Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Pantagruel : des encodeurs auto-supervisés unifiés pour le texte et la parole

Phuong-Hang Le[1,9] Valentin Pelloin[2] Arnault Chatelain[4] Maryem Bouziane[3]
Mohammed Ghennai[1] Qianwen Guan[5] Kirill Milintsevich[2]
Salima Mdhaffar[3] Aidan Mannion[1] Nils Defauw[6] Shuyue Gu[5]
Alexandre Audibert[1] Marco Dinarelli[1] Yannick Estève[3] Lorraine Goeuriot[1]
Steffen Lalande[2] Nicolas Hervé[2] Maximin Coavoux[1] François Portet[1]
Étienne Ollion[4] Marie Candito[5] Maxime Peyrard[1] Solange Rossato[1]
Benjamin Lecouteux[1] Aurélie Nardy[7] Gilles Sérasset[1] Vincent Segonne[8]
Solène Evain[10] Diandra Fabre[1] Didier Schwab[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] INA (Institut National de l'Audiovisuel), 4 Avenue de l'Europe, 94366 Bry-sur-Marne, France
[3] Avignon Université, LIA, France
[4] CREST (École Polytechnique, ENSAE, CNRS), 5 avenue Le Chatelier, 91120 Palaiseau, France
[5] LLF (Université Paris Cité and CNRS), UFRL Olympe de Gouges,
13 place Paul Ricoeur, 75013 Paris, France
[6] Univ. Grenoble Alpes, EFELIA-MIAI, IUT2 Grenoble, LIG, 38000 Grenoble, France
[7] Univ. Grenoble Alpes, Lidilem, 38000 Grenoble, France
[8] Université Bretagne Sud, CNRS, IRISA, France
[9] Saclay AI, France
[10] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
didier.schwab@univ-grenoble-alpes.fr

RÉSUMÉ

Nous publions les modèles Pantagruel, une nouvelle famille d'encodeurs autosupervisés pour le texte et la parole en français. Plutôt que de prédire des cibles adaptées à chaque modalité, comme des tokens textuels ou des unités de parole, Pantagruel apprend des représentations cibles contextualisées dans l'espace des caractéristiques, ce qui permet à des encodeurs spécifiques à chaque modalité de mieux capturer les régularités linguistiques et acoustiques. Des modèles distincts sont préentraînés sur de vastes corpus français, notamment Wikipedia, OSCAR et CroissantLLM pour le texte, ainsi que MultilingualLibriSpeech, LeBenchmark et INA-100k pour la parole. INA-100k est un nouveau corpus de 100 000 heures d'audio en français, issu des archives de l'Institut national de l'audiovisuel (INA), dépôt national des émissions de radio et de télévision françaises, fournissant des données audio très diversifiées. Nous évaluons Pantagruel sur un large éventail de tâches aval couvrant les deux modalités, y compris celles des principaux benchmarks français tels que FLUE ou LeBenchmark. Sur l'ensemble de ces tâches, les modèles Pantagruel obtiennent des performances compétitives, voire supérieures, par rapport à de solides bases de référence françaises comme CamemBERT, FlauBERT et LeBenchmark 2.0, tout en conservant une architecture commune capable de traiter de manière transparente des entrées de parole ou de texte. Ces résultats confirment l'efficacité d'objectifs auto-supervisés dans l'espace des caractéristiques pour l'apprentissage de représentations en français et mettent en évidence Pantagruel comme une base robuste pour la compréhension multimodale

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 38–39.

parole-texte.

ABSTRACT

Pantagruel : Unified Self-Supervised Encoders for French Text and Speech

We release Pantagruel models, a new family of self-supervised encoder models for French text and speech. Instead of predicting modality-tailored targets such as textual tokens or speech units, Pantagruel learns contextualized target representations in the feature space, allowing modality-specific encoders to capture linguistic and acoustic regularities more effectively. Separate models are pre-trained on large-scale French corpora, including Wikipedia, OSCAR and CroissantLLM for text, together with MultilingualLibriSpeech, LeBenchmark, and INA-100k for speech. INA-100k is a newly introduced 100000-hours corpus of French audio derived from the archives of the Institut National de l'Audiovisuel (INA), the national repository of French radio and television broadcasts, providing highly diverse audio data. We evaluate Pantagruel across a broad range of downstream tasks spanning both modalities, including those from the standard French benchmarks such as FLUE or LeBenchmark. Across these tasks, Pantagruel models show competitive or superior performance compared to strong French baselines such as CamemBERT, FlauBERT, and LeBenchmark 2.0, while maintaining a shared architecture that can seamlessly handle either speech or text inputs. These results confirm the effectiveness of feature-space self-supervised objectives for French representation learning and highlight Pantagruel as a robust foundation for multimodal speech–text understanding.

MOTS-CLÉS : apprentissage auto-supervisé, JEPA, data2vec, modèles de langue français, encodeurs parole et texte, apprentissage de représentations multimodales, architecture prédictive à embeddings joints, modélisation prédictive.

KEYWORDS: self-supervised learning, JEPA, data2vec, French language models, speech and text encoders, multimodal representation learning, joint-embedding predictive architecture, predictive modeling.

ARTICLE ACCEPTÉ À : LREC 2026 : The Fifteenth biennial Language Resources and Evaluation Conference, Palma, Mallorca, Spain, May 11-16, 2026. .
URL : https://lrec.elra.info/lrec2026-main-799

Remerciements

Ces travaux de recherche ont été partiellement financés par l'Agence nationale de la recherche (ANR), projet « PANTAGRUEL », ANR-23-IAS1-0001. Ils ont également bénéficié du soutien du projet CREMA (Coreference REsolution into MAchine translation), financé par l'ANR, contrat n° ANR-21-CE23-0021-01. Ils ont par ailleurs reçu un financement public géré par l'ANR dans le cadre de la stratégie France 2030, référence ANR-23-IACL-0006. Ces travaux ont également été soutenus par l'ANR à travers la chaire MIAI « IA et Langage » (ANR-19-P3IA-0003), la chaire MIAI « Socialisation et Langage à l'école », et la chaire AugmentIA sous le mécénat d'Artelia et de Grenoble INP (ANR-23-IACL-0006).

Ce travail a été réalisé en utilisant les ressources HPC de GENCI à IDRIS et CINES dans le cadre des allocations 2022-A0131013801, 2023-A0151013801, 2024-A0171013801, 2024-A0161015074 et 2025-A0191013801 sur les supercalculateurs Jean Zay et Adastra.