



Paper Link

# Delta-KNN: Improving Demonstration Selection in In-Context Learning for Alzheimer's Disease Detection



Chuyuan Li\*, Raymond Li\*, Thalia S. Field<sup>^</sup>, Giuseppe Carenini\*

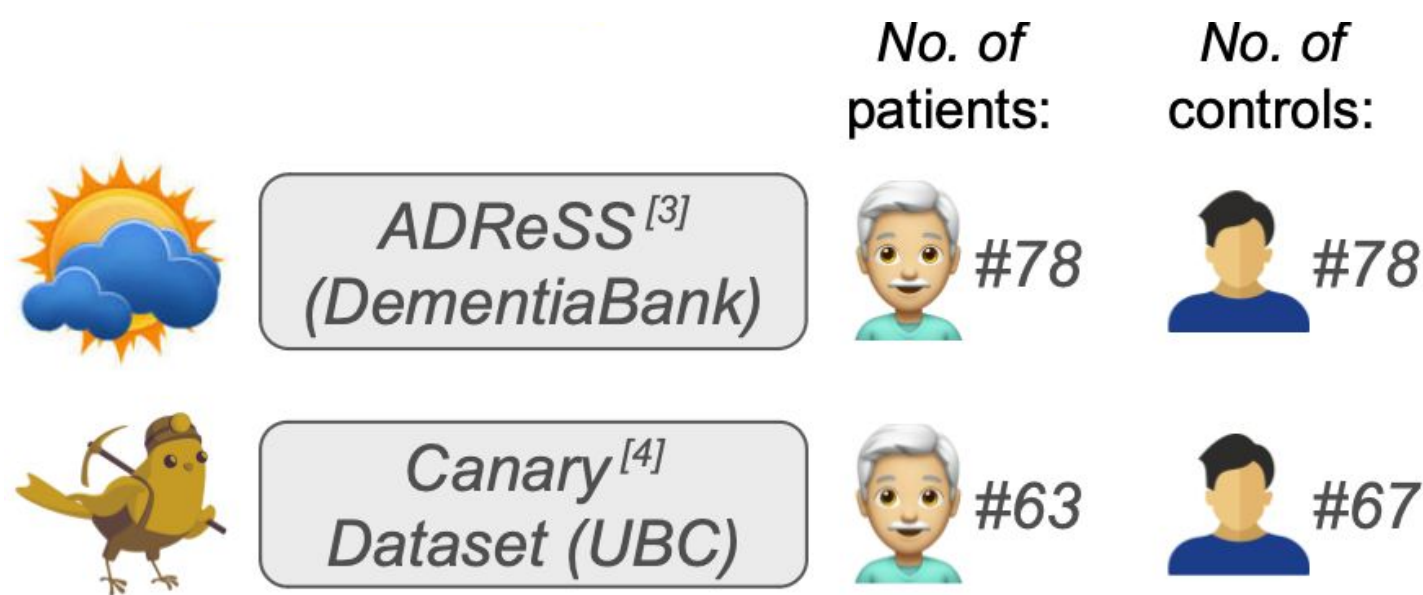
\*Department of Computer Science, <sup>^</sup>Faculty of Medicine, The University of British Columbia  
{chuyuan.li, thalia.field}@ubc.ca, {raymondli, carenini}@cs.ubc.ca

## Context & Motivation

- **Emergent Capabilities:** Large proprietary Language Models (LLMs) such as GPT-4 have shown impressive performance on professional benchmarks in the health domain.
- **Interpretable Explanations:** LLMs can generate interpretable explanations to their predictions, providing clinical doctors with valuable insights into their reasoning.
- **Very limited data in Healthcare Domain:** In-Context Learning (ICL)—a model performs a new task by conditioning on a few input-output pairs during inference time—emerged as a powerful and widely adopted strategy.
- **Existing Search-Based ICL Approaches:** Similarity-based (Liu et al., 2022) or understanding-based (Peng et al., 2024) performs poorly on AD detection from text → We introduce a novel demonstration selection method for ICL.

## Task & Datasets

- Task: **Cookie Theft** Picture Description.

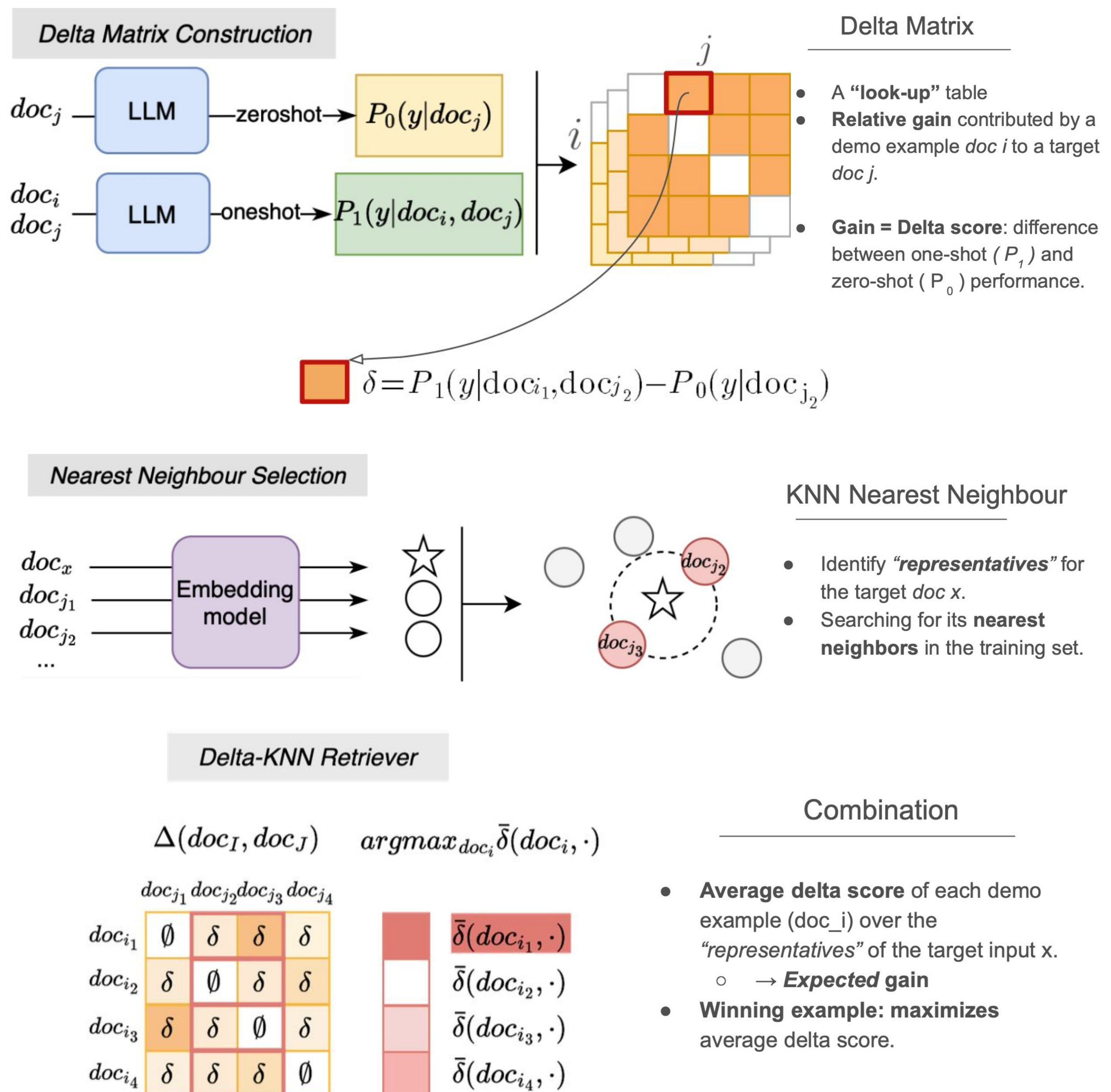


[4] Hyeju Jang et al. 2021. Classification of Alzheimer's Disease leveraging multi-task machine learning analysis of speech and eye-movement data. *Frontiers in Human Neuroscience*.

## Experiments

- ICL Baselines
  - Zero-shot
  - Random Sampling
  - Similarity-based Top-k Selection (Liu 2022)
  - Text-understanding-based Conditional Entropy Selection (Peng et al., 2024)
- Supervised baselines
  - Statistical ML Classifiers: SVM, RF, LR
  - Transfer learning-based LM: BERT, GPT-3
  - Supervised Fine-Tuning
- LLMs: Llama3.1-8B, Mistral-7B, Qwen2.5-7B
- Prompt Template: Full context+Guided CoT (Li et al., 2025)

## Method: Delta-KNN



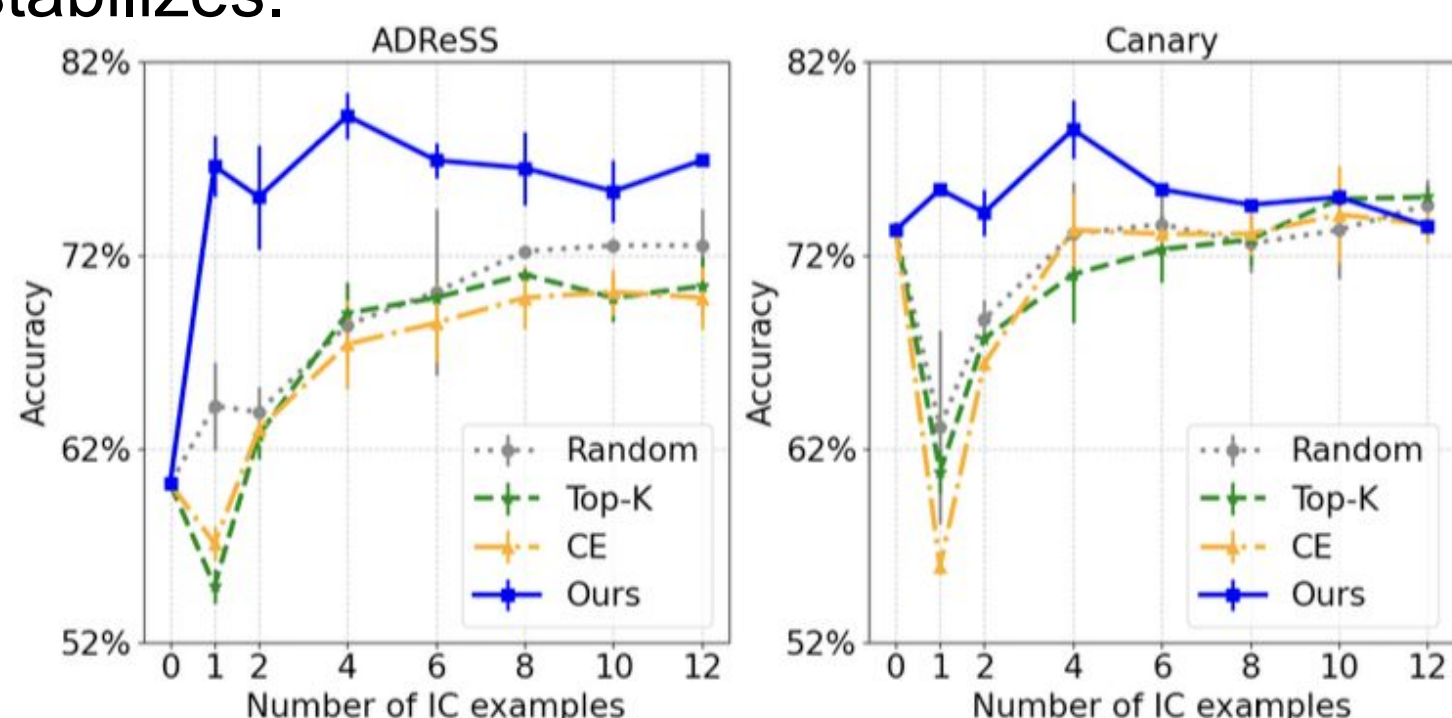
## Results and Analysis

- On **Llama**, **Mistral**, and **Qwen** LLMs, our proposed method **consistently outperforms** all selection methods on both datasets, achieving a 5-10% and 5% accuracy improvement on ADReSS and Canary, respectively.

Method	ADReSS-train				ADReSS-test				Canary			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
Zero-shot	62.2 <sub>0.0</sub>	60.1 <sub>0.0</sub>	98.1 <sub>0.0</sub>	22.2 <sub>0.0</sub>	57.6 <sub>1.0</sub>	57.6 <sub>1.0</sub>	100.0 <sub>0.0</sub>	15.3 <sub>2.0</sub>	73.3 <sub>0.4</sub>	72.1 <sub>1.0</sub>	79.4 <sub>0.0</sub>	67.7 <sub>0.7</sub>
Random	68.4 <sub>2.2</sub>	71.9 <sub>3.1</sub>	84.0 <sub>2.3</sub>	48.8 <sub>6.3</sub>	75.7 <sub>4.3</sub>	81.5 <sub>2.6</sub>	93.1 <sub>2.0</sub>	58.3 <sub>9.0</sub>	73.1 <sub>2.7</sub>	75.3 <sub>3.7</sub>	72.0 <sub>3.3</sub>	74.1 <sub>2.5</sub>
Top-k Select.	69.0 <sub>1.6</sub>	71.9 <sub>2.5</sub>	88.3 <sub>2.3</sub>	45.7 <sub>1.7</sub>	70.1 <sub>2.0</sub>	80.0 <sub>0.8</sub>	91.7 <sub>3.4</sub>	48.6 <sub>2.0</sub>	71.0 <sub>2.5</sub>	75.0 <sub>2.2</sub>	76.7 <sub>0.7</sub>	65.7 <sub>4.2</sub>
ConE* Select.	67.4 <sub>2.3</sub>	74.5 <sub>1.3</sub>	85.2 <sub>1.5</sub>	45.7 <sub>3.1</sub>	70.1 <sub>1.0</sub>	76.4 <sub>2.6</sub>	93.1 <sub>2.0</sub>	47.2 <sub>2.0</sub>	73.3 <sub>1.9</sub>	78.4 <sub>0.9</sub>	79.9 <sub>2.0</sub>	67.2 <sub>4.4</sub>
Delta-KNN (ours)	79.2 <sub>1.2</sub>	78.9 <sub>1.3</sub>	69.1 <sub>0.9</sub>	85.2 <sub>1.5</sub>	80.5 <sub>3.9</sub>	85.8 <sub>0.9</sub>	70.8 <sub>5.9</sub>	86.1 <sub>2.0</sub>	78.5 <sub>1.5</sub>	79.8 <sub>0.9</sub>	70.6 <sub>0.8</sub>	85.8 <sub>2.2</sub>

	ADReSS-train	ADReSS-test	Canary
<b>Mistral-7B-Instruct-v0.3</b>			
Zero-shot	52.3 <sub>0.5</sub>	67.7 <sub>1.0</sub>	63.1 <sub>0.8</sub>
Random	62.0 <sub>2.8</sub>	70.8 <sub>2.1</sub>	55.0 <sub>0.4</sub>
Top-k Select.	53.2 <sub>2.3</sub>	63.5 <sub>3.1</sub>	62.3 <sub>0.0</sub>
ConE Select.	61.1 <sub>1.9</sub>	66.7 <sub>4.2</sub>	58.8 <sub>3.5</sub>
Ours	69.9 <sub>1.4</sub>	76.0 <sub>5.2</sub>	72.3 <sub>0.4</sub>
<b>Qwen2.5-7B-Instruct</b>			
Zero-shot	61.6 <sub>0.5</sub>	66.8 <sub>2.2</sub>	63.5 <sub>0.4</sub>
Random	62.0 <sub>2.8</sub>	57.3 <sub>1.0</sub>	64.6 <sub>3.8</sub>
Top-k Select.	58.8 <sub>1.4</sub>	66.7 <sub>2.1</sub>	53.1 <sub>6.2</sub>
ConE Select.	58.8 <sub>0.5</sub>	65.8 <sub>5.3</sub>	60.0 <sub>1.5</sub>
Ours	63.4 <sub>0.5</sub>	67.7 <sub>0.0</sub>	66.1 <sub>2.7</sub>

- Impact of **in-context examples N**: Delta-KNN (ours) shows immediate advantage at N=1, peaking at N=4, after it stabilizes.



- Impact of **Demonstration Ordering**: Ours achieves higher maximum and average accuracy across 24 possible orderings in the 4-shot setting, with lower standard deviation.
- Impact of **Prompt Engineering**: Seven prompt variations, ours consistently outperforms ICL baselines.
- Impact of **k value in Delta-KNN**: Varying k from 1 to 20 on train sets, found k=13 yields the best results on both datasets.
- Further Investigation in the paper such as:
  - Text encoders: LLM hidden states vs. OpenAI embeddings
  - Comparison with supervised baselines