

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

Chuyuan Li*, Raymond Li*, Thalia S. Field^, Giuseppe Carenini*

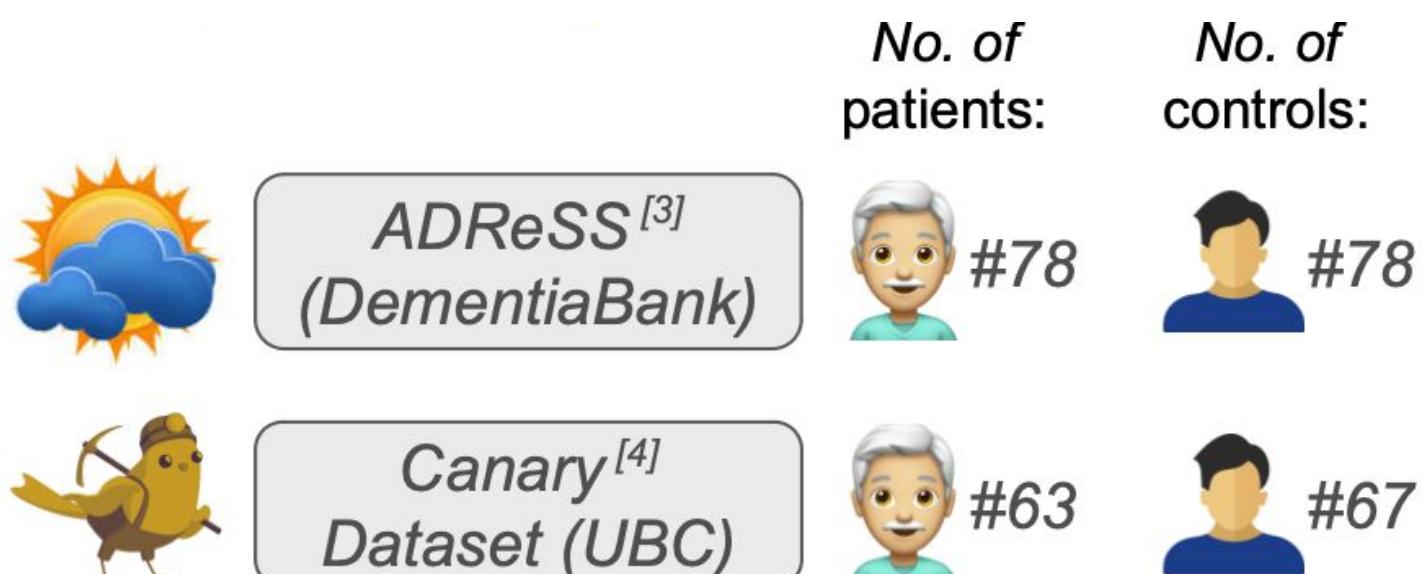
*Department of Computer Science, ^Faculty of Medicine, The University of British Columbia
{chuyuan.li, thalia.field}@ubc.ca, {raymondli, carenini}@cs.ubc.ca

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.

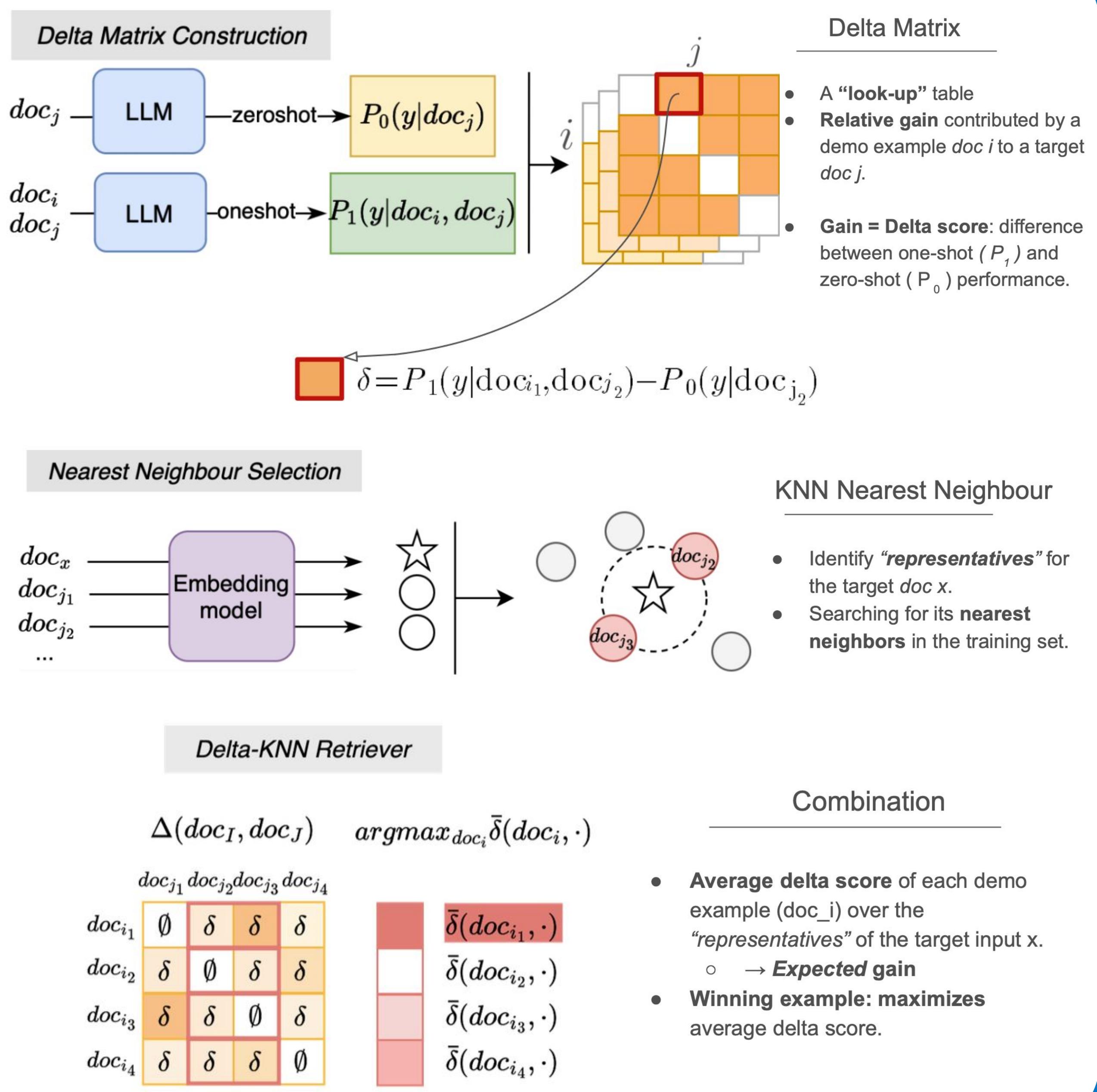
Task & Datasets

- 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*.

Method: Delta-KNN



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)

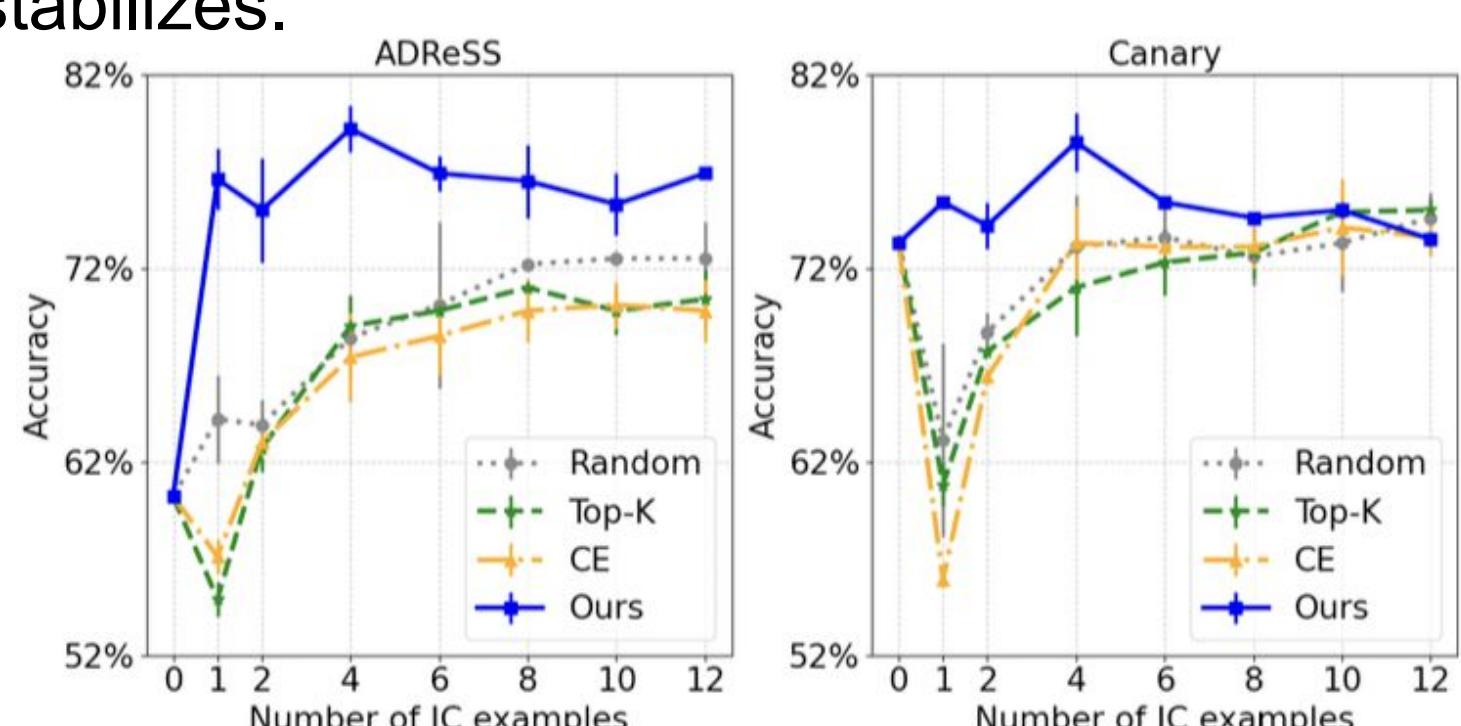
Results and Analysis

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

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

- 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