LLM4II

WP4

Boosting LLMs with other data

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Loria Laboratoire lorrain de recherche en informatique et ses applications

Inria

LINAGORA

October 24th, 2024

Plan

1. ASR Domain Adaptation by **Jérôme Louradour**
2. Clinical Notes by **Ivan Lerner**
3. QUARTZ by **Mohamed Imed Eddine Ghebriout**
4. Medical Domain Adaptation by **Gaël Guibon**

ASR Domain Adaptation

by Jérôme Louradour
(work jointly done with Edgar Laugeais)

Preliminary results

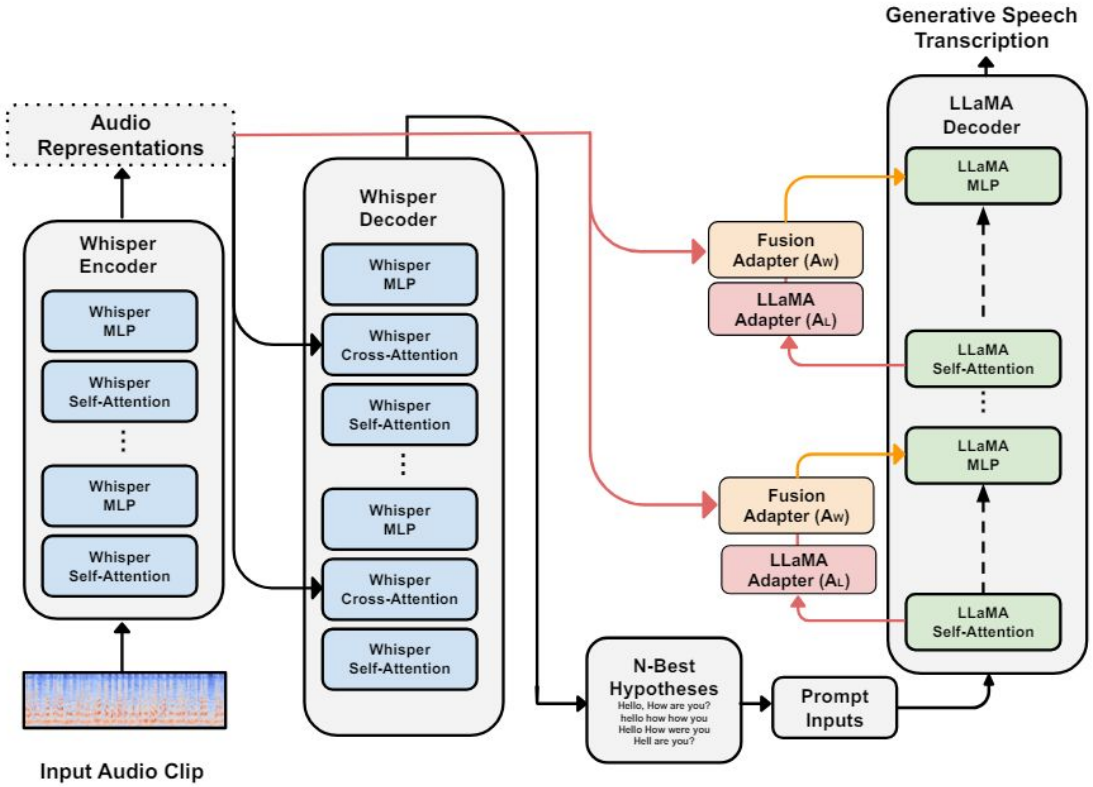
- **SimSamu dataset**
  - ~ 10% speech zone with very low volume (can be removed by VAD)
  - ~ 10% speech disfluencies
  - ~ 15% WER remaining (some spelling errors)

pre-Whisper VAD	disfluent prompt	WER	Deletion	Insertion	Substitution
_	_	30.3%	18.3%	2.8%	9.2%
Silero	_	35.1%	28.0%	**1.7%**	**5.4%**
Auditok	_	26.8%	18.6%	2.2%	6.0%
_	***"Bon. Ben…"***	**21.5%**	**10.8%**	2.8%	7.9%
Silero	***"Bon. Ben …"***	28.1%	19.2%	2.3%	6.6%
Auditok	***"Bon. Ben …"***	**21.9%**	**11.6%**	2.6%	7.7%

ASR Domain Adaptation

- Motivations
  - Whisper is state-of-the-art in French and English
  - Classical ASR systems used to separate Acoustic Model / Language Model
  - Whisper is an end-to-end encoder-decoder transformer : it is impossible to isolate a Language Model that can be tuned with text only
- Existing approaches to adapt Whisper to a new lexical domain
  - Whispering Llama / TCPGen with GPT-2
  - need transcribed audio with that vocabulary
- We propose an approach which only needs text
  - A lexicon with “new” words
  - A Language Model (n-gram or transformers) trained on the medical domain

ASR Domain Adaptation (Whispering Llama)



Model Input

Instruction:

You are an ASR transcript selector. You have a few transcripts generated by an automatic speech recognition model. Your task is to generate the most likely transcript from them. If the generated transcripts have grammatical or logical errors, you will modify them accordingly to produce the most accurate and coherent transcript.

Input:
so that it can carry its momentum to the logic
that he can carry his moment on tour with the logic
that he can carry his momentum through with the logic
that he can carry his momentum to with an logic
that he can carry his momentum true within logic
that he carries momentum through with logic
that it can carry a moment and through with the logic
that it can carry a moment of truth with the logic
that it can carry a moment on tour with the logic
that it can carry a moment on true with the logic
that it can carry a momentum tool within logic
that it can carry as a moment on tour with the logic
that it can carry at moments and through with the logic
that it can carry his moment on tour with a logic
that it can carry his moment on tour with the logic

Response:

Model Output

that it can carry its momentum through with a logic.

Whisper Top-N hypotheses (N~100)

public fork of openai-whisper: github.com/linagora-labs/whisper_nbest

- access to tokens visited in the beam-search, along with probabilities
- heuristics to avoid uninteresting diversity (punctuation marks variants, start/end timestamps…)

Speech segment alternatives	Avg. token prob.
Dactarin 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.771
Dactarin 2% sous forme de gel buccal, à appliquer **deux** fois par jour pendant **huit** jours.	0.753
Dactaryn 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.701
Dactarins 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.700
Dac Tarin 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.691
Dacta Rhin 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.678
DAC-TARIN 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.674
Daktarin 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.670
Dactar 1 2% sous forme de gel buccal à appliquer **deux** fois par jour pendant **huit** jours.	0.670
Dactarun 2% sous forme de gel buccal, à appliquer **2** fois par jour pendant **8** jours.	0.670
…	…
Daktarin 2% sous forme de gel buccal, à appliquer **deux** fois par jour pendant **huit** jours.	0.644
…	…

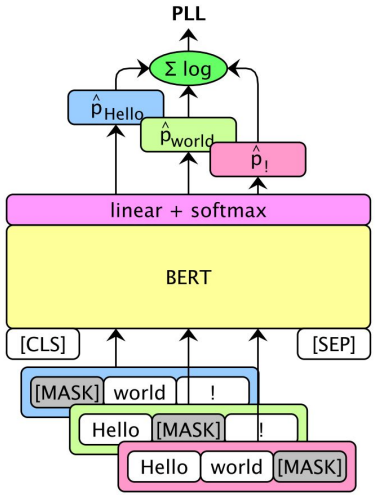
Approach 1: Rerank Top-N hypotheses using a LM

- Top-N re-ranking using Language Model prior probabilities

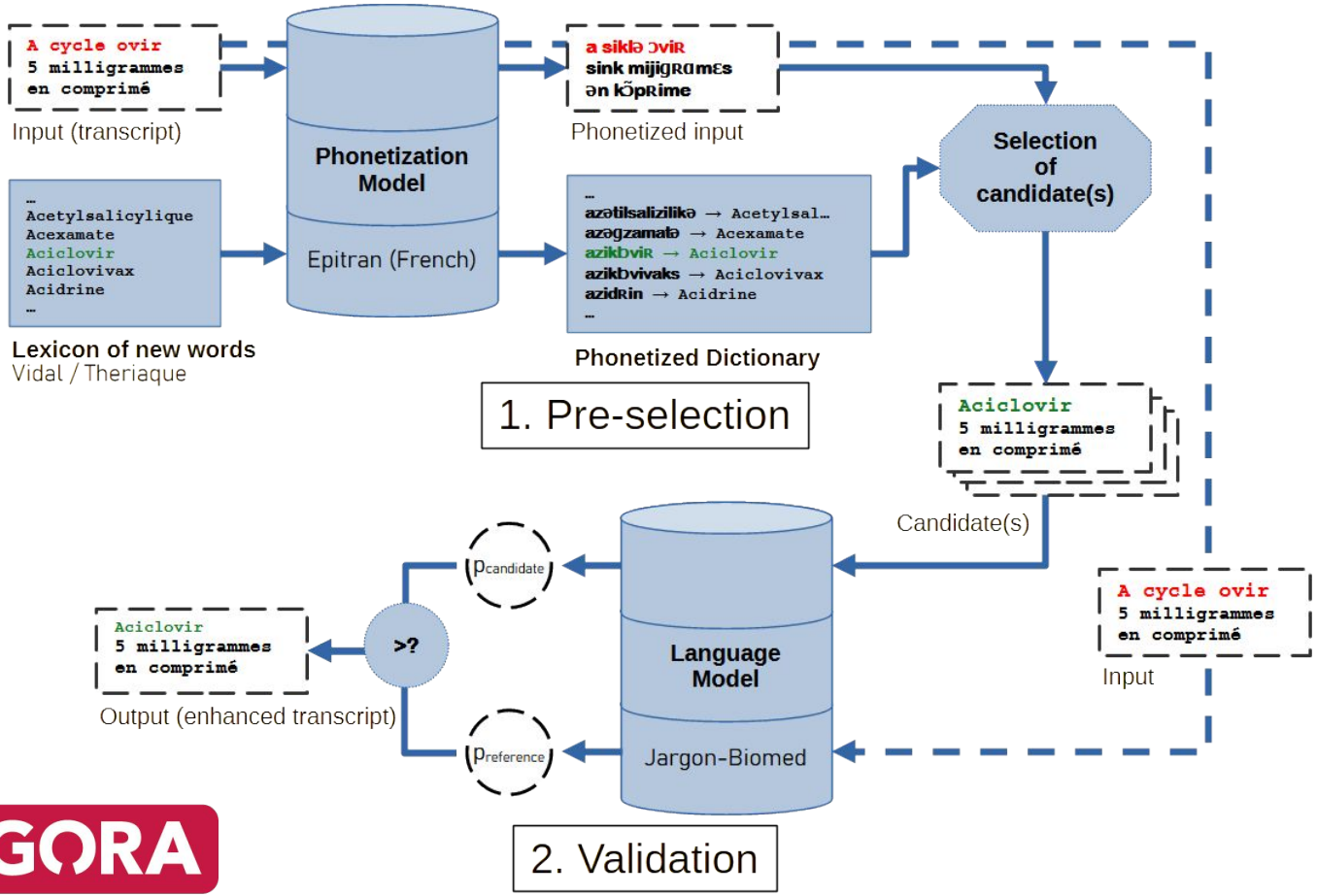
  $$\textit{score to be maximized} = (1-\alpha) \cdot \log P_{Whisper}(\text{text}|\text{audio}) + \alpha \cdot \log P_{LanguageModel}(\text{text})$$

- BERT-like transformers LM specialized in medical data:
  - Jargon-Biomed (best)
  - Dr-BERT 7B
  - Camembert-Bio

- Pseudo log-probabilities



Approach 2: Phonetic Selection and LM Validation (PS-LMV)



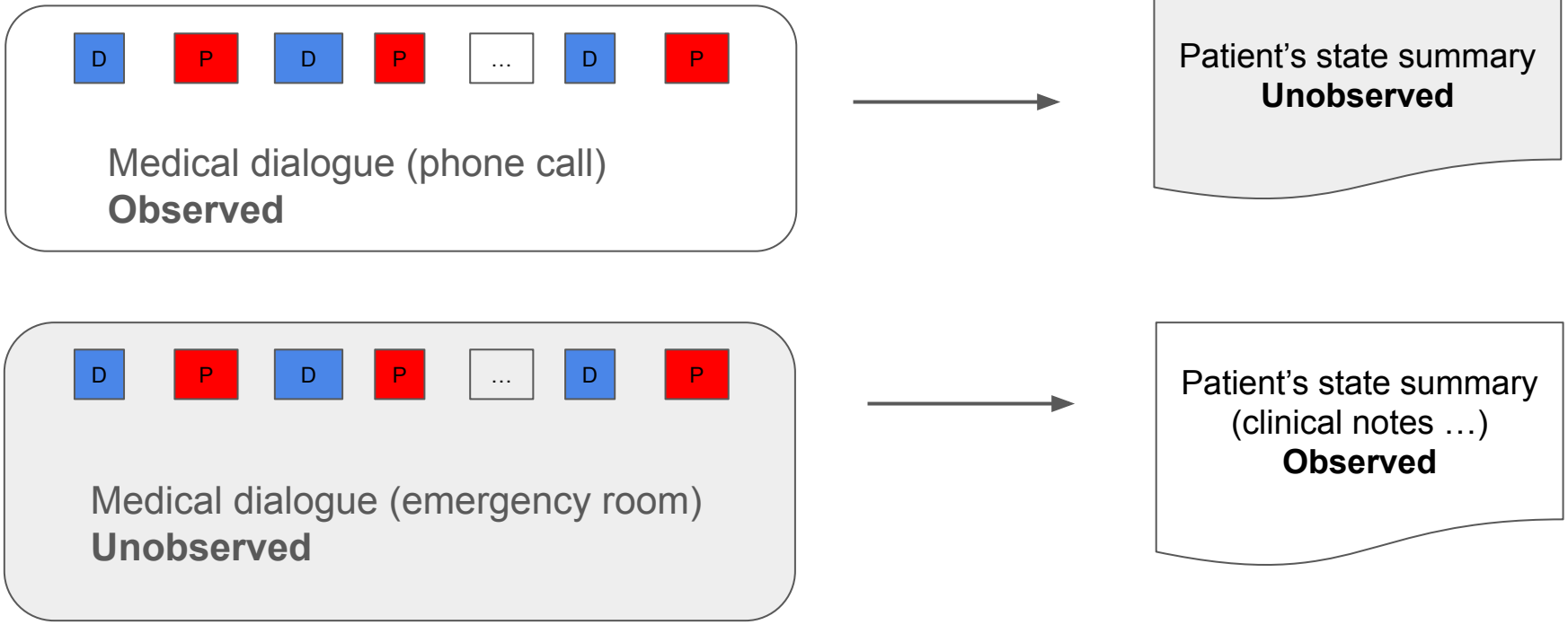
Results on PxSLU (drug prescriptions and misc.)

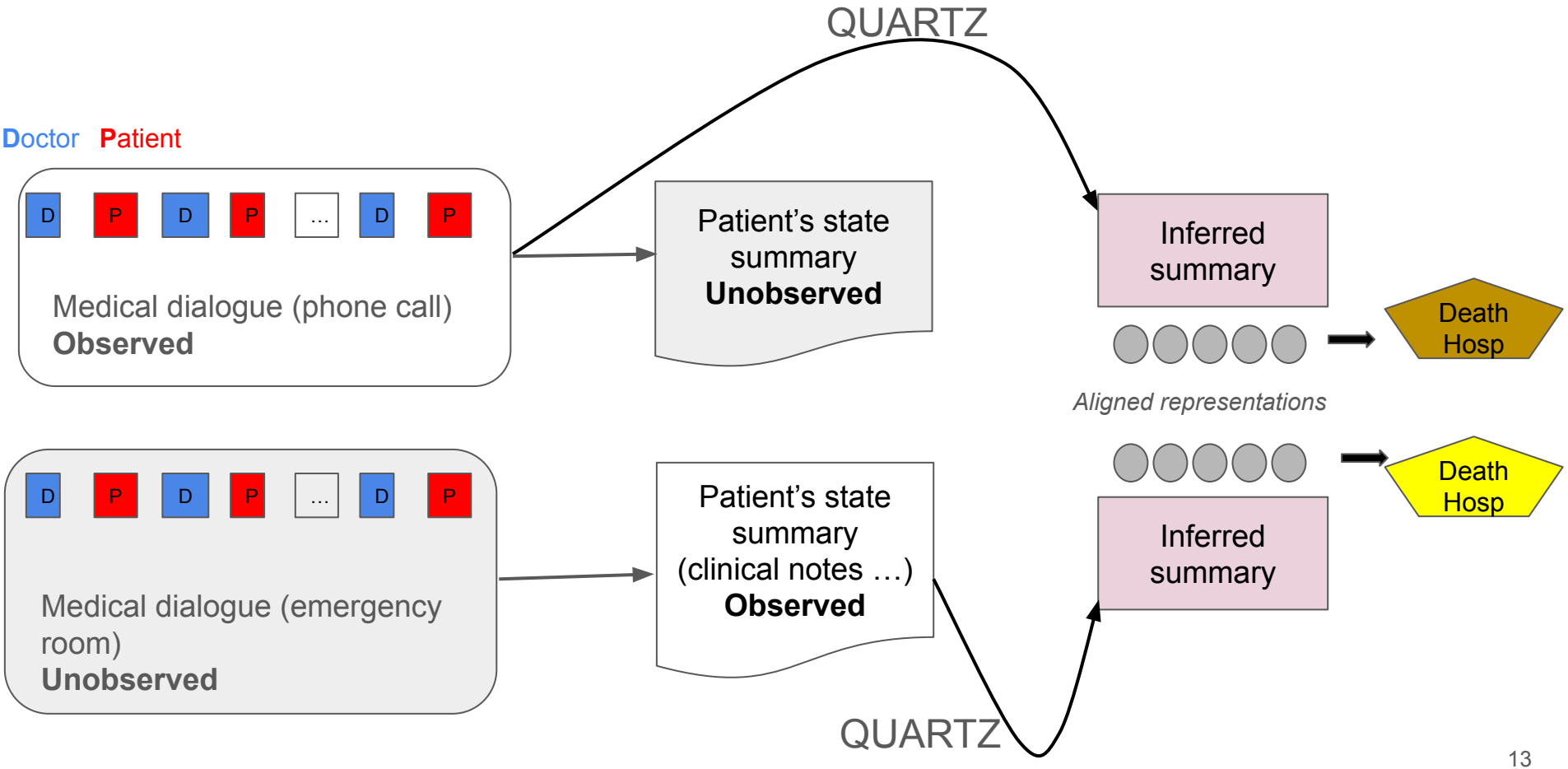
	WER↘	drug names detection		
		recall↗	F1↗	precision↗
Whisper Baseline	15.57%	31.21%	45.89%	86.63%
Approach 1: Whisper top-100 + Jargon-Biomed (α=0.1)	15.84%	45.56%	60.13%	**88.39%**
Approach 2: Whisper + PS (Theriaque)	**14.33%**	53.37%	61.69%	73.09%
Approach 2: Whisper + PS-LMV (Theriaque, Jargon-Biomed)	14.66%	46.44%	57.45%	75.31%
Hybrid approach: Whisper top-100 + PS-LMV (α=0.1)	15.21%	**59.46%**	**66.31%**	74.94%

Clinical Notes Dataset

The data generating process

Doctor Patient




QUARTZ
Doctor
Patient
D
P
D
P
...
D
P
Medical dialogue (phone call)
Observed
Patient's state summary
Unobserved
Inferred summary
Death Hosp
Aligned representations
D
P
D
P
...
D
P
Medical dialogue (emergency room)
Unobserved
Patient's state summary (clinical notes …)
Observed
Inferred summary
Death Hosp
QUARTZ

	Alive D30 (N=388067)	Death D30 (N=2192)	Total (N=390259)
Sex			
F	185364 (47.8%)	986 (45.0%)	186350 (47.8%)
M	202703 (52.2%)	1206 (55.0%)	203909 (52.2%)
Age			
Median [IQR]	38.2 (26.2, 57.1)	80.1 (67.9, 88.0)	38.3 (26.2, 57.4)
death_d1			
0	388067 (100.0%)	1798 (82.0%)	389865 (99.9%)
1	0 (0.0%)	394 (18.0%)	394 (0.1%)
death_d3			
0	388067 (100.0%)	1477 (67.4%)	389544 (99.8%)
1	0 (0.0%)	715 (32.6%)	715 (0.2%)
death_d7			
0	388067 (100.0%)	1091 (49.8%)	389158 (99.7%)
1	0 (0.0%)	1101 (50.2%)	1101 (0.3%)
Year			
<2015	141573 (36.5%)	862 (39.3%)	142435 (36.5%)
<2020	146457 (37.7%)	690 (31.5%)	147147 (37.7%)
<2025	100037 (25.8%)	640 (29.2%)	100677 (25.8%)

QUARTZ by Imed Ghebriout

QUARTZ :
QA-based Unsupervised Abstractive Refinement for Task-oriented Dialogue Summarization

GHEBRIOUT Mohamed Imed Eddine
1st year PhD student at Loria (Multispeech / Synalp)
imed-eddine.ghebriout@loria.fr

Fine-tuning LLMs have demonstrated super-human performance for dialogue summarization (Van Veen et al., 2024).

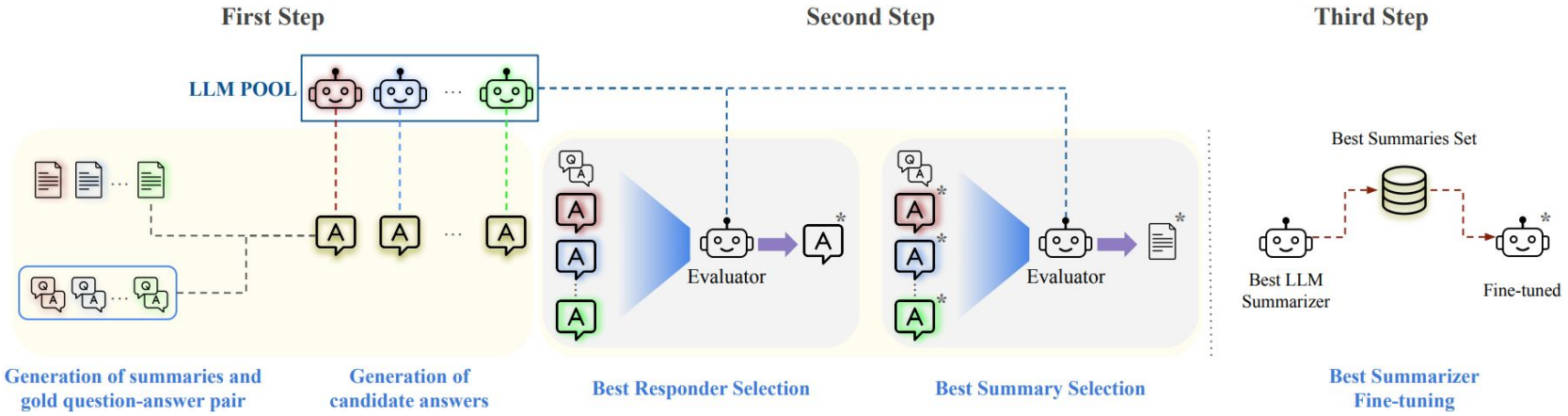
High cost for human annotation (data collection, labeling, etc)

LLMs still lack robust mechanisms for maintaining topic focus and factual consistency (Tonmoy et al., 2024).

Abstractive Dialogue Summarization Task-oriented Unsupervised

[1] Dave Van Veen, Cara Van Uden, Louis Blanke701 meier, Jean-Benoit Delbrouck, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine.
[2] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. preprint

QUARTZ : QA-based Unsupervised Abstractive Refinement for Task-oriented Dialogue Summarization.



A framework aimed at improving LLMs' capability for dialogue summarization in an unsupervised manner.

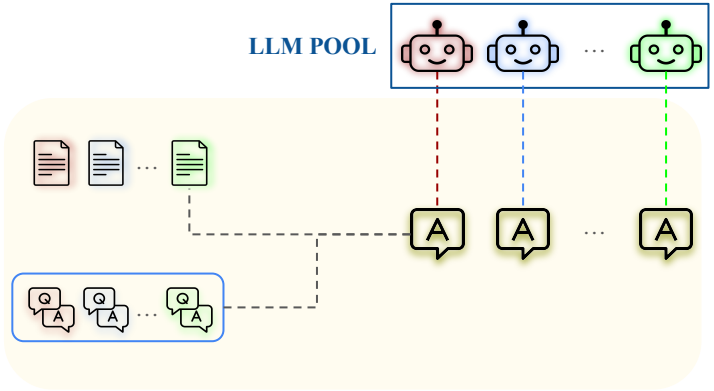
Two-stage approach for evaluating the quality of generated summaries.

We demonstrate the effectiveness of *QUARTZ* through fine-tuning on the selected summaries.

First Step : ***Generation***

Input Dialogue

Doctor: I'm Dr. DAMANI from SAMU 93, you're calling about your wife who hurt herself, right?
Patient: Yeah yeah and well she's I don't know she fell there on a she fell into the kitchen.
Doctor: So, she has a wound at what level?
Patient: Uh, so I think it's a wound on the arm.
.
.
.
Doctor: Good-bye, sir.
Patient: Good-bye.



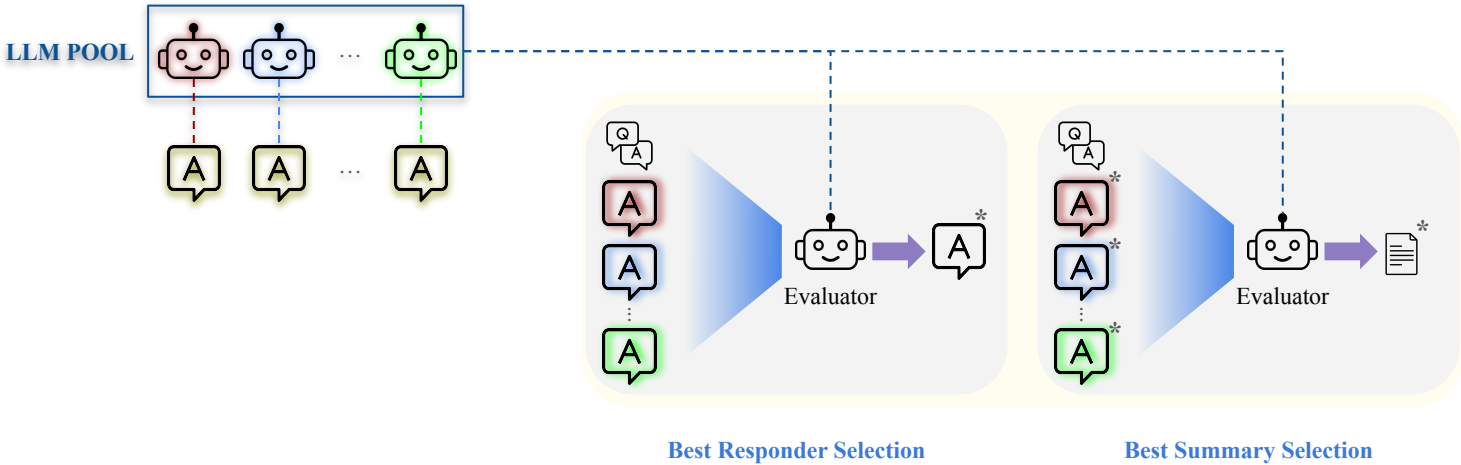
Generation of summaries and gold question-answer pair

Generation of candidate answers

LLMs Self-bias

(Xu et al., 2024) : An LLM may easily respond to its own generated question but may struggle to do so when responding to questions generated by other LLMs.

Second Step : ***Evaluation***



We use LLMs from LLM_POOL as rankers :

- **Best Responder Selection** (given the L_S , which L_R have answered correctly ?)
- **Best Summary Selection** (given a dialogue, which L_S is the best ?)

Second Step : *Evaluation*

#1 Prompting the LLM to rank different responses :

You will be given a question, a ground truth answer and a list of possible answers.
Your goal is to rank the generated answers indexes based on their correctness and closeness to the ground truth answer.
Question: <QUESTION>
Ground Truth Answer: <GT_ANSWER>
Possible answers:
1) <ANSWER_1>
2) <ANSWER_2>
3) <ANSWER_3>

Positional Bias

Self-preference Bias

The ranking is : [2, 1, 3]

[1] Panickssery, A., Bowman, S. R., & Feng, S. (2024). Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.
[2] Tang, R., Zhang, C., Ma, X., Lin, J., & Türe, F. (2024). Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics.

Second Step : ***Evaluation***

#2 Aggregating the ranking across the LLMs :

A ranking σ is a |L|-permutation. $\sigma : \{1, \ldots, |\mathcal{L}|\} \mapsto \{1, \ldots, |\mathcal{L}|\}$

The optimal ranking σ is obtained as follows :

$$\sigma_{i,j,l_\mathrm{S},l_\mathrm{E}} := \arg\min_{\sigma} \sum_{1 \le n \le N} d_\kappa(\hat{\sigma}_{i,j,l_\mathrm{S},l_\mathrm{E},n}, \sigma)$$

$$d_\kappa(\sigma_1, \sigma_2) := \sum_{k=1}^{|\mathcal{L}|} \mathrm{inv}(\sigma_1^{-1} \circ \sigma_2)_k$$

$$\mathrm{inv}(\sigma)_k := \#\{k' : \sigma(k') > \sigma(k), k' < k\}$$

Example (N=4) :

$\hat{\sigma}_1$ = [‘Answer_1’, ‘Answer_2’, ‘Answer_3’]

$\hat{\sigma}_2$ = [‘Answer_2’, ‘Answer_1’, ‘Answer_3’]

$\hat{\sigma}_3$ = [‘Answer_3’, ‘Answer_2’, ‘Answer_1’]

$\hat{\sigma}_4$ = [‘Answer_3’, ‘Answer_1’, ‘Answer_2’]

σ = [‘Answer_1’, ‘Answer_2’, ‘Answer_3’]

[1] Panickssery, A., Bowman, S. R., & Feng, S. (2024). Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.
[2] Tang, R., Zhang, C., Ma, X., Lin, J., & Türe, F. (2024). Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics.

Second Step : ***Evaluation***

#2 Selecting the best answer :

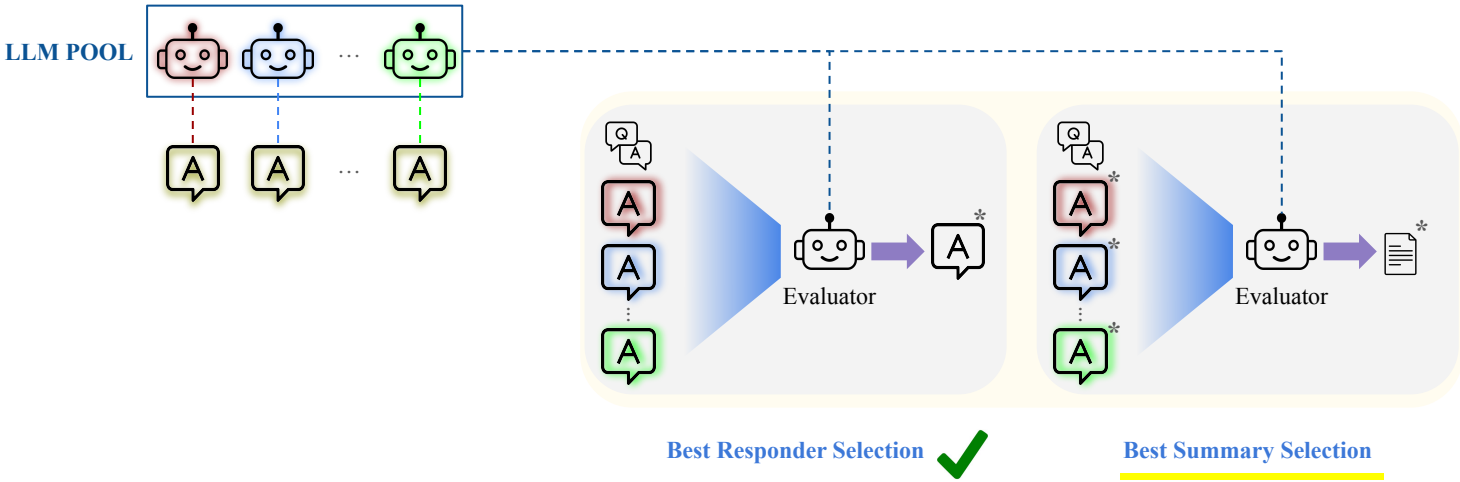
We score score each LLM based on the mean reciprocal rank (MRR)

$$\text{Score}_{i,l_\text{S},l_\text{R}} = \sum_{l_\text{E} \in \mathcal{L}} \alpha_{l_\text{R},l_\text{E}} \, \text{MRR}_{i,l_\text{S},l_\text{R},l_\text{E}}$$

$$\text{MRR}_{i,l_\text{S},l_\text{R},l_\text{E}} = \frac{1}{J_i} \sum_{j=1}^{J_i} \frac{1}{\sigma^{-1}_{i,j,l_\text{S},l_\text{E}}(l_\text{R})} \qquad \alpha_{l_\text{R},l_\text{E}} = \begin{cases} 0.8 & \text{if } l_\text{R} = l_\text{E} \\ 1 & \text{otherwise} \end{cases}$$

[1] Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02).

Second Step : ***Evaluation***



We use LLMs from LLM_POOL as rankers :

- **Best Responder Selection** (given the L_S , which L_R have answered correctly ?)
- **Best Summary Selection** (given a dialogue, which L_S is the best ?)

Third Step : ***Fine-tuning***

After selecting the best-summaries, we fine-tune the best LLM_Summarizer (the one that produced the majority of the selected summaries) using LoRA and we evaluate on the test set.

Experimental Setup :

LLM Pool = { Llama-3.1-8B-Instruct, Gemma-2-9b-it, Qwen2-7B-Instruct }

LoRA fine-tuning : $\Gamma_{LoRA} = 8$, $\propto_{LoRA} = 16$

N = 5 (the number of trials per ranking)

[1] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Evaluating QUARTZ : Datasets :

Dialogue
Blair: Remember we are seeing the wedding planner after work Chuck: Sure, where are we meeting her? Blair: At Nonna Rita's Chuck: Can I order their seafood tagliatelle or are we just having coffee with her? I've been dreaming about it since we went there last month Blair: Haha sure why not Chuck: Well we both remmber the spaghetti pomodoro disaster from our last meeting with Diane Blair: Omg hahaha it was all over her white blouse Chuck: :D Blair: :P
Summary
Blair and Chuck are going to meet the wedding planner after work at Nonna Rita's. The tagliatelle served at Nonna Rita's are very good.

SAMSum dialogue example

Dialogue:
Doctor: My chart here says that you're eighty three years old, is that correct, ma'am?
Patient: Yes doctor, that's correct, I just had my birthday.
Doctor: Happy belated birthday! How have you been doing since your last visit?
Patient: Well, my cancer hasn't needed phlebotomies for several months now, which is good.
Doctor: That's great, you have been treated for polycythemia vera, correct?
Patient: Yes, that's the one.
Doctor: I also see you're unassisted today, which is also great.
Patient: Yeah, having some independence is nice.
Section header: History of Present Illness
Section text: The patient is an 83-year-old female with a history of polycythemia vera. She comes in to clinic today for followup. She has not required phlebotomies for several months. The patient comes to clinic unaccompanied.

MTS-Dialog dialogue example

Dataset		# Dial.	Avg. Len.	Avg. Turns
SAMSum	Train	14,732	93.8	11.2
	Valid	818	91.6	10.8
	Test	819	95.5	11.3
MTS-Dialog	Train	1201	87.99	8.9
	Valid	100	77.46	7.7
	Test	200	87.69	9.1

Dataset statistics

[1] Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. EMNLP-IJCNLP 2019.
[2] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. EACL 2024.

Evaluating QUARTZ :

Results :

Method	R-1	R-2	R-L	BLEU	BERT-Score
Supervised Dialogue Summarization					
MoE (Tian et al., 2024)	**55.93**	30.86	**52.02**	**26.03**	**75.66**
InstructDS (Wang et al., 2023)	55.30	**31.30**	46.70	-	55.50
MoCa (Zhang et al., 2022)	55.13	30.57	50.88	-	-
Unsupervised Dialogue Summarization					
Llama 3.1	40.76	18.97	31.76	10.68	64.97
QUARTZ (Best summarizer: Llama 3.1)	**61.34**	**38.53**	**53.11**	**31.07**	**77.95**

Performance evaluation on the SAMSum dataset.

Method	R-1	R-2	R-L	BERT-Score	BLEURT
Supervised Dialogue Summarization					
BART-GS-DA (Abacha et al., 2023)	42.52	17.50	34.90	40.80	51.23
GPT4-ICL (Giorgi et al., 2023)	**44.66**	**22.82**	**38.37**	**73.07**	**55.93**
Unsupervised Dialogue Summarization					
Llama 3.1	28.17	10.23	21.11	55.82	39.16
QUARTZ (Best summarizer: Llama 3.1)	**34.69**	**14.00**	**27.75**	**61.18**	**48.11**

Performance evaluation on the MTS-Dialog dataset

[1] Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. ACL 24".
[2] Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. EMNLP 23".
[3] Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei.2022. Momentum calibration for text generation.
[4] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. EACL 24".
[5] John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In Proceedings of the 5th Clinical Natural Language Processing Workshop.

Medical Domain Adaptation in Text

What about vocabulary?

Medical DA in Text: Context

Context

- In-Domain data is blurry:
  - Domain labels are not available / exhaustive ⇒ domain overlap
- Medical DA mainly considered in Images

Main Text Datasets

- DrugBank (Lu et al., 2014): Medical Translation
- WMT: sub-domain (Koehn et al., 2017)
- PubMed (Cohan et al., 2018)
- MIMIC (Johnson et al., 2019): free-text reports
- French:
  - FrenchMedMCQA (Labrak et al., 2022) : Multiple choice
  - CAS (Grabar et al., 2018) : clinical cases
- …

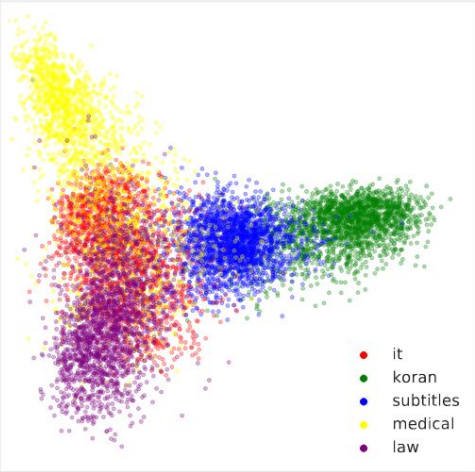


Figure 1: A 2D visualization of average-pooled BERT hidden-state sentence representations using PCA. The colors represent the domain for each sentence.

(Aharoni, Goldberg, 2020)

Medical DA in Text: Considered Evaluations

Shallow Metrics from Machine Translation

- BLEU (Papineri, 2002), SacreBLEU (Post, 2018)
- ROUGE (Lin, 2004)

Embedding-based Similarity Metrics

- BERTscore (Zhang, 2019)
- COMET (Rei, 2020)
- Domain Vocabulary Overlap (Yu, 2021)
- Domain Token Distribution Shift (Lin, 2023)
  - KL-Divergence
  - Token-Shift Rate (Afzal et al., 2024)

LLM-based Evaluation

- Reference-free GPT4 + CoT

Manual Evaluation by Annotators

Token Distribution Shift (Lin et al., 2023)

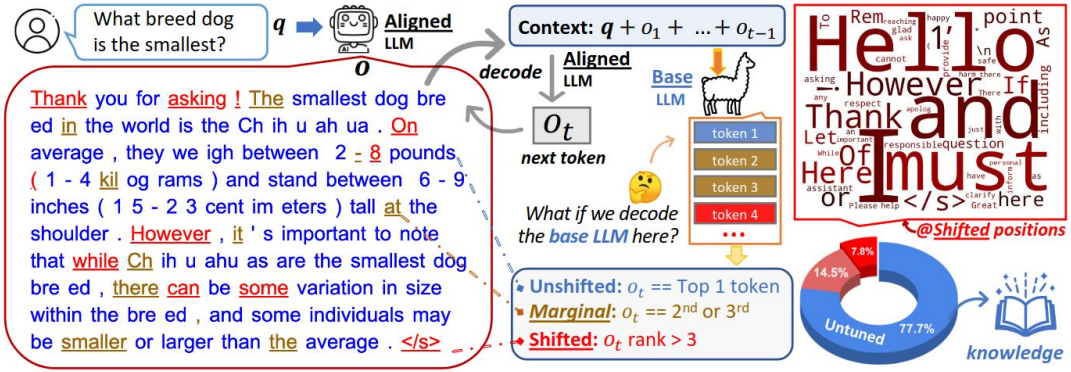


Figure 2: **Analyzing alignment with token distribution shift.** An aligned LLM (`llama-2-chat`) receives a query **q** and outputs a response **o**. To analyze the effect of alignment tuning, we decode the untuned version (`llama-2-base`) at each position t. Next, we categorize all tokens in **o** into three groups based on o_t's rank in the list of tokens sorted by probability from the base LLM. On average, 77.7% of tokens are also ranked top 1 by the base LLM (**unshifted** positions), and 92.2% are within the top 3 (+ *marginal*). Common tokens at **shifted** positions are displayed at the top-right and are mostly stylistic, constituting discourse markers. In contrast, knowledge-intensive tokens are predominantly found in unshifted positions. (More in Fig. 8 and Appendix. C)

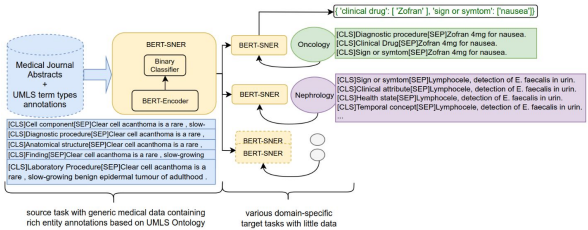
Lingering Questions and Challenges

Domain Adaptation or Specific Models (Jargon, DrBERT, etc.)?

- Smaller LLMs ⇒ challenging domain-shift (Azfal et al., 2024)
- Token Distribution Shift : Fine-tuning impacts only the stylistic tokens, not the vocabulary domain

Recent Vocabulary Based approaches

- RL-HF+SFT (Tian, 2024), Ontology (Kwon, 2022; Liang, 2023)



What about conversations?

- Task Adaptation ⇒ Additional context + Fillers & Hedges (Dinkar, 2023;Abulimiti 2023)
  - May be interpreted as noise in behavior and evaluation