# Software Project: Learning to Diagnose

Stefan Riezler

WiSe 2024/25

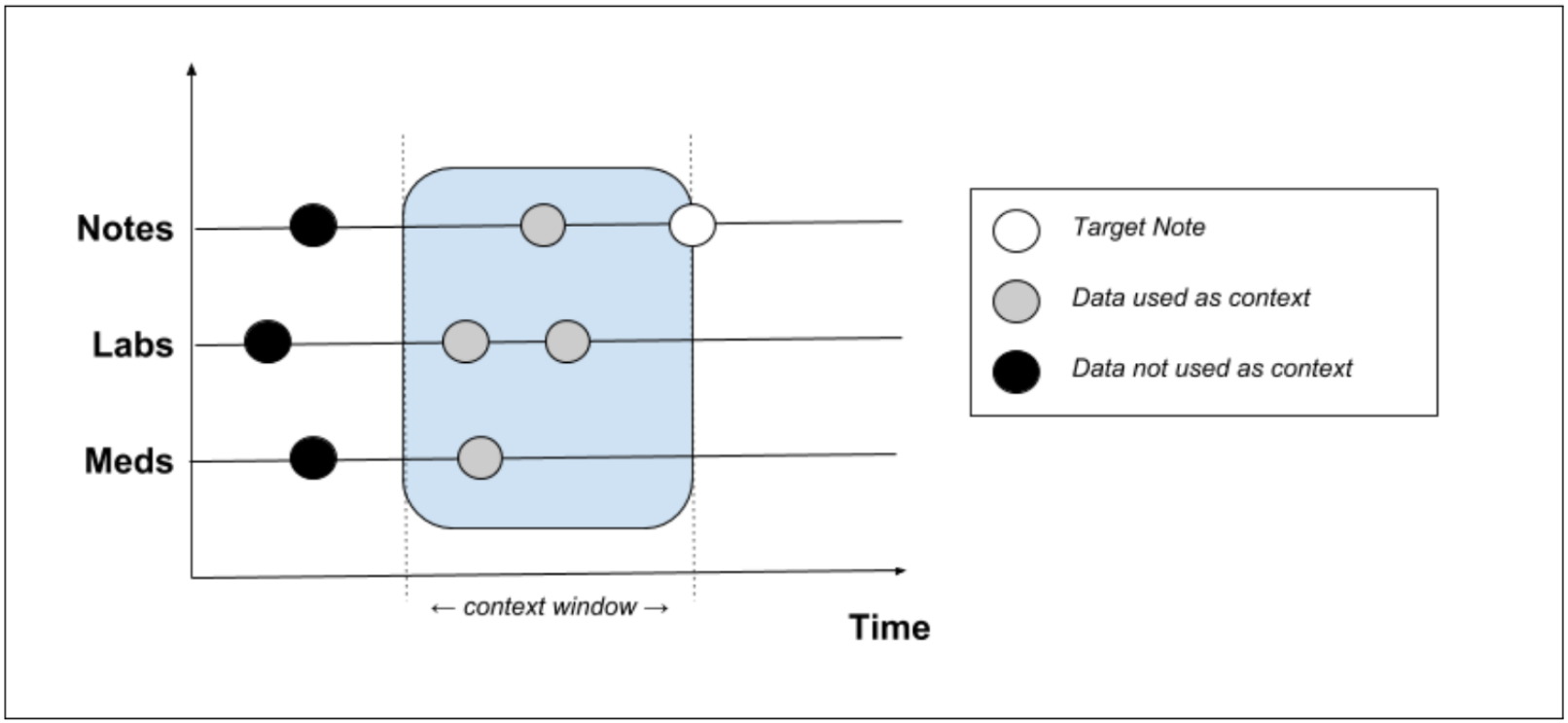
# Learning to Write Notes in Electronic Health Records

- Pioneering work from [Liu, 2018]
- Central idea: Encoder-decoder conditional language model $p_\theta$ that factorizes the prediction of a clinical note $w$ into conditional probabilities over tokens $w_1, w_2, \ldots, w_T$ where

$$p_\theta(w_1, w_2, \ldots, w_T|c) = \prod_{t=1}^{T} p_\theta(w_t|c, w_{<t}), \tag{1}$$

$w_{<t}$ is the context of past predicted notes, and the additional context data $c = (v_1, v_2, \ldots, v_m)$ consists of clinical measurements for clinical variables and of documented interventions.

# Learning to Write Notes in Electronic Health Records



# Learning to Write Notes in Electronic Health Records

- From-scratch training on MIMIC-III [Johnson et al., 2016] data
- Evaluation by comparison with clinical notes in dataset
  - Per-token perplexity
  - Next-token accuracy
  - ROUGE

# Project Proposal

- **Research question:** How do pre-trained LLMs perform on this task?
- **Project tasks:**
  - Get to know the data, especially the “nursing notes” `https://physionet.org/content/mimiciii/1.4/`
  - Use open-source pretrained LLM in few-shot learning mode, searching for robust prompts [Lu et al., 2022, Mizrahi et al., 2024]
  - Circumvent the “shaky foundations” [Wornow et al., 2023] of clincial AI by focusing on a proper evaluation, e.g., using medically-informed BERT-score and BARTscore [Ben Abacha et al., 2023]

# Project Proposal

- **Synthetic Data:**[1]
  - We added analyses of SOFA scores for 6 organ systems [Vincent et al., 1996], suspected infection, and Sepsis-3 assessment [Singer et al., 2016, Seymour et al., 2016] to the data

    *Organ dysfunction can be identified as an acute change in total SOFA score $\geq$ 2 points consequent to the infection. The baseline SOFA score can be assessed to be zero in patients not known to have preexisting organ dysfunction.*
  - Goal: Evaluate if LLM understands Sepsis-3 definition based on clinical measurements and SOFA-scores.

[1] `staniek@login:/workspace/mitarb/staniek/clinical_databases/physionet.org/files`

# Project Proposal

**Table 1. Sequential [Sepsis-Related] Organ Failure Assessment Score[a]**

| System | Score 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Respiration | | | | | |
| $Pao_2/Fio_2$, mm Hg (kPa) | ≥400 (53.3) | <400 (53.3) | <300 (40) | <200 (26.7) with respiratory support | <100 (13.3) with respiratory support |
| Coagulation | | | | | |
| Platelets, $\times 10^3/\mu L$ | ≥150 | <150 | <100 | <50 | <20 |
| Liver | | | | | |
| Bilirubin, mg/dL (μmol/L) | <1.2 (20) | 1.2-1.9 (20-32) | 2.0-5.9 (33-101) | 6.0-11.9 (102-204) | >12.0 (204) |
| Cardiovascular | MAP ≥70 mm Hg | MAP <70 mm Hg | Dopamine <5 or dobutamine (any dose)[b] | Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤0.1[b] | Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1[b] |
| Central nervous system | | | | | |
| Glasgow Coma Scale score[c] | 15 | 13-14 | 10-12 | 6-9 | <6 |
| Renal | | | | | |
| Creatinine, mg/dL (μmol/L) | <1.2 (110) | 1.2-1.9 (110-170) | 2.0-3.4 (171-299) | 3.5-4.9 (300-440) | >5.0 (440) |
| Urine output, mL/d | | | | <500 | <200 |

Abbreviations: $Fio_2$, fraction of inspired oxygen; MAP, mean arterial pressure; $Pao_2$, partial pressure of oxygen.

[a] Adapted from Vincent et al.[27]

[b] Catecholamine doses are given as μg/kg/min for at least 1 hour.

[c] Glasgow Coma Scale scores range from 3-15; higher score indicates better neurological function.

# References


Ben Abacha, A., Yim, W.-w., Michalopoulos, G., and Lin, T. (2023).
An investigation of evaluation methods in automatic medical note generation.
In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada.


Johnson, A. E., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016).
MIMIC-III, a freely accessible critical care database.
*Scientific Data*, 3(1):160035.


Liu, P. J. (2018).
Learning to write notes in electronic health records.
*arXiv*, abs/1808.02622.


Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022).
Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.
In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland.


Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. (2024).
State of What Art? A Call for Multi-Prompt LLM Evaluation.

*Transactions of the Association for Computational Linguistics (TACL)*, 12:933–949.

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., Deutschman, C. S., Escobar, G. J., and Angus, D. C. (2016).
Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (Sepsis-3).
*JAMA*, 315(8):762–774.

Singer, M., Deutschman, C. S., and Seymour, C. W. (2016).
The third international consensus definitions for sepsis and septic shock (Sepsis-3).
*JAMA*, 315(8):801–810.

Vincent, J., Moreno, R., Takala, J., Willatts, S., Mendonça, A. D., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996).
The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure.
*Intensive Care Medicine*, 22(7):707–710.

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., and Shah, N. H. (2023).
The shaky foundations of large language models and foundation models for electronic health records.
*npj Digital Medicine*, 6(135).