Semantic and Pragmatic Properties of LLM's "Hallucinations" in the Medical Field

Ioana Buhnila[1]* **Georgeta Cislaru[2]** **Amalia Todirascu[3]**

[1] ATILF UMR 7118 (CNRS – University of Lorraine)

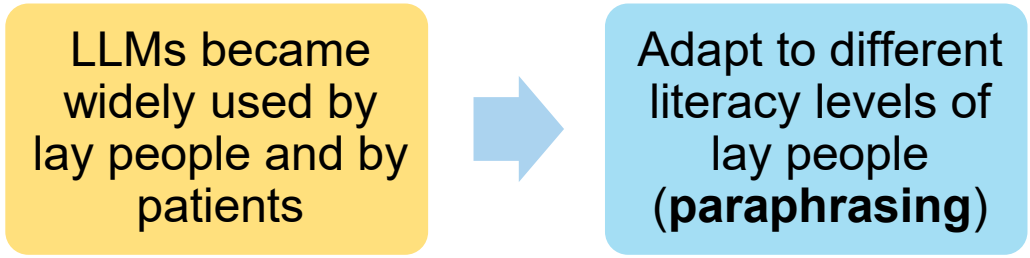
[2] MoDyCo UMR 7114 (CNRS – University Paris Nanterre)

[3] LiLPa UR 1339 (University of Strasbourg)

*ioana.buhnila@univ-lorraine.fr

**Agency and Intentions in AI, University of Göttingen,
May 16-17, 2024**

1. Context of the Study: LLMs in the Medical Field (I)



- **Medical paraphrases** = explain and simplify medical terms and make medical knowledge accessible to lay people (Grabar & Hamon, 2015; Buhnila, 2022)

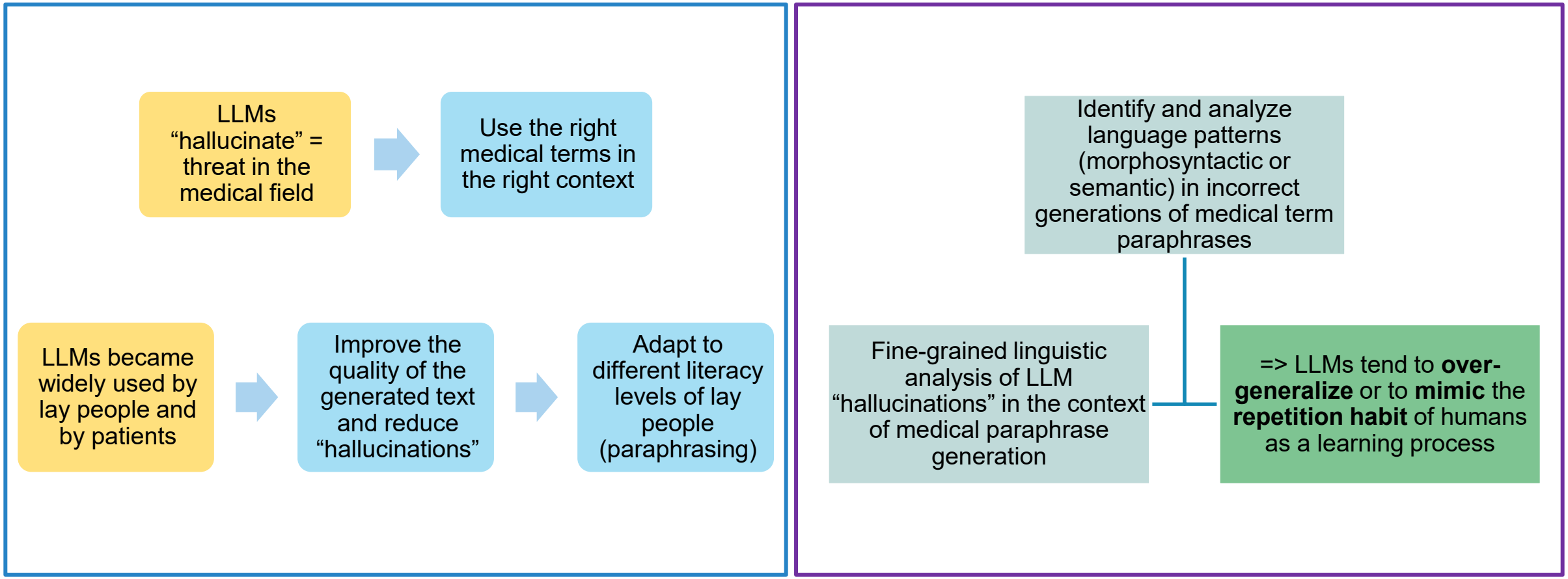
  Examples:

  [1] *placebo* (placebo) – *faux médicament* (fake medicine)

  [2] *hypotension* (hypotension) – *faible tension arterielle* (low blood pressure)

- **Simplified medical** content can also facilitate communication with patients (Pecout et al., 2019; Koptient & Grabar, 2020)

1. Context of the Study: Motivation and Goals (II)



1. Context of the Study: Resource Overview (III)

Paraphrase Dataset	Corpora	Type	Tokens	Sentences with terms and paraphrases	Tokens	Sentences with correct paraphrases
RefoMed **(Buhnila, 2023)**	**ClassYN** (Todirascu et al., 2012)	Scientific	1 007 049	**2 689**	**88 407**	**1 195**
		Popularization	772 374	**4 871**	**139 320**	**2 234**
	CLEAR Cochrane (Grabar et Cardon, 2018)	Scientific	2 840 003	**4 687**	**173 616**	**2 528**
		Popularization	1 515 051	**3 980**	**123 249**	**2 669**
	Total		***6 134 477***	***16 227***	***524 592***	***8 626***

1. Context of the Study: Data (IV)

- 2266 paraphrases generated for 480 terms

Generated Paraphrases	Correct Paraphrases	Incorrect Paraphrases	Partially correct *	Repetitions**	No Tag ***	Abbreviations****
2266 (100%)	725 (31,99 %)	779 (34,38 %)	211 (9,31 %)	5 (0,22 %)	545 (24,05 %)	1 (0,05 %)

* **Invented words**: *l'oxyde d'oxalate* (oxylate oxyde)

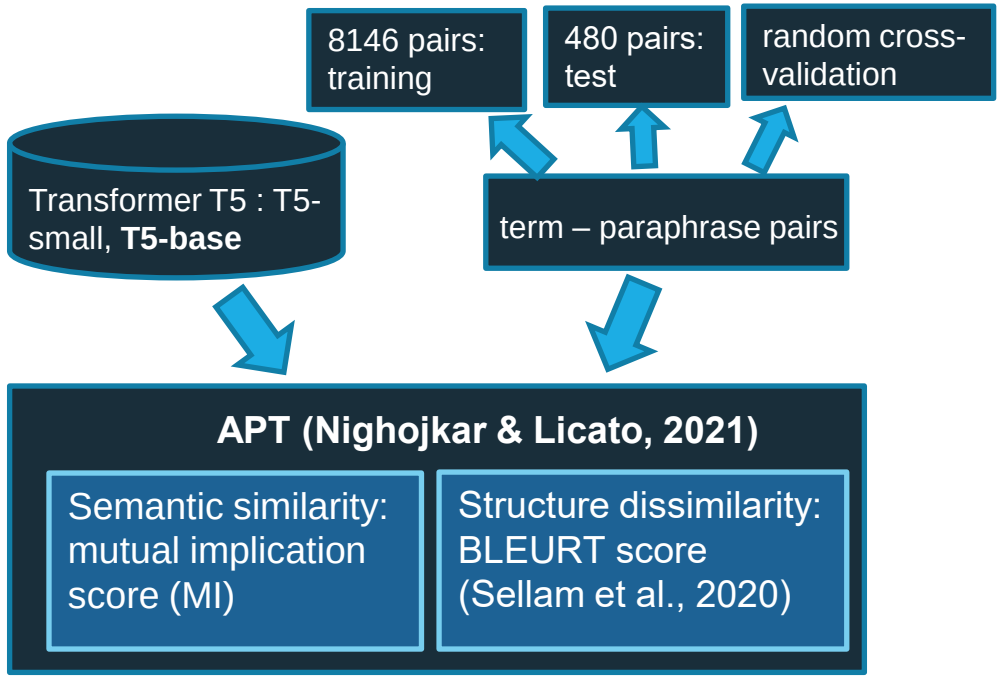
** **Repeated words**: *maladie maladie de Parkinson* (Parkinson disease disease)

*** **Abbreviations excluded from the linguistic annotation**: *médecine traditionnelle chinoise (MCT)* (traditional chinese medicine (TCM))

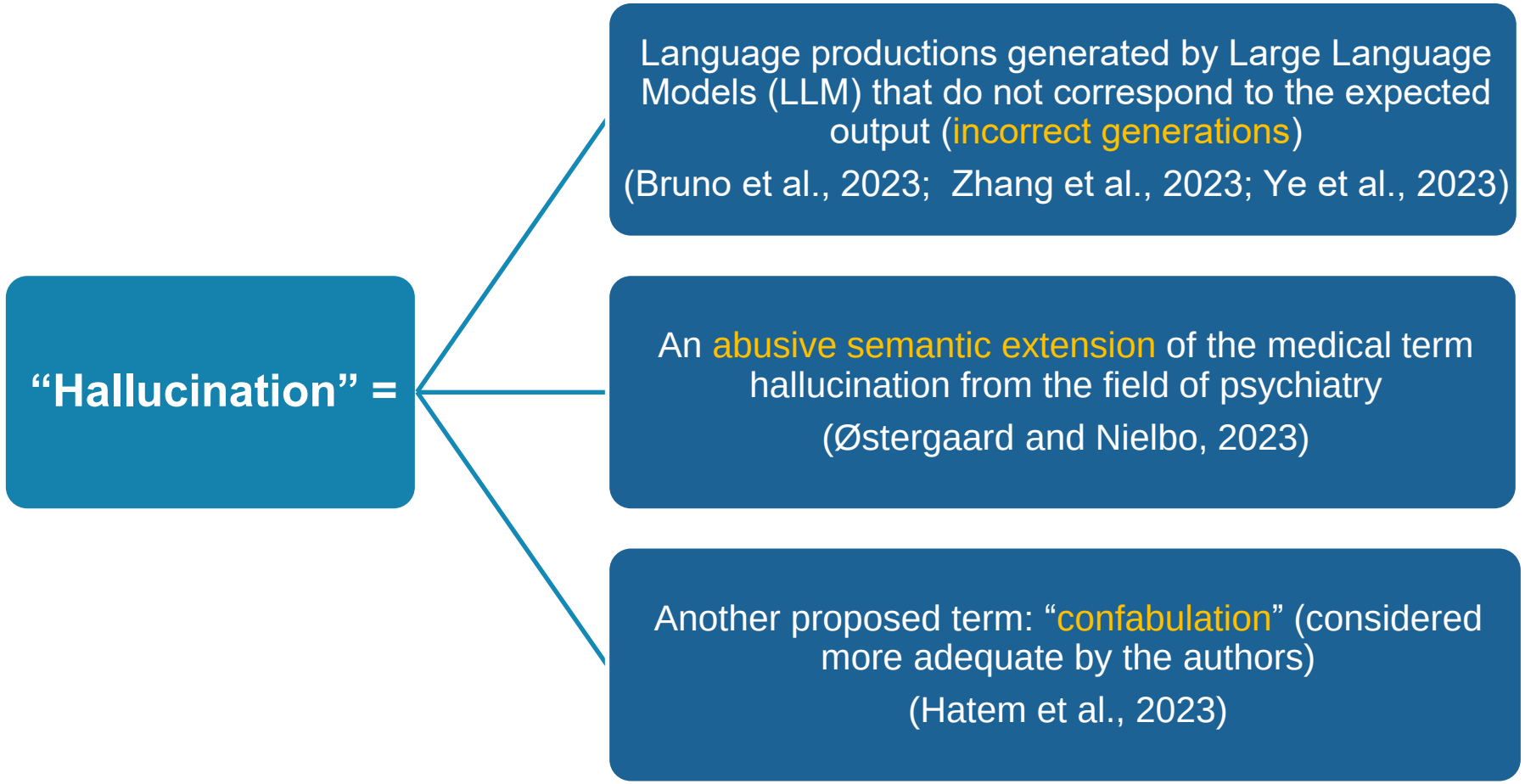
**** **Resting abbreviations**

1. Context of the Study: NLP Task and Tools (V)

- **NLP Task : Paraphrase Generation (PG) :** create new texts from data and LLMs (Gupta et al., 2018 ; Bowman et al., 2016).
  - **APT architecture** (*Adversarial Paraphrasing Task*) (Nighojkar & Licato, 2021)
  - **Transformer T5 LLM** (*Text-to-Text Transformer*) (Raffel et al., 2020)



2. LLMs' "Hallucinations": Definition and Concept Discussion (I)



2. LLMs "Hallucinations" (II)

[1] **Term**: *acetylcholine* (acetylcholine)

Correct paraphrase: *connue sous le nom de substance chimique* (known as a chemical substance)

"Hallucination": *c'est-à-dire l'oxyde d'oxalate de glycosyle de l'oxalate* (i.e. the glycosyl oxalate oxide of the oxalate)

[2] **Term**: *strabisme* (strabismus)

Correct paraphrase: *est une affection dans laquelle les yeux ne sont pas alignés normalement* (is a condition in which the eyes are not aligned normally)

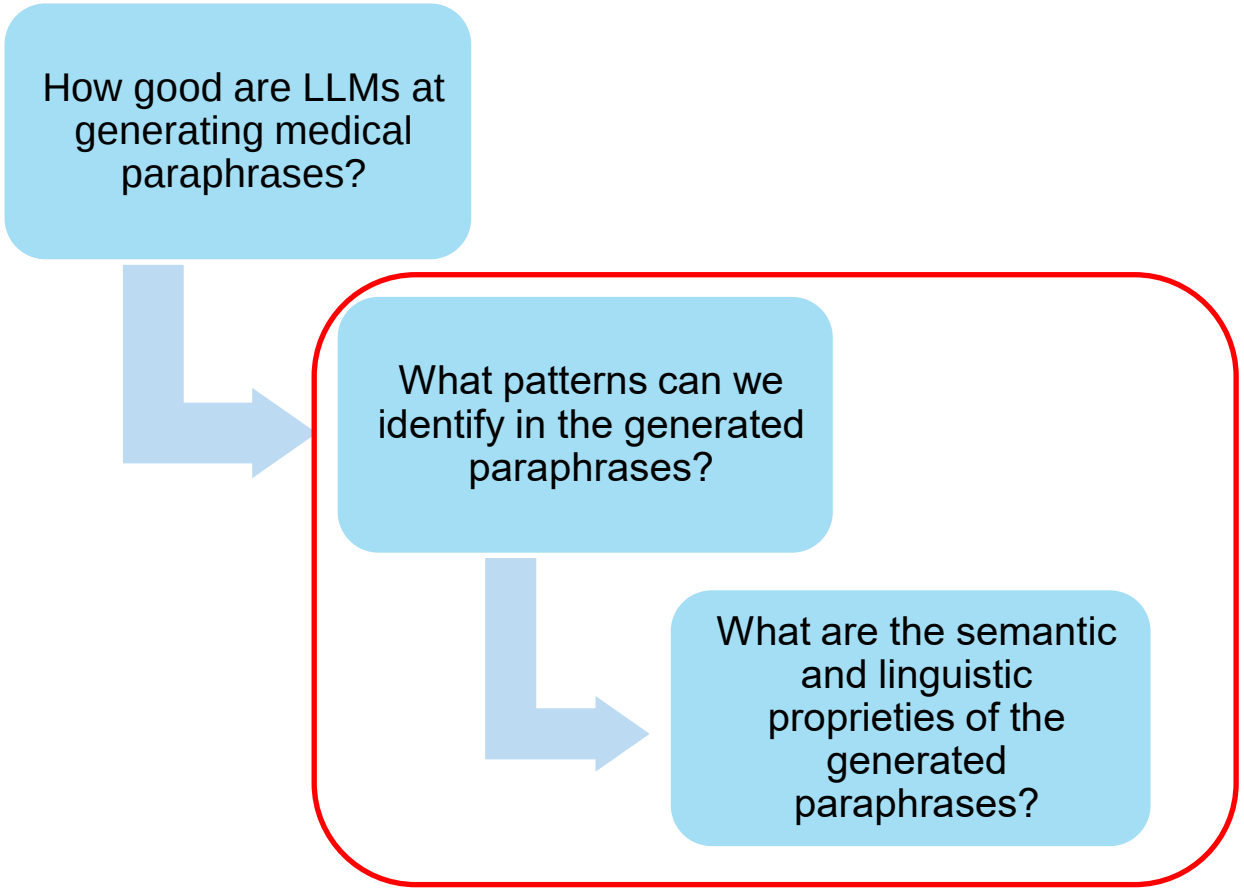
"Hallucination": *est une maladie associée à une stagiaire ayant tendance à se dissoudre rapidement* (is a condition associated with a trainee with a tendency to dissolve rapidly)

[3] **Term**: *sepsis* (sepsis)

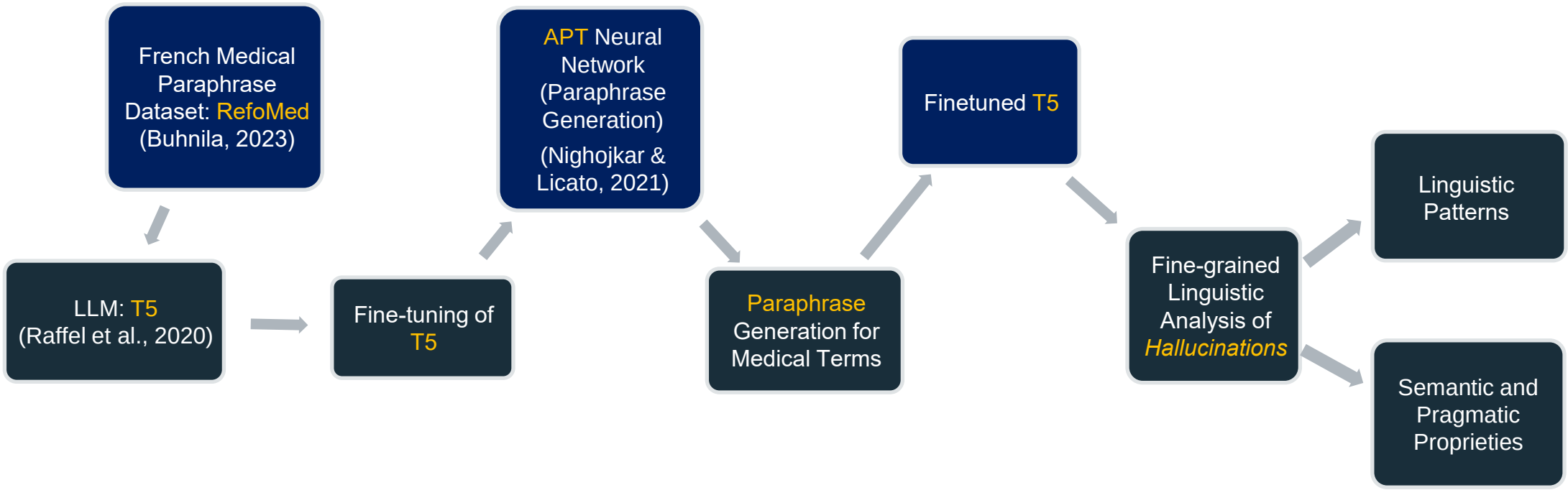
Correct paraphrase: *infection du pied diabétique* (diabetic foot infection)

"Hallucination": *une maladie neurodégénérative courante* (a common neurodegenerative disease)

3. Primary Research Questions



4. Method (I)


French Medical Paraphrase Dataset: RefoMed (Buhnila, 2023)
LLM: T5 (Raffel et al., 2020)
Fine-tuning of T5
APT Neural Network (Paraphrase Generation) (Nighojkar & Licato, 2021)
Paraphrase Generation for Medical Terms
Finetuned T5
Fine-grained Linguistic Analysis of Hallucinations
Linguistic Patterns
Semantic and Pragmatic Proprieties

4. Fine-grained Lingusitic Analysis (II)

- **Quantitative annotation of "hallucinations"**:
  - **Annotation of grammaticality**: correct, incorrect or partially correct;
  - **Type of inadequacy**: incorrect form, but comprehensible; correct form but not equivalent meaning; incorrect meaning;
  - **Lexical relations***: hyperonymy, hyponymy, synonymy, meronymy* (Sapoiu, 2013) (ex. [3]);
  - **Semantic-pragmatic functions:** *definition, exemplification, rephrasing, explanation* (Eshkol-Taravella & Grabar, 2017; Buhnila, 2022);

- **Quantitative analyses of "hallucinations"**:
  - **Degree of grammatical and semantic adequacy** of the generated text: invented words, genre/number mistakes, morphosyntaxctic errors (ex. [1] & [2]) ;
  - **Combinatorial discrepancies:** association of incompatible or foreign terms, mixture of different morphemes in invented words (ex. [1]).

5. Analysis Results (I)

- “Hallucinations” show:
  - incorrect medical knowledge (ex. [4])

[4] **Term**: *myopie* (myopia)

Correct paraphrase: *est un défaut de la vision qui se trouble lorsque des objets sont observés à distance* (is a vision defect that becomes blurred when objects are viewed from a distance)

“Hallucination” : *tel que un sombre myocarde* (such as a dark myocardium)

  - aberrant information (ex. [5])

[5] **Term**: *médecine traditionnelle chinoise* (traditional chinese medicine)

Correct paraphrase: c'est-à-dire une ancienne méthode chinoise (meaning an old chinese method)

“Hallucination” : *c'est-à-dire les méthodes d'admission orales dans les écoles ou les foyers de l'école, par exemple* (meaning oral admission methods into schools or school fosters, for example)

5. Analysis Results (II)

- “Hallucinations” show similarities with the exploitation of **patterns** in natural language practice:
  - Abusive **generalizations** of patterns not validated by usage (hypernyms and hyponyms)
  - Comparable processes to **foreign language learning** (rephrasing and definitions)
  - **Frequency calculations** in the absence of contextualized statistical weights (Piantadosi (2014), Zipf law).
  - Abusive use of **repetition** as a truth effect (Hasher et al., 1977; Dechêne et al., 2010)

5. The Place of Intention in LLMs' Paraphrases (III)

- LLMs are often antropomorphized and associated with an **"illusion of content"** (Ostertag, 2023)
- However, LLMs are **obeying rules** and **not communicative intentions** (while humans respond to both)
- Yet LLMs are oriented towards **intention recognition** (Manias et al., 2024)
- Good calculation of user's intentions may not prevent the generation of "confabulations" or non-appropriate paraphrases
- No specific meaning can be inferred, only **patterns** can be identified
- Intentions ≠ Pattern reproduction ≠ Meaning
- In return, this also questions the place of **pattern reproduction in human languages**

6. Conclusion and Future Work

Computer sciences: "hallucinations" raise questions about the right language model parameters for text generation

Linguistics: linguistic usages of patterns in human language and their reproduction by generative LLMs

Solutions? **RAG** systems for scientific grounding

(Lewis et al., 2020; Asai et al., 2023; Jeong et al., 2024)

7. References

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. arXiv preprint arXiv:2310.11511.

Buhnila, I. (2022). Identifying medical paraphrases in scientific versus popularization texts in French for laypeople understanding. *In Proceedings of the Third Workshop on Scholarly Document Processing, COLING 2022*, pages 69–79.

Buhnila, I. (2023). *Une méthode automatique de construction de corpus de reformulation*. PhD Thesis, University of Strasbourg, June 2023.

Dechêne, A., Stahl, C., Hansen, J. & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review 14(2):* 238-257. doi:10.1177/1088868309352251.

Eshkol-Taravella, I. & Grabar, N. (2017). Taxonomy in reformulations from a corpus linguistics perspective. *Syntaxe et semantique, 18(1),* 149-184.

Grabar, N. & Cardon, R. (2018). CLEAR - simple corpus for medical French. *In Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Association for Computational Linguistics.

Grabar, N. & Hamon, T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. *TALN*, pages 182–195.

Hasher, L., Goldstein, D. & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior 16(1):* 107-112. doi:10.1016/S0022-5371(77)80012-1.

Hatem R., Simmons B. & Thornton JE. (2023). Chatbot Confabulations Are Not Hallucinations. *JAMA Intern Med. 2023;183(10)*:1177.

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. arXiv preprint arXiv:2401.15269. doi:10.1001/jamainternmed.2023.4231

Koptient, A. & Grabar, N. (2020). Fine-grained text simplification in French: steps towards a better grammaticality. *In International Symposium on Health Information Management Research*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Nighojkar, A. & Licato, J. (2021). Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint.* arXiv:2106.07691.

Manias, D. M., Chouman, A., & Shami, A. (2024). Semantic Routing for Enhanced Performance of LLM-Assisted Intent-Based 5G Core Network Management and Orchestration. arXiv preprint arXiv:2404.15869.

Østergaard, SD. & Nielbo, KL. (2023). False Responses From Artificial Intelligence Models Are Not Hallucinations. *Schizophrenia Bulletin, Volume 49, Issue 5,* September 2023, pages 1105–1107, https://doi.org/10.1093/schbul/sbad068

Ostertag, G. (2023). Meaning by Courtesy: LLM-Generated Texts and the Illusion of Content. The American Journal of Bioethics, 23(10), 91–93. https://doi.org/10.1080/15265161.2023.2249851

Pecout, A. Tran TM, & Grabar, N. (2019). Améliorer la diffusion de l'information sur la maladie d'Alzheimer: étude pilote sur la simplification de textes médicaux. *Etudes de linguistique appliquée, 3(195):* 325–341.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S, Matena, M., Zhou, Y., Li, W. & Liu, PJ. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21(1):*5485–5551.

Todirascu, A., Padó, S., Krisch, J., Kisselew, M. & Heid, U. (2012). French and German corpora for audience-based text type classification. *In LREC*, volume 2012, pages 1591–1597.

Ye H., Liu, T., Zhang, A., Hua, W. & Jia, W. (2023). Cognitive Mirage: A Review of Hallucinations in Large Language Models. *arXiv:2309.06794v1* [cs.CL]

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, AT, Bi, W., Shi, F. & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv cs.CL eprint 2309.01219*, https://doi.org/10.48550/arXiv.2309.01219

Thank you for your attention !