Bridging the Domain Divide: Supervised vs. Zero-Shot Clinical Section Segmentation from MIMIC-III to Obstetrics

Anonymous ACL submission

Abstract

Clinical notes contain vital patient information organized into sections such as "*History of Present Illness*" and "*Medications*". Recognizing these sections supports clinical decision-making, yet most existing segmentation approaches rely on supervised models trained on large public corpora (e.g., MIMIC-III), which may not generalize effectively to specialized domains such as obstetrics. In this paper, we advance clinical section segmentation through three key contributions: (1) we introduce a novel, de-identified dataset of obstetrics clinical notes; (2) we systematically evaluate transformer-based supervised models on both in-domain (MIMIC-III) and out-of-domain (obstetrics) data; and (3) we present the first head-to-head comparison with zero-shot large language models (Llama, Mistral, and Qwen). Our results show that while supervised models significantly outperform large language models (LLMs) on in-domain MIMIC-III data, their performance degrades substantially in the out-of-domain setting—where the best zero-shot LLM (Llama 3.3-70B-Instruct) surpasses all supervised baselines, even before applying our hallucination correction step. Once hallucinated section headers are corrected, zero-shot performance improves further, with three out of four LLMs outperforming the best supervised model, demonstrating the viability of zero-shot models for specialized clinical domains. These findings underscore the challenge of transferring models trained on broad public corpora to underexplored clinical subdomains and highlight the strong potential of zero-shot approaches when labeled data is scarce.

1 Introduction

Electronic Health Records (EHRs) are widely used in modern healthcare to provide detailed records of patient encounters and their interactions within the healthcare system (Holmes et al., 2021). EHR data often contain free text clinical notes, which are typically organized into sections such as "*Chief Complaint*" and "*History of Present Illness*". Accurately identifying these sections is crucial for downstream natural language processing (NLP) tasks, including entity extraction, information retrieval, and word sense disambiguation (Denny et al., 2008). However, clinical documentation is highly variable and often lacks standardized formatting. For example, the section "*Social History*" can appear as "*Social Hx*" or contain typographical errors like "*Chief Complain*" instead of "*Chief Complaint*" (Ganesan and Subotin, 2014). Such inconsistencies complicate rule-based solutions and motivate more robust machine-learning and deep-learning approaches.

Although transformer-based models trained on large public corpora (e.g., MIMIC-III (Johnson et al., 2016)) have shown promise for section segmentation, their domain adaptation capabilities remain uncertain. In particular, obstetrics represents a specialized clinical subdomain with unique documentation styles and limited annotated data. This variability is evident in how physician counseling information is structured, where the same content may be documented under diverse section headers such as "*Impression and Plan*", "*Assessment and Plan*", "*Assessment*", or even abbreviated formats like "*A/P*", "*A&P*" and "*A: P:*". For instance, in one of our obstetrics notes, the counseling section appears under an abbreviated heading, as shown in Figure 1.

Such variability often introduces out-of-vocabulary or unseen section headers that can significantly hinder deep learning approaches trained on more standardized data. Recent research suggests that LLMs can handle out-of-domain tasks through zero-shot or few-shot prompting, but it is unclear how they fare against traditional supervised methods in specialized domains.

To address these gaps, we make the following key contributions:

a/p:
"This is 28 y/o G6P4013 presents for repeat cesarean delivery and bilateral tuba ligation. admit to labor and delivery cbc, type and screen iv fluids consented for repeat cesarean delivery, possible hysterectomy, possible blood transfusion, and bilateral tubal ligation. Hgb 11.1 patient alst ate at 2100on <*DATE*> discussed with the team <*NAME*> MD Obstetrics and Gynecology Resident Attending note I saw patient prior to scheduled repeat c/section this morning, reviewed cerner and counseled patient r/b/a. Patient with history of prior c/s x 1 here for RCD and BTL. Labs this morning includes Hgb 11.1, plat 207, bl gp O+ve. Singlton cephalic, posterior placenta. Writtent consent obtained, will proceed when OR is ready. Patient voiced understanding of the plan. <*NAME*>, MD"

Figure 1: Counseling section from a sample obstetrics note (includes typographical errors).

1. **A Novel Obstetrics Dataset:** We introduce a de-identified dataset of 100 *History & Physical (H&P)* obstetrics notes, annotated in collaboration with a domain expert. This dataset provides a new and realistic benchmark for investigating section segmentation in underexplored clinical subdomains.

2. **Domain-Specific Evaluation of Supervised Models:** We assess whether transformer-based supervised models —originally trained on public datasets— can effectively generalize to obstetrics notes. By comparing them on both in-domain (MedSecId (Landes et al., 2022)) and out-of-domain (Obstetrics) data, we highlight the difficulties in transferring knowledge across clinical sub-specialties.

3. **Systematic Comparison With Zero-Shot LLMs:** We present the first head-to-head comparison of supervised transformer models and zero-shot LLMs (i.e., Llama, Mistral and Qwen) for clinical section segmentation. Our experiments reveal challenges (e.g., hallucinated section headers) as well as the potential benefits of zero-shot strategies, especially when annotated data are scarce.

The paper is organized as follows: we discuss related work in Sec. 2, our datasets in Sec. 3, and our proposed approaches in Sec. 4. We present our experimental results in Sec. 5, and conclude with future directions (Sec. 6) and limitations (Sec. 7).

2 Related Work

Before the emergence of advanced machine learning and NLP techniques, early approaches to clinical section segmentation primarily relied on rule-based methods. Denny et al. (2008), for instance, extracted candidate section header strings from a large corpus of "history and physical" (H&P) notes through pattern-based matching (e.g., detecting strings that end with punctuation or follow certain capitalization patterns). These candidates were then refined in collaboration with clinicians to build a terminology of section headers. However, purely rule-based methods tend to be inflexible and often fail to handle unexpected variations in unstructured, non-standardized text, which constitutes approximately 80% of the content in Electronic Health Records (EHR)(Kong, 2019).

To overcome the limitations of rule-based approaches, researchers proposed machine learning-based solutions for section segmentation, often framing it as a sequence-labeling task. Li et al. (2010) trained a Hidden Markov Model (HMM) on a clinical corpus to segment 15 predefined section types. Ganesan and Subotin (2014) employed an L1-regularized multi-class Logistic Regression model to classify each line of a clinical note into one of five roles—start header, continue header, start section, continue section, or footer—and then used the Viterbi algorithm (Forney, 1973) to determine the most probable sequence of labels.

More recent work has been grounded in transformer-based architectures. Zhang et al. (2022) presented a multi-task transformer model that simultaneously identifies section boundaries and assigns medically relevant labels. Saleh et al. (2024) leveraged BioClinicalBERT embeddings (Alsentzer et al., 2019) and framed section title and subtitle detection as a named entity recognition (NER) task. While not fully transformer-based, Landes et al. (2022) incorporated BERT embeddings as sentence-level representations, which were then processed using a BiLSTM model for sequence modeling and further refined with a Conditional Random Field (CRF) layer to enforce structured predictions across section boundaries.

Most of this research relies on large publicly available datasets such as MIMIC-III. Since producing high-quality annotated data is very resource-intensive, recent work has explored large language models (LLMs) for clinical section segmentation in zero-shot settings. Zhou and Miller (2024) evaluated several LLMs, both zero-shot and fine-tuned, across multiple corpora to assess their section-segmentation effectiveness; but these LLMs were still tested on common public datasets (e.g., MIMIC-III and i2b2 (Özlem Uzuner et al., 2011)) rather than more specialized clinical domains, such as obstetrics.

Hence, it remains unclear how well these ap-

proaches generalize to specialized and underutilized domains like obstetrics. Moreover, existing comparative studies often evaluate supervised methods solely against each other or LLM-based methods solely against each other, leaving a gap in cross-method comparisons in specialized settings.

In this paper, we address this gap by introducing a small yet informative dataset of obstetrics-related H&P narratives. We propose both supervised and zero-shot approaches for clinical section segmentation, and then evaluate their performance against each other on our newly collected dataset as well as on publicly available annotated corpora. This comprehensive evaluation sheds new light on how different models perform in a specialized medical domain.

3 Data

We utilize the publicly available MedSecId corpus introduced by Landes et al. (2022) to train and evaluate our models. MedSecId comprises 2,002 fully annotated clinical notes from MIMIC-III, specifically designed for clinical section segmentation. Additionally, we introduce a novel, de-identified dataset of 100 History & Physical (H&P) notes from 50 vaginal birth after cesarean (VBAC) and 50 repeat cesarean section (RCS) patients to evaluate model performance in obstetrics, an underrepresented clinical domain.

MedSecId spans five note types—*Discharge summary (1,254)*, *Physician (288)*, *Radiology (205)*, *Echo (198)* and *Consult (57)*—and segments each note into 50 section categories, plus a "<*none*>" label for text outside of any predefined section (Landes et al., 2022). To prepare the dataset, we first extracted section spans from MedSecId and split each note into section segments. Next, we tokenized the context of each section into lists of sentences using the NLTK sentence tokenizer (Bird et al., 2009), ensuring each sentence was correctly assigned to its respective section and appeared in the correct order.

Our obstetrics dataset was collected and managed using REDCap (Harris et al., 2009, 2019). Since the notes contained protected health information (PHI), we transferred them to a HIPAA-secure environment and applied automatic de-identification using the Spark NLP framework (Kocaman and Talby, 2021). This framework masked entities, including *NAME*, *LOCATION (address, city, zip code)*, *DATE*, *CONTACT (phone numbers, email addresses)*, and *ID (social security number, medical record number)*. We then manually reviewed all notes to ensure PHI removal was complete.

Due to annotation resource constraints, we focused on 100 high-quality, full-length H&P notes from distinct patients across both delivery groups (VBAC and RCS). Annotations were performed in collaboration with a midwifery domain expert. As with MedSecId, we split each section into sentences, using the dataset solely for evaluation due to its limited size. Unlike MedSecId, which includes a mix of general-purpose clinical note types, our dataset is obstetrics-specific and incorporates domain-relevant sections headers (e.g., "*Pregnancy History*", "*Gynecologic History*") that capture obstetric-specific content such as gravida/para notation and neonatal outcomes. Rather than normalizing to MedSecId's schema, we retained these specialized headers to preserve the narrative structure and semantics of obstetric H&P narratives.

To enable fair cross-domain evaluation, we excluded specialized headers when testing supervised models trained on MedSecId. This allows us to isolate the models' ability to generalize to a clinically distinct domain. Table A1 (Appendix) compares section headers across both datasets, highlighting shared and domain-specific labels. Table A2 (Appendix) presents the frequency distribution of section spans observed in the Obstetrics corpus.

4 Methodology

We explore two approaches for clinical section segmentation: Supervised Learning and Zero-shot Learning via LLMs. In this section, we provide an overview of both approaches; highlighting model architectures and design choices. Detailed implementation, training configurations, and computational resource usage are provided in Appendix A.

4.1 Supervised Learning Approach

We first develop a supervised approach to clinical section segmentation using pre-trained Transformer-based models, widely used in text classification and sequence labeling tasks (Vaswani, 2017; Devlin et al., 2019). While these models do not surpass existing systems such as Landes et al. (2022), they provide competitive and robust supervised baselines to evaluate the zero-shot LLM approach on this task. We fine-tune the models using two architectures:

1. **Transformer-based Classification:** Each line is treated as an independent input and classified into one of the predefined section headers.

2. **Transformer + CRF:** A Conditional Random Field (CRF) layer is added on top of the Transformer to model label dependencies between consecutive lines, framing the task as sequence labeling.

4.1.1 Transformer-based Section Segmentation

We approach section segmentation as a 51-way classification task (including the label *"<none>"*) using an IO-like encoding scheme: lines within labeled sections are tagged as *"I_section_name"*, while lines outside any labeled section are tagged as *"<none>"* (Landes et al., 2022). Throughout this work, we use the terms *"line"* and *"sentence"* interchangeably, as each unit in our dataset corresponds to a single textual span separated by newlines in clinical notes. We experiment with BERT-base, a widely used Transformer model pre-trained on general-domain English corpora (Devlin et al., 2019), and three models trained on biomedical text. BioBERT (Lee et al., 2020) extends BERT via further pretraining on PubMed abstracts and PubMed Central (PMC) articles. BiomedBERT (formerly PubMedBERT) (Gu et al., 2021) is trained from scratch exclusively on PubMed abstracts, making it fully domain-specialized. GatorTron-base (Yang et al., 2022) is trained on a diverse corpus comprising de-identified clinical notes from a university hospital, PubMed articles, and Wikipedia, totaling 90 billion words. We exclude models primarily trained on MIMIC-III (e.g., BioClinicalBERT (Alsentzer et al., 2019)) to avoid evaluation bias.

Line-Level Representation We represent clinical notes as sequences of independent lines (rather than full-length notes) to comply with Transformer token limits and reduce computational overhead. Each line is treated as a separate example, capturing local context without modeling sequential dependencies. Consequently, we flatten each note—originally a list of labeled lines—into a dataset of individual line-label pairs. Because the model does not leverage inter-line context, we perform train-test splitting at the line level, consistent with the model's independence assumption and eliminating the need to preserve note boundaries. This process yields 175,703 lines from 2,002 clinical notes, with 80% (140,140 lines) used for training and 20% (35,563 lines) for evaluation. While this setup ignores document-level structure, it provides a fair supervised baseline for comparison with zero-shot LLMs, which —despite accessing the full note—do not explicitly model label transitions or structured dependencies across lines.

Token Length Analysis Before tokenization, we analyzed the distribution of token lengths per line. Approximately 97% of lines (across all models) contained fewer than 100 subword tokens. We therefore truncate each line to 100 tokens to optimize memory usage and format inputs using the standard HuggingFace (Wolf et al., 2020) convention: *input_ids* and *attention_mask* for training.

Training configurations, hyperparameters, and evaluation metrics are provided in Appendix A.3.

4.1.2 Transformer + CRF based Section Segmentation

Unlike the line-level approach in Section 4.1.1, we retain note-level structure to model sequential dependencies between lines. Each note is treated as a single training instance, allowing the CRF layer to learn label transitions (e.g., from *"History of Present Illness"* to *"Review of Systems"*).

Custom Collator and Data Preparation To accommodate varying note lengths, we implement a custom collator for note-level batching:

- **Dynamic Line Dimensions:** For each batch, let L be the maximum number of lines among notes; each line is truncated or padded to a maximum token length S.
- **Batch-Size Constraint:** To preserve note-level context and avoid inefficient padding across variable-length sequences, we set the batch size to $B = 1$, which simplifies training and reduces GPU memory usage while retaining the CRF's ability to model label transitions.
- **Final Tensor Shape:** Each note is arranged into a tensor of shape (B,L,S), preserving full note structure. This allows the CRF to model label transitions across all lines within a note.

Model Architecture We use the same Transformer backbones in Section 4.1.1 (BERT-base, BioBERT, BiomedBERT and GatorTron-base) combined with *torchcrf*, a CRF library for PyTorch (Paszke et al., 2019).

```
<|begin_of_text|><|start_header_id|>system<|
end_header_id|>
You are a clinical assistant specializing in
segmenting clinical notes.

<|eot_id|><|start_header_id|>user<|end_header_id
|>
Your task is to assign section headers to each
line of a clinical note. Most of the section
headers will likely span multiple lines, so
headers should be assigned sequentially and
consistently.

Clinical Note:
{enumerated clinical note text}

Select the most appropriate section header for
each line from the following options:
{string of 30 potential headers}

Return your answer as a list of section headers,
 one for each line, in the same order.

Example Output:
Line 0: <none>
Line 1: imaging
Line 2: <none>
Line 3: chief-complaint
Line 4: history-of-present-illness
Line 5: history-of-present-illness
Line 6: history-of-present-illness
Line 7: history-of-present-illness
Line 8: history-of-present-illness
Line 9: history-of-present-illness
...

The output must contain **exactly the same
number of lines** as the clinical note, i.e
number of lines SHOULD BE EQUAL TO {number of
note lines}

<|eot_id|><|start_header_id|>assistant<|
end_header_id|>
Section Headers:
```

Listing 1: Zero-shot prompt snippet for Llama Instruct models

The Transformer + CRF architecture consists of the following steps:

1. **Flatten Input:** We reshape *(B,L,S)* to *(B x L,S)* so each line can be processed independently by the Transformer.
2. **Contextual Embeddings:** We extract the *[CLS]* representation for each line.
3. **Logit Projection:** We apply a linear layer to project contextual embeddings into logits of shape *(B x L, num_labels)* for each section label where *num_labels = 51*.
4. **CRF Reshaping:** We reshape logits back to *(B, L, num_labels)*, so the CRF can model line-level transitions across the entire note.
5. **Viterbi Decoding:** At evaluation, we apply Viterbi decoding (Forney, 1973) to obtain the most likely label sequence for each note.

Training hyperparameters and evaluation details are provided in Appendix A.4.

4.2 Zero-Shot Learning via LLMs

Unlike supervised approaches that require labeled training data, we explore zero-shot learning for clinical section segmentation using pre-trained LLMs. Our primary goal is to evaluate whether instruction-tuned LLMs—without domain-specific fine-tuning—can accurately assign section labels by leveraging general contextual understanding.

Model Selection We selected four instruction-tuned, open-source LLMs for evaluation: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama 3.1-8B-Instruct (Touvron et al., 2023), Qwen-2.5-32B-Instruct (Yang et al., 2024), and Llama 3.3-70B-Instruct (Touvron et al., 2023). These models support extended context windows (32k-128k tokens), enabling full-note inference without truncation. Their varied sizes (7B-70B) allow us to assess how model scale affects performance in long-form clinical narratives.

Although our dataset is de-identified, real-world clinical documents often contain protected health information (PHI). Closed-source models such as GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023) can pose security and privacy risks, as they require sending user data to third-party servers and thus increase the likelihood of unauthorized access or misuse of sensitive information (Kim et al., 2025). In contrast, open-source models can be deployed on-premises, offering a more secure pathway for integrating LLMs into clinical workflows. This practical consideration further motivates our use of open-source models.

Prompt Engineering We adopt an instruction-style prompt to assign section labels to each line in a clinical note, without any task-specific fine-tuning. All four models are chat-based and support system/user prompting. The Llama models use a unified template with explicit system and user roles (see Listing 1); for Mistral and Qwen, we adapt the prompt format to match their respective syntax conventions (e.g., *[INST]* or <|*im_start*|>). We designate the model as a "clinical assistant specializing in segmenting clinical notes" and provide it with a list of valid section labels. Each line in the note is numbered (e.g., "1. line1," "2. line2") to

Model	MP	MR	MF1	wP	wR	wF1
Supervised Models						
$BERT_{base}$	0.71	0.67	0.68	0.78	0.78	0.77
BioBERT	0.72	0.68	0.68	0.78	0.78	0.77
BiomedBERT	0.72	0.69	0.68	0.79	0.79	0.78
$GatorTron_{base}$	0.73	0.69	**0.69**	0.80	**0.80**	0.78
$BERT_{base}$+CRF	0.72	0.69	0.68	0.79	0.77	0.77
BioBERT+CRF	0.74	0.69	0.68	0.79	0.77	0.76
BiomedBERT+CRF	**0.75**	**0.70**	**0.69**	0.79	0.79	0.78
$GatorTron_{base}$+CRF	0.74	0.65	0.67	**0.81**	**0.80**	**0.79**
Zero-Shot Models (Results with Hallucinations)						
Mistral-7B-$Instruct_{raw}$	0.03	0.02	0.02	0.54	0.17	0.22
Llama 3.1-8B-$Instruct_{raw}$	0.17	0.19	0.14	0.70	0.48	0.52
Qwen-2.5-32B-$Instruct_{raw}$	0.25	0.23	0.21	0.68	0.45	0.49
Llama 3.3-70B-$Instruct_{raw}$	0.23	0.29	0.23	0.76	0.61	0.64
Zero-Shot Models (Results after Mitigating Hallucinations)						
Mistral-7B-$Instruct_{corrected}$	0.21	0.19	0.16	0.41	0.20	0.23
Llama 3.1-8B-$Instruct_{corrected}$	0.46	0.54	0.39	0.70	0.49	0.52
Qwen-2.5-32B-$Instruct_{corrected}$	0.47	0.48	0.41	0.61	0.46	0.49
Llama 3.3-70B-$Instruct_{corrected}$	0.47	0.61	0.48	0.73	0.62	0.64

Table 1: Performance metrics on MedSecId: MP = macro precision, MR = macro recall, MF1 = macro F1; wP = weighted precision, wR = weighted recall, wF1 = weighted F1.

Model	MP	MR	MF1	wP	wR	wF1
Supervised Models						
$BERT_{base}$	0.66	0.37	0.39	0.76	0.43	0.47
BioBERT	0.54	0.39	0.39	0.75	0.45	0.48
BiomedBERT	0.61	0.39	0.40	0.76	0.46	0.49
$GatorTron_{base}$	**0.73**	0.48	0.49	0.85	0.58	0.61
$BERT_{base}$+CRF	0.68	0.49	0.47	0.80	0.61	0.62
BioBERT+CRF	0.55	0.45	0.43	0.74	0.59	0.57
BiomedBERT+CRF	0.56	0.51	0.50	0.76	0.65	0.66
$GatorTron_{base}$+CRF	0.65	0.51	0.49	0.79	0.65	0.65
Zero-Shot Models (Results with Hallucinations)						
Mistral-7B-$Instruct_{raw}$	0.05	0.04	0.04	0.72	0.45	0.52
Llama 3.1-8B-$Instruct_{raw}$	0.35	0.33	0.32	0.84	0.70	0.74
Qwen-2.5-32B-$Instruct_{raw}$	0.34	0.39	0.34	0.88	0.79	0.83
Llama 3.3-70B-$Instruct_{raw}$	0.61	0.59	0.58	**0.90**	**0.85**	**0.86**
Zero-Shot Models (Results after Mitigating Hallucinations)						
Mistral-7B-$Instruct_{corrected}$	0.38	0.45	0.37	0.56	0.47	0.49
Llama 3.1-8B-$Instruct_{corrected}$	0.58	0.56	0.54	0.83	0.71	0.74
Qwen-2.5-32B-$Instruct_{corrected}$	0.61	**0.71**	0.61	0.88	0.82	0.84
Llama 3.3-70B-$Instruct_{corrected}$	0.70	0.67	**0.67**	**0.90**	**0.85**	**0.86**

Table 2: Performance metrics on Obstetrics: MP = macro precision, MR = macro recall, MF1 = macro F1; wP = weighted precision, wR = weighted recall, wF1 = weighted F1.

ensure independent prediction while preserving sequence order. This structure allows the model to reference neighboring lines during inference, enabling implicit modeling of section transitions. To clarify output formatting (rather than teach section content), we include a single one-shot-style example (e.g., "Line 0: <none>, Line 1: imaging"). This preserves a near-zero-shot setup, relying solely on the model's pretrained knowledge to infer appropriate section labels. See Appendix A.6 for inference details.

Post Processing We parse model outputs using regular expressions to isolate predicted section headers (e.g., removing "Line 0:" prefixes). Predictions are evaluated against gold labels in the MedSecId and Obstetrics datasets using precision, recall, F1, and hallucination rate—defined as the percentage of lines assigned to non-existent section headers. To reduce label fragmentation, we normalize semantically equivalent labels. In collaboration with the midwifery expert who assisted with annotations, we consolidated *impression-and-plan* and *plan* into the standardized label *assessment-and-plan*, following clinical convention. This label aligns with terminology adopted in prior clinical section segmentation work (Denny et al., 2009; Landes et al., 2022), supporting its use as a canonical form for evaluation.

5 Experiments

5.1 Evaluation and Experimental Setup

We evaluate the performance of both our supervised models and zero-shot LLMs on two datasets: MedSecId and Obstetrics. Since the supervised models were trained on MedSecId, we excluded the training portion to avoid evaluation bias. Specifically, we removed 80% (1,601 notes) of the original MedSecId corpus used for training. From the remaining 401 notes, we further excluded those with more than 100 lines to maintain a tractable sequence length for evaluation, resulting in a final subset of 251 notes comprising 11,528 lines. For the Obstetrics dataset, we used all 100 notes (5,352 lines).

5.2 Hallucinations in Zero-Shot LLMs

Despite receiving clear instructions, all four zero-shot models—Mistral-7B-Instruct-v0.3, Qwen-2.5-32B-Instruct, Llama 3.1-8B-Instruct, and Llama 3.3-70B-Instruct—exhibited hallucinations during inference by generating section headers not present in the ground truth. We define hallucination in this context as the assignment of a section header that does not appear in the predefined list of valid labels. For example, Mistral frequently labeled *substance-abuse* as a distinct section, although it should be subsumed under the broader *social history*. Such mislabeling risks fragmenting semantically related

content, potentially compromising clinical workflows.

As shown in Table 3, hallucination rates varied across models, with Mistral producing the highest rates on both datasets (22.21% for MedSecId; 17.64% for Obstetrics), followed by Qwen, Llama 3.1-8B and Llama 3.3-70B. Interestingly, this ranking diverges from those reported in general-domain hallucination benchmarks (e.g., Hughes et al. (2023)), underscoring the importance of evaluating model reliability within the specific context of clinical tasks. These findings suggest that hallucination behavior is highly sensitive to domain, task formulation, and prompting strategy—and cannot be reliably extrapolated from general-purpose evaluations. Further research is needed to address the factual consistency of LLM outputs in the healthcare domain (Nori et al., 2023).

To better characterize model behavior, Table A3 (Appendix) lists the five most frequently hallucinated section headers for each model on the Obstetrics dataset.

Model	HL	TL	H%	HS
MedSecId				
Mistral-7B-Instruct-v0.3	2,560	11,528	22.21%	433
Llama 3.1-8B Instruct	452	11,528	3.92%	89
Qwen-2.5-32B-Instruct	497	11,528	4.31%	54
Llama 3.3-70B Instruct	404	11,528	3.50%	57
Obstetrics				
Mistral-7B-Instruct-v0.3	944	5,352	17.64%	136
Llama 3.1-8B Instruct	115	5,352	2.15%	19
Qwen-2.5-32B-Instruct	177	5,352	3.31%	23
Llama 3.3-70B Instruct	5	5,352	0.09%	4

Table 3: Hallucination Analysis on MedSecId and Obstetrics. *HL* = number of hallucinated lines; *TL* = total lines; *H%* = hallucination rate; *HS* = number of hallucinated sections types are not in the original label set.

Mitigating Hallucinations To mitigate hallucinations, we implemented a post-processing correction step using GPT-4o (Achiam et al., 2023). For each hallucinated section header—i.e., one not present in the predefined set of valid labels—we prompted GPT-4o to map it to the most semantically appropriate label from the valid list. Because this task involved only generic section names (e.g., *labs*, *social-history*) and no patient-level content, we could safely use an API-based model without violating privacy constraints. We selected GPT-4o over embedding-based heuristics (e.g., Sentence-BERT cosine similarity (Reimers and Gurevych, 2019)) due to its superior contextual reasoning, particularly for ambiguous or sparsely descriptive headers.

While some edge cases remain challenging, this procedure substantially reduced the number of non-standard predictions and improved alignment with the target schema. Importantly, the correction accuracy may underestimate true semantic alignment: in some cases, hallucinated headers (e.g., *ultrasound*) may be semantically closer to a different valid label (e.g., *imaging*) than to the gold-standard label used for evaluation (e.g., *review-of-systems*). In such cases, lower correction scores may reflect initial label misalignment rather than a failure of the mapping strategy. We report the post-correction mapping results in Table A4 (Appendix). The prompt used for GPT-4o hallucination correction is provided in Listing 2 (Appendix).

5.3 Qualitative Error Analysis for LLM Predictions

After correcting hallucinations, we analyzed remaining section labeling errors through a qualitative evaluation of outputs from the best-performing model, Llama 3.3-70B-Instruct. To scale this process, we employed the same model in an LLM-based classification framework to automatically assign errors to one of four categories: (1) *Omission*—the model incorrectly predicted <*none*> for a span that should have received a valid label; (2) *Label confusion*—the predicted label was clearly incorrect relative to the gold label; and (3) *Valid local interpretation*—the predicted label differed but was semantically justifiable given the local span; and (4) *Other*—ambiguous or uncategorizable cases.

The classification prompt handled categories 2–4, while *omission* was identified separately using rule-based logic. The prompt is shown in Listing 3 (Appendix). Figure 2 summarizes error type distributions, and Table A5 provides representative examples; both appear in the Appendix.

5.4 Results and Discussion

Tables 1 and 2 present the performance of all models on the MedSecId and Obstetrics datasets, respectively. For zero-shot LLMs, we report both raw and corrected results to highlight the impact of hallucination mitigation. Notably, post-correction macro F1 scores increase by 9% to 33%, confirming that hallucinations are a major source of error in zero-shot predictions.

As expected, supervised models outperform zero-shot LLMs on MedSecId due to their direct

training on that dataset. Among the supervised models, performance is largely comparable across Transformer variants. However, the addition of a CRF layer yields modest but non-negligible gains for some models. Specifically, macro F1 scores for BERT-based models improve by 4% to 10% with CRF integration, suggesting that modeling inter-line dependencies offers measurable benefits. In contrast, GatorTron shows no improvement with a CRF layer, indicating that larger models may already encode sufficient contextual information for accurate line-level predictions. Meanwhile, zero-shot LLMs display large discrepancies between macro and weighted F1 scores due to macro F1's sensitivity to hallucinated labels. Once hallucinated headers are corrected, spurious labels are mapped to valid ones, leading to substantial improvements in macro scores.

While supervised models maintain a strong lead over zero-shot LLMs on MedSecId, they struggle to generalize to the newly introduced Obstetrics dataset. This result suggests that models trained on large public corpora, such as MIMIC, may not transfer effectively to narrower clinical subdomains like Obstetrics. Although GatorTron-base initially outperforms the other supervised models, the addition of a CRF layer allows others—particularly BioMedBERT—to close the gap or even surpass it. Notably, BioMedBERT+CRF outperforms GatorTron+CRF by approximately 1% in both macro and weighted F1 on the Obstetrics dataset.

Interestingly, zero-shot LLMs perform relatively better on Obstetrics, partly due to the smaller label space (28 vs. 51). To quantify the robustness of model performance across notes, we report 95% confidence intervals over per-note macro and weighted F1 scores (Appendix A.8). Llama 3.3-70B-Instruct achieves the highest overall performance, outperforming all supervised baselines. To assess the consistency of this advantage, we conducted Wilcoxon signed-rank tests on per-note macro F1 scores. Even in its hallucinated form, Llama 3.3-70B-Instruct significantly outperforms the strongest supervised model ($p < 4.88 \times 10^{-17}$), with further gains after correction ($p < 3.75 \times 10^{-17}$). These results suggest that the LLM's advantage reflects robust generalization, not merely post-hoc label correction. While Llama 3.3-70B-Instruct slightly outperforms Qwen-2.5-32B-Instruct on average, the difference is not statistically significant ($p \approx 0.11$), indicating comparable performance between the two strongest zero-shot models.

Overall, these findings highlight the flexibility of zero-shot LLMs in adapting to novel domains without requiring additional annotation or fine-tuning. While supervised Transformer models remain state-of-the-art for in-domain tasks, instruction-tuned LLMs—especially when paired with simple hallucination correction—offer a statistically robust and scalable alternative for clinical NLP in under-explored subdomains.

6 Conclusions and Future Work

We addressed clinical section segmentation in a specialized obstetrics domain by introducing a curated dataset of obstetrics-related H&P narratives. We evaluated both supervised and zero-shot LLM approaches on this dataset and existing public corpora. While supervised models perform well in-domain, they struggle to generalize to unfamiliar clinical subdomains. In contrast, zero-shot LLMs demonstrate greater adaptability, particularly when domain-specific fine-tuning is unavailable.

Despite these advances, several challenges remain. First, our dataset's limited size may not capture the full variability of obstetrics documentation. Second, although zero-shot LLMs reduce reliance on labeled data, they remain prone to domain-inconsistent predictions, including hallucinated section headers and omissions of clinically important spans. These issues are especially concerning in specialized domains, where mislabeling critical content undermines reliability and interpretability.

Future work includes expanding the dataset to cover a wider range of conditions, procedures, and patient profiles, improving clinical diversity. We also aim to explore further LLM adaptation strategies, such as few-shot learning and parameter-efficient fine-tuning (PEFT), to more effectively tailor models to specialized domains while retaining computational efficiency (Han et al., 2024). Finally, integrating domain knowledge bases or medical ontologies may enhance performance and interpretability by guiding segmentation and label assignment. These efforts aim to support the development of robust, domain-aware clinical NLP systems.

7 Limitations

Our dataset currently includes 100 H&P narratives—50 from VBAC patients and 50 from RCS patients—randomly selected from a larger pool. While this subset provides an initial look at obstetrics-focused documentation, it may not capture the full variability of patients in this domain. For section annotation, we adopted a set of obstetrics-specific headers developed in collaboration with a certified midwifery expert. While these labels offer improved clinical relevance over general-purpose schemas such as MedSecId (Landes et al., 2022), they may introduce subjectivity, as other experts might define or group sections differently. This lack of standardization may limit comparability across datasets or models. Future work should explore building consensus-driven or ontology-aligned section schemas tailored to obstetrics, as well as expanding dataset coverage to better reflect diverse clinical structures and documentation styles.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*. Available only as a preprint on arXiv.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard III. 2008. Development and evaluation of a clinical note section header terminology. In *AMIA annual symposium proceedings*, volume 2008, page 156. American Medical Informatics Association.

Joshua C Denny, Anderson Spickard III, Kevin B Johnson, Neeraja B Peterson, Josh F Peterson, and Randolph A Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806–815.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

G David Forney. 1973. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Kavita Ganesan and Michael Subotin. 2014. A general supervised approach to segmentation of clinical texts. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 33–40. IEEE.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*. Available only as a preprint on arXiv.

Paul A. Harris, Robert Taylor, Brenda L. Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, Jacqueline Kirby, and Stephany N. Duda. 2019. The redcap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95:103208.

Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G Conde. 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381.

John H Holmes, James Beinlich, Mary R Boland, Kathryn H Bowles, Yong Chen, Tessa S Cook, George Demiris, Michael Draugelis, Laura Fluharty, Peter E Gabriel, et al. 2021. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine*, 60(01/02):032–048.

Simon Hughes, Minseok Bae, and Miaoran Li. 2023. Vectara hallucination leaderboard. Dataset available under the Apache-2.0 license.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*. Available only as a preprint on arXiv.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

S.H. Kim, S. Schramm, L.C. Adams, et al. 2025. Benchmarking the diagnostic performance of open source llms in 1933 eurorad case reports. *npj Digital Medicine*, 8:97.

Veysel Kocaman and David Talby. 2021. Spark nlp: Natural language understanding at scale. *Software Impacts*, 8:100058.

Hyoun-Joong Kong. 2019. Managing unstructured big data in healthcare system. *Healthcare informatics research*, 25(1):1–2.

Paul Landes, Kunal Patel, Sean S Huang, Adam Webb, Barbara Di Eugenio, and Cornelia Caragea. 2022. A new public corpus for clinical section identification: Medsecid. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3709–3721.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM international health informatics symposium*, pages 744–750.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*. Available only as a preprint on arXiv.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Majd Saleh, Sarra Baghdadi, and Stéphane Paquelet. 2024. Tocbert: Medical document structure extraction using bidirectional transformers. *arXiv preprint arXiv:2406.19526*. Available only as a preprint on arXiv.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. Available only as a preprint on arXiv.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. Available only as a preprint on arXiv.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*. Available only as a preprint on arXiv.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Fan Zhang, Itay Laish, Ayelet Benjamini, and Amir Feder. 2022. Section classification in clinical notes with multi-task transformers. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 54–59.

Weipeng Zhou and Timothy A Miller. 2024. Generalizable clinical note section identification with large language models. *JAMIA Open*, 7(3):ooae075.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association (JAMIA)*, 18(5):552–556.

A Appendix

A.1 Section Header Comparison Between MedSecId and Obstetrics

Section Header	MedSecId	Obstetrics
<none>	✓	✓
24-hour-events	✓	✗
addendum	✓	✗
allergies	✓	✓
assessment-and-plan	✓	✓
chief-complaint	✓	✓
clinical-implications	✓	✗
consent	✗	✓
code-status	✓	✗
communication	✓	✗
comparison	✓	✗
conclusions	✓	✗
contrast	✓	✗
critical-care-attending-addendum	✓	✓
current-medications	✓	✓
discharge-condition	✓	✗
discharge-diagnosis	✓	✗
discharge-disposition	✓	✗
discharge-instructions	✓	✗
discharge-medications	✓	✗
disposition	✓	✗
facility	✓	✗
family-history	✓	✓
findings	✓	✗
flowsheet-data-vitals	✓	✗
gestational-age	✗	✓
gynecological-history	✗	✓
history	✓	✗
history-of-present-illness	✓	✓
history-of-present-pregnancy	✗	✓
hospital-course	✓	✗
image-type	✓	✗
imaging	✓	✓
impression	✓	✗
impression-and-plan	✗	✓
indication	✓	✗
infusions	✓	✗
labs	✓	✓
labs-imaging	✓	✓
last-dose-of-antibiotics	✓	✗
major-surgical-or-invasive-procedure	✓	✗
medical-condition	✓	✗
medication-history	✓	✗
obstetrical-and-gynecological-history	✗	✓
obstetrical-history	✗	✓
other-medications	✓	✗
past-medical-history	✓	✓
past-surgical-history	✓	✓
patient-test-information	✓	✗
physical-examination	✓	✓
plan	✗	✓
pregnancy-history	✗	✓
prenatal-care	✗	✓
prenatal-history	✗	✓
prenatal-screens	✓	✓
problem-list	✗	✓
procedure	✓	✗
procedure-history	✗	✓
reason	✓	✗
review-of-systems	✓	✓
social-and-family-history	✓	✗
social-history	✓	✓
technique	✓	✗
wet-read	✓	✗

Table A1: Comparison of Section Headers in MedSecId vs. Obstetrics Dataset (✓ = Present, ✗ = Absent)

A.2 Section Header Distribution of Obstetrics Dataset

Section Header	Total Spans	Overall %
social-history	119	7.89
current-medications	114	7.56
allergies	114	7.56
physical-examination	102	6.76
family-history	97	6.43
history-of-present-illness	96	6.37
impression-and-plan	83	5.50
chief-complaint	79	5.24
review-of-systems	79	5.24
problem-list	79	5.24
pregnancy-history	79	5.24
gestational-age	78	5.17
procedure-history	64	4.24
past-medical-history	61	4.05
labs	51	3.38
past-surgical-history	49	3.25
obstetrical-history	46	3.05
gynecological-history	46	3.05
assessment-and-plan	19	1.26
critical-care-attending-addendum	12	0.80
labs-imaging	11	0.73
imaging	11	0.73
prenatal-history	11	0.73
obstetrical-and-gynecological-history	2	0.13
plan	2	0.13
prenatal-screens	1	0.07
consent	1	0.07
history-of-present-pregnancy	1	0.07
prenatal-care	1	0.07

Table A2: Frequency distribution of section headers in the Obstetrics dataset (excluding <none>).

A.3 Transformer-Based Section Segmentation—Training and Evaluation

Training Configuration We use the Trainer class from HuggingFace (Wolf et al., 2020) with the following hyperparameters (tuned within our GPU/memory constraints):

- **Learning rate:** 2e-5
- **Epochs:** 5
- **Batch size:** 32
- **Mixed precision:** Training is accelerated with bf16 precision
- **Max token length:** 100
- **Warmup steps:** 500
- **Weight decay:** 0.05

Evaluation Metrics We compute standard classification metrics—accuracy, precision, recall, F1—along with macro-F1 (class-agnostic) and weighted-F1 (weighing classes by frequency) to assess how class imbalance affects performance.

A.4 Transformer + CRF-Based Section Segmentation—Training and Evaluation

Training Details The training details for our experiments are as follows:

- **Learning rate:** 2e-5
- **Epoch:** 5
- **Batch size:** $B = 1$
- **Mixed precision:** Training is accelerated with bf16 precision
- **Optimizer:** AdamW (Loshchilov and Hutter, 2019) (updates Transformer + CRF parameters)
- **Max token length:** 100 for BERT-base, BioBERT, BiomedBERT; 64 for GatorTron (due to higher memory consumption)

Evaluation Metrics As in Section A.3, we compute precision, recall, macro-F1, and weighted-F1 to evaluate note-level segmentation performance.

A.5 Hallucination Analysis

Model	Top 5 Hallucinated Sections
Mistral-7B-Instruct-v0.3	*substance-abuse*, *neurologic*, *psychiatric*, *psychosocial-history*, *integumentary*
Llama 3.1-8B-Instruct	*review / management*, *review-and-management*, *health maintenance*, *psychosocial-history*, *obstetrical-examination*
Qwen-2.5-32B-Instruct	*basic-information*, *substance-abuse*, *psychosocial-history*, *obstetric-exam*, *postoperative-information*
Llama 3.3-70B-Instruct	*health-maintenance*, *risk-factors*, *psychosocial-history*, *comments*

Table A3: Top 5 most frequently hallucinated section headers generated by each model. (Llama 3.3-70B-Instruct produced only four hallucinated headers in total.)

```
Suppose the following are the valid section
headers:
{set of valid actual headers}

And the following are the hallucinated headers:
{set of hallucinated headers}

Please map each hallucinated header to the most
suitable or semantically similar valid header.
If no valid header is an appropriate match,
return "<none>".
```

Listing 2: Zero-shot prompt used to align hallucinated headers with valid section labels.

Model	CH	TH	S%
Mistral-7B-Instruct-v0.3	479	944	50.74%
Llama 3.1-8B-Instruct	23	115	20.00%
Qwen-2.5-32B-Instruct	116	177	65.54%
Llama 3.3-70B-Instruct	3	5	60.00%

Table A4: Correction success rates for hallucinated section headers. CH: number of corrected hallucinations that matched the gold-standard label; TH: total hallucinated lines; S%: success rate.

A.6 Zero-Shot Learning via LLMs—Inference Details

Inference Details To generate section labels, we perform a forward pass in inference-only mode with the following parameters:

- **temperature = 0.0**: Forces greedy decoding, prioritizing the most probable token at each step for consistent and deterministic output.
- **do_sample = False**: Disables random sampling, ensuring reproducible outputs for identical prompts.
- **num_beams = 1**: Avoids complex beam search, reducing computational overhead.
- **pad_token_id = tokenizer.eos_token_id**: Uses the end-of-sequence token for padding, preventing extraneous tokens in the output.

By combining these settings, our inference procedure remains deterministic and focused, yielding consistent line-by-line label predictions for each clinical note.

A.7 Qualitative Error Analysis

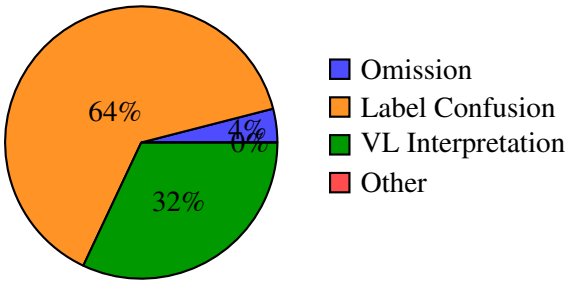

Figure 2: Proportional distribution of section labeling errors for Llama 3.3-70B-Instruct.

Text Span	Error Type	Explanation
Review / Management Results review: GBS (group B streptococcal): negative Hepatitis B : negative Syphilis screen: NR x2 Rubella: Immune HIV: NR STDs: Neg Blood type: O+ 1-hr GTT: 94 Genetic Screening Tests (First/Sequential/QUAD) : normal.	Label Confusion	The predicted label "assessment-and-plan" is a valid label, but it clearly differs from the gold label "labs-imaging", as the text span primarily discusses laboratory results, which aligns more closely with "labs-imaging".
In addition patient was incidentally found to have 4cm arrachnoid cyst of her left temporal fossa.	Valid Local Interpretation	The predicted label *physical-examination* makes sense given the text span, which describes a medical finding, even though the gold label is *assessment-and-plan*. The sentence could be part of a physical examination section, but in the context of the entire clinical note, it might be more appropriately classified under assessment and plan due to the incidental finding mentioned.

Table A5: Representative examples of Llama 3.3-70B-Instruct prediction mismatches categorized by LLM, with associated reasoning.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are analyzing a prediction error in a clinical note section classification task. A sentence was
assigned a gold-standard label. A language model attempted to predict this label.


<|eot_id|><|start_header_id|>user<|end_header_id|>
Your task is to assign section headers to each line of a clinical note. Most of the section headers
will likely span multiple lines, so headers should be assigned sequentially and consistently.

You are given:
    - Gold label: {gold_label}
    - Predicted label: {predicted_label}
    - Text span: {span_text}

Your task is to decide what type of error this is.

Choose only one of the following categories:
1. **Label Confusion** The model predicted a valid but clearly different label from the gold.
2. **Valid Local Interpretation** The predicted label is different from gold, but makes semantic
sense given the span alone.
3. **Other** This case is ambiguous or doesn't fit the above categories.

Respond exactly in the following format:
Label: <one of the 3 options above>
Reason: <your brief explanation>


<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Section Headers:
```

Listing 3: Zero-shot prompt used to classify qualitative errors.

A.8 Confidence Intervals for Model Performance on Obstetrics Data

Model	Macro F1 (±95% CI)	Weighted F1 (±95% CI)
Llama 3.3-70B-Instruct$_{\text{corrected}}$	0.800 ± 0.024	0.851 ± 0.024
Qwen-2.5-32B-Instruct$_{\text{corrected}}$	0.764 ± 0.032	0.818 ± 0.038
BioMedBERT+CRF	0.604 ± 0.026	0.646 ± 0.028

Table A6: Comparison of per-note macro and weighted F1 scores ± 95% bootstrap confidence intervals across 100 obstetric notes for the top two zero-shot LLMs and the best-performing supervised model (BioMedBERT+CRF).

A.9 Computational Resources

All training and inference experiments were conducted on *NVIDIA A100 GPUs (80GB VRAM)*. We used 3 GPUs for Llama 3.3-70B-Instruct, while all other models fit on a single GPU. Detailed training configurations, including batch sizes and epoch settings for both supervised and zero-shot experiments, are provided in their respective sections in Appendix A.