55318.

[17] X. Meng, et al., "The application of large language models in medicine: A scoping review", iScience, Vol. 27, No. 5, pp. 109713, Apr. 2024. https://doi.org/10.1016/j.isci.2024.109713.

[18] J. Maharjan, \et al., "OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models", Sci. Rep., Vol. 14, No. 14156, Jun. 2024. https://doi.org/10.1038/s41598-024-64827-6.

[19] Y. Hu, J. Chen, S. Li, and F. Wang, "Improving large language models for clinical named entity recognition via prompt engineering", Journal of the American Medical Informatics Association, Vol. 31, No. 4, Jan. 2024. https://doi.org/10.1093/jamia/ocad259.

[20] H. Zhao, H. Zhao, B. Shen, A. Payani, F. Yang, and M. Du, "Beyond single concept vector: Modeling concept subspace in LLMs with Gaussian distribution", pp. 1-29, Sep. 2024. https://doi.org/10.48550/arXiv.2410.00153.

[21] https://www.kaggle.com/datasets/dheerajmpai/covid19india. Original source: https://www.covid19india.org/. [accessed: Apr. 02, 2025].

[22] https://www.kaggle.com/datasets/minhajulislamakib/diabates. [accessed: Apr. 03, 2025].

## 저자소개

구 서 연 (Seoyon Koo)

2023년 3월 ~ 현재 : 강남대학교 인공지능융합공학부 학사과정
관심분야 : 인공지능, 자연어처리, 딥 러닝, 컴퓨터비전

최 권 택 (Kwon-Taeg Choi)

2006년 2월 : 연세대학교 컴퓨터공학과(공학석사)
2011년 2월 : 연세대학교 컴퓨터공학과(공학박사)
2016년 3월 ~ 현재 : 강남대학교 소프트웨어응용학부 교수
관심분야 : 가상현실, 증강현실, 모바일컴퓨팅, 기계학습, HCI