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

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

Abstract

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 100 000-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.

Keywords: self-supervised, data2vec, JEPA, French, speech, text, representation learning, predictive modeling

1. Introduction

Mirroring trends in other languages, self-supervised encoders have become the standard backbone for French speech and language processing. Text models such as FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020), together with LeBenchmark (Evain et al., 2021b; Parcollet et al., 2024) for speech, have established strong baselines across a variety of downstream tasks. Yet most prior work relies on token-level reconstruction (Devlin et al., 2019; Baevski et al., 2020; Warner et al., 2025), which can under-utilize the structural regularities of continuous signals and hinders a unified treatment of text and speech. Recent predictive approaches (LeCun, 2022; Assran et al., 2023; Baevski et al., 2023) that learn contextualized targets in feature space offer a promising alternative, enabling modality-specific encoders to capture richer linguistic and acoustic structure beyond the surface form.

In this paper, we introduce **Pantagruel**, a family of French self-supervised encoders for text and speech, trained separately using the data2vec 2.0 architecture (Baevski et al., 2022), an instance of the Joint Embedding Predictive Architecture (JEPA) framework (LeCun, 2022). Pantagruel follows a teacher–student training paradigm in which the student predicts masked latent representations produced by a teacher that observes the full, unmasked input. Such representation-based objectives have proven highly effective for vision (Assran et al., 2023; Mo and Tong, 2024) and audio (Fei et al., 2023; Tuncay et al., 2025; Yuksel et al., 2025),

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10168–10191
11–16 May 2026.

yet they remain underexplored for text. Recent work suggests that textual tokens are compact, semantically dense units with minimal low-level variability, leaving limited room for embedding-based methods to improve over input-level approaches (Van Assel et al., 2025). Our experiments on text-based models also confirm this hypothesis. Therefore, for text, we propose to augment the feature-space objective with masked language modeling (MLM, Devlin et al., 2019) to better capture fine-grained syntactic and semantic information, yielding stronger textual representations while retaining the benefits of contextualized target prediction. Our self-supervised models are released on the HuggingFace hub[1].

To support large-scale pre-training in French, we curate substantial corpora for each modality. For text, we use Wikipedia, OSCAR (Martin et al., 2020) and CroissantLLM (Faysse et al., 2024) datasets. For speech, we assemble a diverse collection spanning read, spontaneous, professional, and broadcast speech audio data, including 14 000 audio hours from LeBenchmark (Evain et al., 2021b) and 100 000 audio hours from INA-100k, a new corpus derived from the archives of France's National Audiovisual Institute (INA) that we introduce in this paper. We evaluate Pantagruel on a broad suite of downstream tasks in both modalities and compare it to three main French baselines: FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020) for text, and LeBenchmark (Evain et al., 2021a,b; Parcollet et al., 2024) for speech.

Contributions First, we release Pantagruel, a family of French self-supervised encoders for speech and text based on the data2vec 2.0 and JEPA frameworks, which allows training models on different modalities using the same framework. Second, we investigate embedding-based prediction objectives for text, an underexplored regime, and show that combining feature-space prediction with MLM yields competitive results for French text encoders. Third, we study the impact of the large-scale INA-100k broadcast corpus on the models' performances. Finally, we provide a unified evaluation across speech and text tasks, where Pantagruel consistently matches or improves over established French baselines.

2. Related Work

Self-supervised learning (SSL) has driven rapid progress in several domains, in particular in text and speech processing. In text, encoder-only models such as BERT (Devlin et al., 2019) learn bidirectional representations through the MLM objective, with subsequent refinements improving attention mechanisms, efficiency, and context length (Clark et al., 2020; He et al., 2020; Lan et al., 2020; Warner et al., 2025), while the GPT family (Radford et al., 2018, 2019; Brown et al., 2020) popularized autoregressive pre-training for generative tasks. In speech, wav2vec2.0 (Baevski et al., 2020) introduced contrastive learning over quantized representations, while HuBERT (Hsu et al., 2021) adopted masked prediction of discrete clusters, and WavLM (Chen et al., 2022) improved robustness through denoising objectives. Initially developed for English, these self-supervised frameworks were rapidly extended to multilingual settings (Devlin et al., 2019; Conneau et al., 2020; Scao et al., 2022; Shliazhko et al., 2024; Babu et al., 2022).

For French, many studies have adapted these self-supervised architectures to both modalities. In text, FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020) were the first French BERT variants, trained on large-scale corpora and demonstrating that language-specific pre-training outperforms multilingual models on downstream tasks. Subsequent efforts explored efficiency trade-offs with compact models such as FrALBERT (Cattan et al., 2022), LePetit (Micheli et al., 2020), and D'AlemBERT (Gabay et al., 2022). Encoder-decoder and decoder-only approaches have also been explored with BARThez (Eddine et al., 2021), PAGnol (Launay et al., 2022), and Cedille (Müller and Laurent, 2022), while more recent large-scale models such as CroissantLLM (Faysse et al., 2024) further bridge the gap with multilingual systems. Continued refinement of French encoders has led to improved variants including CamemBERTa (Antoun et al., 2023), CamemBERT 2.0 (Antoun et al., 2024), and CamemBERTav2 (Antoun et al., 2024), which revisit data filtering, tokenization, and training recipes to yield stronger representations. For speech, the LeBenchmark initiative (Evain et al., 2021b; Parcollet et al., 2024) extends the pre-training paradigm to spoken French, providing wav2vec2.0-based models and evaluation resources for speech understanding in French. For spoken language understanding, FlauBERT-Oral (Pelloin et al., 2022) provides textual representations adapted to transcribed spoken French. Altogether, these efforts contribute to a comprehensive ecosystem of French models covering written, specialized, and oral modalities.

Orthogonal to the above advances, an active line of research has explored embedding-level predictive objectives, where models learn to predict latent representations of masked regions from visible context. A notable example is data2vec (Baevski et al., 2022, 2023), which introduced a modality-agnostic framework for predicting contextualized representations across speech, vision, and text. Closely related ideas were formalized in the JEPA

[1] `https://huggingface.co/PantagrueLLM`

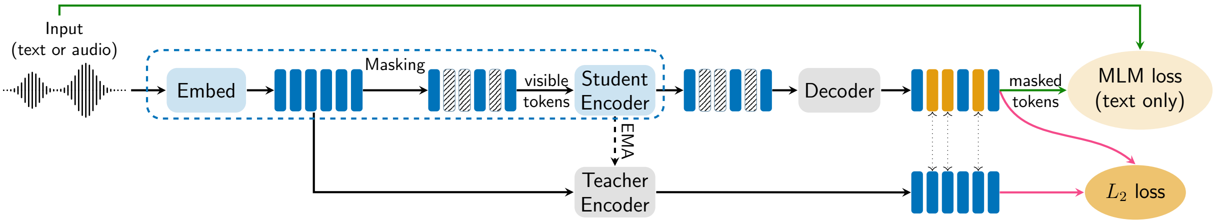

Figure 1: Overview of the Pantagruel model architecture. The network starts with a modality-specific pre-net to extract feature vectors from the input text/speech sequence. These features are input to a teacher encoder, while randomly chosen visible tokens (in blue) are input to a student encoder. A lightweight decoder predicts the teacher's latent representations from the student's outputs. For text input, an additional masked language modeling (MLM) loss is used. The teacher's parameters are updated as an exponential moving average (EMA) of the student's. After training, only the embedding layer and the student encoder are used for fine-tuning on downstream tasks.

framework (LeCun, 2022), which emphasizes prediction in the latent space. This paradigm has since enabled successful applications across domains: I-JEPA (Assran et al., 2023), V-JEPA (Bardes et al., 2024; Assran et al., 2025) for image and video; A-JEPA (Fei et al., 2023), Audio-JEPA (Tuncay et al., 2025), and WavJEPA (Yuksel et al., 2025) for audio. Similar to data2vec, most of these work rely on the teacher-student setup to prevent representation collapse. Recently, LeJEPA (Balestriero and LeCun, 2025) introduced a theoretically grounded formulation based on Latent-Euclidean regularization, removing the need for such heuristics. We leave its exploration to future work. Beyond data2vec, JEPA-style pre-training remains largely underexplored for text, where the input-level MLM objective continues to dominate. In this work, we investigate JEPA-style representation learning for French speech and text using data2vec2.0 architecture (Baevski et al., 2023). For speech, we adopt a pure feature-based objective. For text, we propose a hybrid approach that combines JEPA-style representation prediction with MLM to leverage their complementary strengths, inspired by (Huang et al., 2025).

3. Models and Pre-training Framework

Our SSL models for speech are entirely based on data2vec 2.0 (Baevski et al., 2023). For text, we extended the data2vec 2.0 loss to include the MLM objective (Devlin et al., 2019; Liu et al., 2019), which we found to produce better textual representations. Training is performed separately for each modality, following the data2vec approach.

3.1. Framework overview

An overview of our pre-training framework is presented in Figure 1. Given an input audio or a sequence of text tokens, a modality-specific **pre-net** is used to extract a sequence of feature vectors. For speech, this pre-net is typically a small CNN while for text it is a standard embedding layer. Then, the obtained sequence is fed into a **teacher encoder**. At the same time, randomly selected items (called *visible tokens*) of this sequence are fed into the **student encoder** (the unselected ones are called *masked tokens*). A lightweight, modality-specific convolutional **decoder** predicts the teacher's latent representations from the student's encoder outputs and is used only in pre-training. Both teacher and student encoders share the same Transformer encoder architecture (Vaswani et al., 2017). The teacher's parameters are maintained as an exponential moving average (EMA) of the student's, gradually stabilizing during training. Conceptually, this approach can be viewed as an instance of the predictive modeling paradigm (LeCun, 2022; Assran et al., 2023), sharing its principle of learning by predicting latent representations rather than reconstructing raw inputs. The idea of using *masked feature prediction* within a *self-distillation* setup is applicable across diverse modalities, motivating our exploration of this framework as a step toward building a truly multi-modal encoder.

Masked feature prediction The training loss is defined as the L_2 distance between the student predictions (*i.e.*, the outputs of the decoder), and the teacher encoder's representations, computed over the masked regions. The teacher representations are not obtained from the last Transformer layer but by averaging the last K layers, for better contextualization. To improve efficiency, multiple masked versions of each example share a single teacher forward pass.

Extension for better textual representations While embedding-based objective excels in speech and vision domains (Assran et al., 2023; Fei et al., 2023; Bardes et al., 2024; Assran et al., 2025), token-level objectives remain dominant in text modeling. Unlike continuous signals, text is discrete and sparse, thus predicting global contextual embeddings alone may not adequately capture fine-grained syntactic and semantic information essential for downstream tasks (Mo and Yun, 2024). Combining masked feature prediction with token-level objectives, as explored in recent work (Huang et al., 2025), offers a promising direction. Following this idea, we enhance our non-generative text models by combining the MLM objective with the data2vec loss, encouraging the model to capture local semantics alongside rich contextual representations. We find that using the student decoder to predict masked text tokens (see Figure 1) outperformed the BERT-style implementation, which requires a second forward pass in our framework. We thus adopt the decoder-based approach for our text models.

3.2. Pantagruel model configurations and implementation details

Tokenizers We train our tokenizer on a subset of the CroissantLLM dataset (Faysse et al., 2024) with customized settings tailored for French, such as normalizing apostrophe variants and treating them as single tokens in elided words (*e.g.*, *quelqu'*, *c'*). For comparison, we also experimented with the CamemBERT tokenizer (Martin et al., 2020, denoted *camtok*). While CamemBERT followed the original BERT (Devlin et al., 2019) and used a vocabulary size of 32K tokens, we used a larger size of 50K, following more recent works such as GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), and ModernBERT (Warner et al., 2025). Later in Section 5, we include two more tokenizers that we train with different settings to investigate the effect of the tokenizer on our models.

Model configurations Following standard Transformer architectures, we propose two model configurations: **base** and **large**. The base variant has 12 layers, 8 attention heads, and 768 hidden dimensions. For the large variant, these values are, respectively, 24, 16, and 1024. A summary of our models is given in Table 1. For text models, we only include the base variant as we experienced training instability with the large variant. It should be noted that the data2vec papers (Baevski et al., 2022, 2023) also did not report results for the large variant on text tasks. Investigating large text models is left for future work.

Loss functions Our speech models follow data2vec 2.0 and use only masked feature prediction, $\mathcal{L}_{\text{speech}} = \mathcal{L}_{L_2}$ (see Section 3.1). For text, we propose a hybrid loss $\mathcal{L}_{\text{text}} = \mathcal{L}_{L_2} + \lambda \mathcal{L}_{\text{mlm}}$, where λ decays linearly from λ_{start} to λ_{end} over N_λ training steps and remains fixed thereafter. We find this schedule yields more stable training and better performance than a constant weight, similar to prior work showing that auxiliary objectives are most beneficial early but can interfere with the main objective later (Du et al., 2018; Fu et al., 2019).

	Model name	Param	Trained on
Text	Pantagruel-B-camtok-Wk	110M	Wikipedia
	Pantagruel-B-Wk	125M	Wikipedia
	Pantagruel-B-Wk-MLM	125M	Wikipedia
	Pantagruel-B-Osc-MLM	125M	OSCAR
	Pantagruel-B-Crs-MLM	125M	CroissantLLM
Speech	Pantagruel-B-1k	93M	LibriSpeech
	Pantagruel-B-14k	93M	LeBenchmark
	Pantagruel-L-14k	313M	LeBenchmark
	Pantagruel-L-114k	313M	LeBenchmark + INA

Table 1: Pantagruel model names and configurations. Suffixes: "B/L" mean *base/large* architecture, and "camtok" denotes the CamemBERT tokenizer. For text, the third suffix specifies the dataset name, while for speech it indicates the training dataset size in hours (*e.g.*, "1k" means 1 000 audio hours).

Implementation details Following data2vec 2.0 recipe, we used the Adam optimizer (Kingma and Ba, 2015) with cosine learning rate scheduler (Loshchilov and Hutter, 2017), where the learning rate η is increased linearly from $\eta_{\min}$ to the configured maximum learning rate $\eta_{\max}$ during a warmup of N_{warmup} steps, then decreases smoothly following a cosine curve until training reaches the specified maximum number of updates $N_{\max}$ where $\eta = \eta_{\min}$. Details of the pre-training hyperparameters are provided in Table 10 and 11 for speech and text models, respectively.

4. Datasets and Resources

This section describes data collection and organization for French text and speech. Our goal was to maximize dataset size, including both aligned (audio with transcripts) and unaligned data. In the following, we detail the publicly available datasets, the proprietary INA corpus, and preprocessing steps.

4.1. Text datasets

Table 2 summarizes the data used to train our text models. We leveraged monolingual French data from three public sources: the 2019 Wikipedia dump from FlauBERT (Le et al., 2020) (876M tokens), the French portion of OSCAR (Ortiz Suárez et al., 2019) (32.7B tokens), and the French subset of the CroissantLLM dataset (Faysse et al., 2024), which is composed of web data (292B tokens), legal and administrative (5.3B tokens), cultural (2.7B tokens), encyclopedia (2B tokens), and industrial data (191M tokens). The corresponding models trained on these datasets are listed in Table 1.

Dataset	Tokens	Size	Type
Wikipedia (Jul 2019)	876M	4 GB	Encyclopedia
OSCAR (Nov 2018)	32.7B	138 GB	Web crawl
CroissantLLM	302.2B	1.3 TB	Mix

Table 2: Statistics of pre-training text corpora.

4.2. Speech datasets

The complete list of datasets considered in this study for speech and their statistics are summarized in Appendix C. We used the dataset collection of LeBenchmark for French which gathered a wide variety of French speech corpora covering different accents, acted emotions, telephone dialogues, read speech, spontaneous sentences, and professional speech. This collection includes Multilingual LibriSpeech (MLS) dataset (Pratap et al., 2020). It is worth noting that some of these corpora are transcribed either manually or automatically and are therefore representative of aligned speech/text data. We refer the reader to Parcollet et al. (2024) for the details.

However, these remain far from SOTA models, such as Whisper (Radford et al., 2022, trained on 680k hours of labeled data) or Google USM (Zhang et al., 2023, trained on more than 12 million hours of multilingual data). Thus, to increase the pre-training speech dataset, a corpus of 100 000 hours of audio content is obtained in partnership with INA, the French National Audiovisual Archive which we will refer to as INA-100k in the rest of the paper. In France, INA is in charge of collecting and archiving TV and radio broadcast since 1975. We evaluate that 75% of this audio content is speech.

The pre-processing steps used to prepare INA-100k are described below. An initial corpus of 473k hours of content, broadcast on 113 French TV and Radio channels, from 1940 to 2022 was collected, covering various kinds of audiovisual content: news, adverts, documentaries, game shows, movies, musics, cartoons, sports, etc.

We first applied the audio deduplication tool presented by Chenot and Daigneault (2014) to remove duplicate content. This tool extracts audio fingerprints. When at least four consecutive similar fingerprints are detected, all other matches are discarded. Deduplication helps control distribution and reduce bias (Lee et al., 2022). Multiple standard benchmark evaluation corpora are also based on French TV broadcast content: we used the same tool to remove them from the INA corpus. This include, among others, ESTER1 (Gravier et al., 2004), ESTER2 (Galliano et al., 2009), EPAC (Estève et al., 2010), QUAERO (Boudahmane et al., 2011), ETAPE (Gravier et al., 2012), REPERE (Giraudel et al., 2012). Overall, these two deduplication steps removed 154k hours (33%) from the initial corpus. Finally we randomly sampled 12M audio chunks of 30s to create a 100 000 hours training corpus, denoted as INA-100k.

5. Ablation study

In this section, we investigate the effect of different tokenizers on our text models and the impact of varying pre-training data sizes on our speech models. The ablation results are reported on the validation sets of two representative tasks: natural language inference (XNLI dataset) for text and automatic speech recognition (CommonVoice dataset) for speech. Further details on these datasets are provided in Sections 6.1 and 6.2.

Effects of different text tokenizers In addition to our main “customized” tokenizer (used in Section 6) and the CamemBERT one (see Section 3.2), we trained two other tokenizers, “default” and “basic”, using HuggingFace’s *tokenizers* library. We refer to the documentation of this library for the details of these settings. The ablation study is conducted on the Wikipedia dataset with models trained for 500K steps (approximately 10 epochs over the dataset). As shown in Table 3, our main tokenizer outperforms the others on the XNLI dataset.

Tokenizers	Accuracy
Camtok	76.94$\pm$0.51
Default	76.34$\pm$0.52
Basic	77.02$\pm$0.55
Customized	**77.40**$\pm$0.46

Table 3: Results using different tokenizers on the dev sets of the XNLI dataset (average over 10 runs). “Customized” denotes our main tokenizer.

Effects of pre-training data size for speech We trained four speech models: two models using the base architecture on 1K hours of LibriSpeech and 14K hours of the LeBenchmark dataset, and two models using the large architecture on 14K hours of LeBenchmark and 114K hours, which combines LeBenchmark 14K hours with 100K hours from the INA dataset. Table 4 reports the Word Error Rate (WER) on the CommonVoice dev sets. Increasing the data from 1K to 14K hours for the base architecture yields a modest improvement in WER (from 8.92 to 8.46). For the large architecture, however, increasing the data from 14K to 114K hours slightly hurts the results. We hypothesize that this could be due to the data nature of INA being quite different from CommonVoice, while at the same time the model capacity being potentially too limited to sufficiently fit this large dataset.

Model/Data	Base		Large	
	1K	14K	14K	114K
WER	8.92	8.46	6.95	7.07

Table 4: WER on CommonVoice dev sets across model architectures and data sizes.

6. Benchmarks

This section presents text and speech downstream tasks used for evaluation. Unless otherwise noted, all results are averaged over five runs with different random seeds. We use ↑ (respectively ↓) to indicate that higher (respectively lower) values correspond to better performance.

An overview of the benchmarks, including task descriptions and evaluatuon metrics, is provided in Table 12 for text-based tasks and Table 13 for speech-based tasks. Further details are provided in Appendix B.

6.1. Text models evaluation

Named Entity Recognition (NER) We perform NER on PxCorpus (Kocabiyikoglu et al., 2022), a dataset containing 1 981 French medical prescriptions in the form of speech recordings along with their transcriptions. Each word of the prescription is tagged in the IOB (Inside-Outside-Beginning) format in various chunks such as “drug name”, “frequency”, “dosage”, *etc*. In the case of the text modality, the task involves classifying tokens, aligning them with the corresponding words, and assigning an IOB label to each word by taking the label of the first token of the word. The metric used for evaluating the models is the F1-score from the SeqEval library (Nakayama, 2018) which is designed for IOB-tagging evaluation.

Automatic Coreference Resolution (CR) The task consists of identifying all mentions (*i.e.*, phrases referring to entities) in the text and grouping those that co-refer to the same entity into clusters. Due to the inherent complexity of coreference resolution, recent systems adopt diverse approaches, often tailored to specific datasets or domains. For benchmarking, we selected the WL-Coref model (Dobrovolskii, 2021; D'Oosterlinck et al., 2023; Liu et al., 2024), known for its strong performance and computational efficiency during training and evaluation. All experiments were implemented using the Stanza library (Qi et al., 2020). For evaluation, we used ANCOR, a corpus comprising manually transcribed interviews (Muzerelle et al., 2014). System performance was assessed using the official `corefud-scorer` (Novák et al., 2024), with head mention matching as the evaluation criterion. We report the CoNLL F1 score.

Extractive Question Answering (QA) The task consists in finding the shortest span of words in a context answering a given question. We used PIAF1.2 (Keraron et al., 2020), an extractive question answering dataset for French inspired by SQuAD1.1 (Rajpurkar et al., 2016) which contains 9 224 context/question/answer pairs extracted from French Wikipedia.

The metric used for evaluation is a F1-score on words, where words are categorized in True/False Positives/Negatives depending on their inclusion in the predicted and ground-truth word spans. This F1-score is at the word-level and not the token-level as different models use different tokenizers that segment the same text into varying token counts. As a result using a token-level F1-score would lead to different evaluations for different models.

FLUE benchmark We evaluate the text models on part of the tasks included in the FLUE benchmark (Le et al., 2020). The tasks are as follows: text classification, paraphrase identification, natural language inference (NLI), verb sense disambiguation (VSD), and dependency parsing. The overall methodology and downstream model architecture are based on the FLUE framework described by Le et al. (2020). For all the above tasks, we reuse the same datasets as in FLUE, except for dependency parsing, for which we use a set of treebanks from the Universal Dependencies repository (de Marneffe et al., 2021): Sequoia (Candito et al., 2014), GSD (McDonald et al., 2013), as well as two treebanks for spoken French: Rhapsodie (Lacheret et al., 2014) and ParisStories (Kahane et al., 2021). We use the HOPS parser (Grobol and Crabbé, 2021) and report classic evaluation metrics for the task: Labelled Attachment Score (LAS) and part-of-speech accuracy (POS).

Jargon biomedical benchmark The Jargon biomedical benchmark (Segonne et al., 2024, Table 7) covers three types of downstream tasks: five token-level classification tasks, one sequence classification task, and one semantic textual similarity task. The token-level tasks draw on the CAS (Grabar et al., 2018) and ESSAI (Dalloux et al., 2021) corpora for POS tagging and UMLS semantic group prediction (CAS-POS, ESSAI-POS, CAS-SG), the QUAERO FrenchMed corpus (Névéol et al., 2014) for NER on MEDLINE titles and EMEA drug descriptions, and the E3C corpus (Minard et al., 2021; Magnini et al., 2020) for 3-class BIO entity recognition (E3C-NER), using Layers 1 and 2 for evaluation and fine-tuning respectively. We report macro-averaged F_1 (**MacF**$_1$ ↑) for POS tag-

Model	FLUE Class. (F1↑)	Paraph. (F1↑)	NLI (Acc↑)	VSD (F1↑)	MEDIA SLU (CER↓)	CoNLL CR (F1↑)	PxCorpus NER (F1↑)	PIAF QA (F1↑)
CamemBERT-B-Wk	86.8±1.2	90.7±0.3	76.88±0.3	49.44	10.5±0.6	69.2±0.8	92.0±0.2	47.1±0.3
CamemBERT-B-Osc	93.7±0.4	91.2±0.5	**82.09**±0.6	**50.03**	11.8±1.9	**73.3**±0.8	**93.6**±0.2	**53.1**±0.6
FlauBERT-B	93.3±0.8	89.0±0.1	80.60±N/A	43.93	10.2±0.6	69.8±0.8	87.3±1.6	N/A
Pantagruel-B-camtok-Wk	86.9±1.1	89.0±0.5	77.43±0.6	31.83	12.8±2.5	67.9±0.5	86.5±0.8	43.5±0.8
Pantagruel-B-Wk	87.3±0.6	89.6±0.8	77.90±0.5	29.39	10.6±0.6	65.4±0.5	87.0±1.7	45.8±0.4
Pantagruel-B-Wk-MLM	88.9±0.7	90.0±0.5	78.41±0.5	43.28	10.5±0.5	67.8±0.8	84.7±2.5	47.3±0.7
Pantagruel-B-Osc-MLM	**94.0**±0.3	90.9±0.3	81.50±0.5	47.90	10.5±0.6	72.6±0.7	89.7±0.4	51.9±0.5
Pantagruel-B-Crs-MLM	93.1±0.3	**91.6**±0.4	81.10±0.3	42.30	**10.1**±0.6	72.1±0.7	86.9±1.5	52.9±0.6

Table 5: Text results. Evaluation tasks and metrics: Class. = text classification (F1); Paraph. = paraphrase identification (F1); VSD = Verb Sense Disambiguation (F1); SLU = Spoken Language Understanding (CER - Concept Error Rate); CR = CoNLL (F1); NER = seqeval F1 score evaluated on the PxCorpus dataset; and QA = word-level F1 for extractive question answering evaluated on the PIAF1.2 dataset. FlauBERT was not tested for QA because its tokenizer implementation does not permit word-token alignment.

Model	Sequoia POS/LAS	GSD POS/LAS	Rhapsodie POS/LAS	ParisStories POS/LAS
CamemBERT-B-Wk	99.1±0.1 / 95.0±0.1	98.4±0.1 / 95.0±0.1	97.5±0.1 / 84.6±0.5	97.1±0.1 / 79.2±0.3
CamemBERT-B-Osc	99.0±0.1 / 95.6±0.1	98.4±0.1 / 95.7±0.1	97.3±0.2 / 86.2±0.2	96.8±0.2 / 80.1±0.3
FlauBERT-B	**99.4**±0.0 / **96.0**±0.0	**98.7**±0.1 / **96.0**±0.1	**97.9**±0.1 / **86.7**±0.3	**97.4**±0.2 / **80.9**±0.3
Pantagruel-B-camtok-Wk	99.0±0.1 / 94.3±0.5	98.3±0.1 / 94.3±0.2	97.1±0.5 / 82.9±1.1	96.9±0.1 / 78.5±0.3
Pantagruel-B-Wk	98.5±0.0 / 93.0±0.3	98.0±0.1 / 93.5±0.2	95.8±0.4 / 80.5±0.5	96.3±0.2 / 77.8±0.4
Pantagruel-B-Wk-MLM	99.0±0.1 / 94.4±0.3	98.3±0.1 / 94.8±0.2	97.4±0.2 / 83.8±0.7	96.9±0.1 / 78.5±0.6
Pantagruel-B-Osc-MLM	99.3±0.1 / 95.2±0.4	98.6±0.1 / 95.3±0.2	97.8±0.1 / 85.9±0.3	97.2±0.1 / 80.2±0.3
Pantagruel-B-Crs-MLM	99.2±0.1 / 95.5±0.1	98.5±0.1 / 95.4±0.1	97.7±0.1 / 85.6±0.4	97.1±0.1 / 79.8±0.3

Table 6: Text results: dependency parsing (↑).

ging and weighted F_1 (**WF**$_1$ ↑) for the remaining tasks. FrenchMedMCQA (Labrak et al., 2022) is a multiple-choice QA dataset of 3,105 questions from French medical exams with five answer options each. CLISTER (Hiebel et al., 2022), derived from CAS, comprises 1,000 annotated sentence pairs rated 0–5 for semantic similarity in clinical text, evaluated with Spearman's ρ.

Summary results on text tasks The results on the text tasks (Tables 5–7) show that Pantagruel-B-Wk-MLM and Pantagruel-B-Osc-MLM perform comparably to CamemBERT-B-Wk and FlauBERT-B, while the larger Pantagruel-B-Crs-MLM benefits from its extensive pre-training data. On the FLUE benchmark (Table 5), Pantagruel models lag slightly behind CamemBERT on NLI but match performance on sentiment analysis and Paraphrase tasks, with Pantagruel-B-Osc-MLM and Pantagruel-B-Crs-MLM achieving the best overall balance across tasks, respectively.

Weaker results on VSD and dependency parsing (Table 6) suggest that the feature-space objective of data2vec may yield less effective token-level representations for text, even when using the compound loss integrating the MLM objective function. Conversely, semantic and retrieval-oriented tasks such as classification and extractive QA benefit from large-scale pre-training with Pantagruel-B-Crs-MLM achieving similar results to CamemBERT-B-Osc on the PIAF 1.2 dataset. On the Jargon biomedical benchmark (Table 7), Pantagruel models match or outperform the baselines, except on the **CLISTER** dataset where the best result has still a significant gap behind the CamemBERT-B-Osc baseline.

Overall, models results and analyses suggest that further improvements can be achieved with a better fine-tuning of the balance between data2vec and MLM loss components, and by employing larger datasets such as OSCAR and CroissantLLM in combination with our best training settings. Investigating more effective ways to combine embedding-level and input-level objectives for text models is left for future work.

6.2. Speech models evaluation

Automatic Speech Recognition (ASR) We assess SSL models on ASR in two settings. In the high-resource setting, models are trained on five audiovisual corpora totalling 258 hours:

Dataset→ Metric→	ESSAI-POS MacF_1	CAS-POS MacF_1	MEDLINE WF_1	EMEA WF_1	CAS-SG WF_1	E3C-NER WF_1	FrMMCQA Hamming	CLISTER ρ
CamemBERT-B-Osc	96.2±0.4	96.6±0.5	85.1±0.2	93.7±0.2	73.4±2.5	93.6±0.8	34.1±0.7	**87.6**±0.0
CamemBERT-B-Wk	96.3±0.2	96.5±0.1	84.3±0.3	93.5±0.1	73.9±0.7	94.0±0.2	32.8±1.0	84.6±0.0
FlauBERT-B	66.8±0.4	89.1±0.2	81.4±1.0	79.5±1.5	66.4±0.7	**94.1**±0.1	35.1±0.8	83.1±0.0
Pantagruel-B-camtok-Wk	95.7±0.2	95.9±0.2	82.3±0.3	91.8±0.4	71.3±1.1	93.0±0.1	32.7±0.9	72.2±0.0
Pantagruel-B-Wk	92.8±0.9	89.6±1.4	77.2±0.2	89.9±0.2	68.3±2.3	92.0±0.2	**35.4**±2.0	75.8±0.0
Pantagruel-B-Wk-MLM	96.3±0.1	96.8±0.6	84.2±0.4	93.7±0.1	75.7±0.4	93.6±0.3	33.7±0.9	76.9±0.0
Pantagruel-B-Osc-MLM	96.3±0.1	97.0±0.1	85.5±0.3	**94.2**±0.1	**76.8**±0.7	93.5±0.1	35.1±1.1	84.1±0.0
Pantagruel-B-Crs-MLM	**96.5**±0.1	**96.9**±0.1	**85.9**±0.1	94.0±0.2	76.6±0.4	93.5±0.1	35.0±0.7	85.5±0.0

Table 7: Text results. Evaluation results for the Jargon biomedical tasks (↑)

Model	PxCorpus NER (F1↑)	MEDIA SLU (CER↓)	ETAPE NER (NEER↓)	AlloSat SER (CCC↑)	fr-en ST (BLEU↑)	fr-es ST (BLEU↑)	fr-pt ST (BLEU↑)
LeBenchmark-w2v-B-1k	57.9±3.0	17.9±2.0	65.13±0.2		14.0±0.5	13.2±0.4	8.60±0.3
LeBenchmark-w2v-L-7k	59.9±3.3	13.3±0.6	52.98±0.1	0.65±0.02	18.2±0.5	18.2±0.7	13.4±0.4
LeBenchmark-w2v-L-14k	82.4±0.2	12.6±0.6	55.30±0.1	0.49±0.01	23.1±0.6	24.2±0.6	21.8±0.6
Pantagruel-B-1k	81.7±1.3	14.4±0.7	61.35±0.3	**0.84**±0.02	17.5±0.4	19.0±0.4	16.8±0.3
Pantagruel-L-14k	**84.4**±0.4	**12.2**±0.7	48.14±0.2	0.82±0.02	24.0±0.4	**25.5**±0.4	21.9±0.4
Pantagruel-L-114k	**84.4**±0.6	12.3±0.6	**44.68**±0.4	0.83±0.03	**25.2**±0.4	25.4±0.4	**24.5**±0.5

Table 8: Speech results. Metrics: CER = Concept Error Rate; NEER = Named Entity Error Rate; CCC = Concordance Correlation Coefficient

Antract (Carrive et al., 2021), QUAERO (Boudahmane et al., 2011), EPAC (Estève et al., 2010), ESTER1 (Gravier et al., 2004), and REPERE (Giraudel et al., 2012). In the dataset-specific setting, we fine-tune separate models on French CommonVoice version 6.1 (Evain et al., 2021b; Parcollet et al., 2024) and ETAPE (Gravier et al., 2012) (low-resource audio-visual). We also evaluate adaptation to child speech using data from the DyLNet project (Nardy et al., 2021), comprising 31 hours of fine-tuning data and 4.09/4.16 hours validation/test sets of conversational speech from children aged 3–6. ASR results in Word Error Rate (WER ↓) are reported in Table 9.

Named Entity Recognition (NER) We benchmark our models on the French ETAPE corpus (Gravier et al., 2012), a 30-hour collection of TV and radio broadcasts covering diverse topics and speaking styles, with an emphasis on spontaneous and multi-speaker speech. ETAPE is a standard benchmark for French NER (Mdhaffar et al., 2022). We train end-to-end NER systems combining an SSL encoder with a three-layer linear probe. The PxCorpus dataset (Kocabiyikoglu et al., 2022) is also used for speech NER, with labels assigned directly to words or word groups, without using the IOB scheme.

Speech Emotion Recognition (SER) In this experiment, we focus on SER continuous prediction of affective dimensions (*e.g.*, satisfaction, valence, arousal). We use AlloSat (Macary et al., 2020), a 37-hour corpus of 303 spontaneous French telephone conversations annotated every 250 ms on a frustration → satisfaction axis.

Spoken Language Understanding (SLU) The French MEDIA corpus (Bonneau-Maynard et al., 2006) contains 1 250 human–machine dialogues about hotel reservations in France, annotated with 76 semantic concepts. It has been widely used to benchmark French SLU systems in both pipeline (Quarteroni et al., 2009; Dinarelli et al., 2009c; Dinarelli and Rosset, 2011; Dinarelli et al., 2017; Caubrière et al., 2020; Ghannay et al., 2021) and end-to-end (Dinarelli et al., 2009a,b, 2010; Serdyuk et al., 2018; Caubrière et al., 2019; Pelloin et al., 2021; Evain et al., 2021b; Parcollet et al., 2024) settings. In particular our results are directly comparable to (Evain et al., 2021b; Parcollet et al., 2024). We perform concept extraction from both speech and transcriptions using Transformer-based SLU models with SSL encoders, reporting Concept Error Rate (CER ↓) in Tables 5 and 8.

Speech-to-text translation (ST) The ST task consists in translating speech in a source language into text in another language. We evaluate our speech models on the French-source subsets of the multilingual TEDx corpus (Salesky et al., 2021), covering three translation directions from French (fr) to English (en), Portuguese (pt), and Spanish (es), with training sets of 50 hours, 38 hours, and 25 hours, respectively. Our experiments are performed in the end-to-end finetuning scenario, where

Model	High-resource setting				Dataset-specific settings		
	Antract	QUAERO	ESTER1	REPERE	CV 6.1	ETAPE	DyLNet
LeBenchmark-w2v-B-1k	22.3 ±3.7	32.0 ±3.5	26.7 ±3.4	28.3 ±3.4	12.28±0.1	34.76 ±0.1	55.74±3.7
LeBenchmark-w2v-L-7k	9.3 ±0.1	14.7 ±0.2	11.4 ±0.1	13.5 ±0.1	9.2±0.1	22.83 ±0.1	41.89±0.1
LeBenchmark-w2v-L-14k	8.6 ±0.1	15.0 ±0.3	11.5 ±0.2	13.6 ±0.2	9.0±0.1	26.03 ±0.1	41.23±0.2
Pantagruel-B-1k	13.2 ±0.3	21.3 ±0.4	17.9 ±0.4	19.1 ±0.3	10.5±0.1	30.61 ±0.3	49.71±0.3
Pantagruel-L-14k	7.5 ±0.1	13.7 ±0.7	9.9 ±0.2	11.3 ±0.2	**8.1**±0.1	19.77 ±0.1	39.34±0.2
Pantagruel-L-114k	**7.4** ±0.1	**11.7** ±0.1	**9.7** ±0.1	**10.4** ±0.1	8.2±0.1	**19.09** ±0.2	**37.48**±0.2

Table 9: Speech results for the ASR tasks in Word Error Rate (WER ↓) comparing performance on the high-resource setting (columns 2-5) and dataset specific ones (columns 6-8).

we plug in a 6-layer Transformer decoder to our pre-trained speech SSL models and finetuned the system on each pair. We followed the same settings for ST fine-tuning in Lebenchmark 2.0 paper (Parcollet et al., 2024) to enable fair comparison with the French wav2vec 2.0-based counterparts. The translation performance in BLEU score (Papineni et al., 2002; Post, 2018) is shown in Table 8.

Summary of results on speech tasks Across speech tasks (Tables 8 and 9), Pantagruel models consistently outperform LeBenchmark baselines. The large model Pantagruel-L-114k achieves the best overall results, with clear gains on challenging spontaneous or noisy corpora such as ETAPE and DyLNet, while maintaining strong performance on CommonVoice. For the cross-modal cross-lingual ST task, the Pantagruel-B-1k model outperforms the LeBenchmark-w2v-B-1k model trained using the same pre-training data by a large margin (+5.8 BLEU on average), highlighting the effectiveness of the latent-based objective for speech inputs under medium-resource regimes. Trained on both LeBenchmark and the INA-100k broadcast corpus, it appears particularly robust to acoustic variability.

Regarding task-specific trends, ASR and SLU benefits most from the larger and more diverse pre-training data, with Pantagruel-L-114k outperforming all other models and Pantagruel-L-14k remaining close behind. In SER, Pantagruel encoders again excel, with Pantagruel-B-1k and Pantagruel-L-114k leading, whereas LeBenchmark-w2v-L-14k struggles to generalize to telephone and spontaneous speech. In the low-resource ST setting (25 hours of fine-tuning data for fr–pt), Pantagruel-L-114k improves translation performance over Pantagruel-L-14k by +2.6 BLEU, demonstrating the benefits of larger pre-training datasets. Overall, the Pantagruel family of models demonstrates solid robustness across all speech benchmarks, confirming the benefit of large-scale feature-space pre-training and the diversity brought by the INA-100k corpus.

7. Discussion and Conclusion

Across modalities, Pantagruel models exhibit complementary behaviors. On speech tasks, Pantagruel consistently outperform LeBenchmark baselines, confirming that feature-space prediction is particularly effective for continuous acoustic signals. The inclusion of INA-100k, with its large diversity of broadcast and spontaneous conditions, enhances robustness to noise and variability, yielding strong gains on ETAPE and DyLNet, although it slightly reduces performance on cleaner, read-speech data such as CommonVoice. On text tasks, the hybrid MLM+data2vec objective mitigates some of the weaknesses of purely embedding-based encoders, achieving competitive results on semantic tasks (QA, classification) but lower scores on syntax-sensitive evaluations (VSD, dependency parsing). This suggests that token-level supervision remains valuable for capturing fine-grained linguistic structure.

Overall, Pantagruel models demonstrate that feature-space self-supervision scales efficiently to French. Its unified architecture, trained separately but identically for text and speech, offers a practical foundation for cross-modal modeling. To our knowledge, this is the most resource-intensive purely French SSL encoders to date, pairing tens of billions of French tokens with 100k hours of diverse French audio.

Future work will focus on (i) joint optimization on unaligned or weakly aligned speech–text pairs, i.e. multi-modal training. We note that our system has already been modified with respect to data2vec 2.0 and it can be trained on audio and text jointly in multi-modal setting, both aligned and unaligned data. Experiments are in progress; (ii) scaling model capacity and corpus diversity; (iii) extending evaluation to multimodal downstream tasks such as spoken QA or automatic subtitling. Beyond empirical performance, the release of Pantagruel aims to empower the French research community with transparent, reproducible, and ethically curated resources, paving the way for responsible multimodal AI in French.

8. Acknowledgements

This research has been partially funded by the French National Research Agency (ANR), project "PANTAGRUEL", ANR-23-IAS1-0001. This work was also supported by the CREMA project (Coreference REsolution into MAchine translation) funded by ANR, contract number ANR-21-CE23-0021-01. It also received government funding managed by ANR under France 2030, reference ANR-23-IACL-0006.Implementation It was also supported by ANR through the MIAI "AI & Language" chair (ANR-19-P3IA-0003) and the MIAI "Socialization and Language at School" chair (ANR-23-IACL-0006). This work was performed using HPC resources from GENCI at IDRIS and CINES under the allocations 2022-A0131013801, 2023-A0151013801, 2024-A0171013801, 2024-A0161015074, and 2025-A0191013801 on the Jean Zay and Adastra supercomputers.

Bibliographical References

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. Camembert 2.0: A smarter french language model aged to perfection. *CoRR*, abs/2411.08868.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023. Data-efficient french language modeling with camemberta. In *ACL (Findings)*, pages 5174–5185. Association for Computational Linguistics.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629. IEEE.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, et al. 2025. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *CoRR*, abs/2506.09985.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *INTERSPEECH*, pages 2278–2282. ISCA.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 1416–1429. PMLR.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Randall Balestriero and Yann LeCun. 2025. Lejepa: Provable and scalable self-supervised learning without the heuristics. *CoRR*, abs/2511.08544.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *Trans. Mach. Learn. Res.*, 2024.

H. Bonneau-Maynard, C. Ayache, and et al. 2006. Results of the French evalda-media evaluation campaign for literal understanding. In *In LREC*, Genoa, Italy. European Language Resources Association (ELRA).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2022. On the usability of transformers-based models for a french question-answering task - abstract. In *CIRCLE*, volume 3178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Antoine Caubrière, Sahar Ghannay, Natalia Tomashenko, Renato de Mori, Antoine Laurent, Emmanuel Morin, and Yannick Estève. 2020. Error analysis applied to end-to end spoken language understanding. In *45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, pages 8514–8518, Barcelona, Spain.

Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève. 2019. Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *20th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1198–1202, Graz, Austria.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Jean-Hugues Chenot and Gilles Daigneault. 2014. A Large-Scale Audio and Video Fingerprints-Generated Database of TV Repeated Contents. In *12th International Workshop on Content-Based Multimedia Indexing (CBMI2014)*, pages 1–6, Klagenfurt, Austria.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009a. Concept segmentation and labeling for conversational speech. In *Interspeech*, Brighton, U.K.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009b. Re-ranking models based-on small training data for spoken language understanding. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1076–1085, Singapore. Association for Computational Linguistics.

Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009c. Re-ranking models for spoken language understanding. In *Conference of the European Chapter of the Association of Computational Linguistics*, pages 202–210, Athens,Greece.

Marco Dinarelli and Sophie Rosset. 2011. Hypotheses selection criteria in a reranking framework for spoken language understanding. In *Conference of Empirical Methods for Natural Language Processing*, pages 1104–1115, Edinburgh, U.K.

Marco Dinarelli, Evgeny Stepanov, Sebastian Varges, and Giuseppe Riccardi. 2010. The luna spoken dialog system: Beyond utterance classification. In *International Conference on Acoustic, Speech and Signal Processing*, Dallas, Texeas, U.S.A.

Marco Dinarelli, Vedran Vukotic, and Christian Raymond. 2017. Label-dependency coding in Simple Recurrent Networks for Spoken Language Understanding. In *Interspeech*, Stockholm, Sweden.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. 2018. Adapting auxiliary losses using gradient similarity. *CoRR*, abs/1812.02224.

Moussa Kamal Eddine, Antoine J.-P. Tixier, and Michalis Vazirgiannis. 2021. Barthez: a skilled pretrained french sequence-to-sequence model. In *EMNLP (1)*, pages 9369–9390. Association for Computational Linguistics.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, et al. 2021a. LeBenchmark: A Reproducible

Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021: Conference of the International Speech Communication Association*, Brno, Czech Republic.

Solène Evain, Manh Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, et al. 2021b. Task agnostic and task specific self-supervised learning from speech with LeBenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. 2023. A-JEPA: joint-embedding predictive architecture can listen. *CoRR*, abs/2311.15830.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *NAACL-HLT (1)*, pages 240–250. Association for Computational Linguistics.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From freem to d'alembert: a large corpus and a language model for early modern french. In *LREC*, pages 3367–3374. European Language Resources Association.

Sahar Ghannay, Antoine Caubrière, Salima Mdhaffar, Gaëlle Laperrière, Bassam Jabaian, and Yannick Estève. 2021. Where are we in semantic concept extraction for Spoken Language Understanding? ⋆. In *SPECOM 2021 23rd International Conference on Speech and Computer*, Saint Petersburg, Russia.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. 2012. The etape corpus for the evaluation of speech-based tv content processing in the french language. In *8th International Conference on Language Resources and Evaluation (LREC 2012)*, page na.

Loïc Grobol and Benoît Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés. In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 16:1569–1583.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Hai Huang, Yann LeCun, and Randall Balestriero. 2025. Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a native French question-answering dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. *arXiv preprint arXiv:2207.08292*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference of Learning Representations*.

Gaëlle Laperrière, Valentin Pelloin, Antoine Caubrière, Salima Mdhaffar, Nathalie Camelin, Sahar Ghannay, Bassam Jabaian, and Yannick Estève. 2022. The spoken language understanding MEDIA benchmark dataset in the era of deep learning: data updates, training and evaluation tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1595–1602, Marseille, France. European Language Resources Association.

Julien Launay, E. L. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. Pagnol: An extra-large french generative model. In *LREC*, pages 4275–4284. European Language Resources Association.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for french.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and conjunction-aware word-level coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *ICLR (Poster)*. OpenReview.net.

Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. 2020. Allosat: A new call center french corpus for satisfaction and frustration analysis. In *Language Resources and Evaluation Conference, LREC 2020*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Salima Mdhaffar, Jarod Duret, Titouan Parcollet, and Yannick Estève. 2022. End-to-end model for named entity recognition from speech without paired training data. In *Proc. Interspeech 2022*, pages 4068–4072.

Salima Mdhaffar, Haroun Elleuch, Chaimae Chellaf, Ha Nguyen, and Yannick Estève. 2025. Sense models: an open source solution for multilingual and multimodal semantic-based tasks.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *EMNLP (1)*, pages 7853–7858. Association for Computational Linguistics.

Shentong Mo and Peter Tong. 2024. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. In *NeurIPS*.

Shentong Mo and Sukmin Yun. 2024. DMT-JEPA: discriminative masked targets for joint-embedding predictive architecture. *CoRR*, abs/2405.17995.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive french language model. *CoRR*, abs/2202.03371.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Aurélie Nardy, Hélène Bouchet, Isabelle Rousset, Loïc Liégeois, Laurence Buson, Céline Dugua, and Jean-Pierre Chevrot. 2021. Variation sociolinguistique et réseau social: constitution et traitement d'un corpus de données orales massives. *Corpus*, 22.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, et al. 2024. LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of French speech. *Computer Speech and Language*, 86:101622.

Valentin Pelloin, Nathalie Camelin, Antoine Laurent, Renato de Mori, Antoine Caubrière, Yannick Estève, and Sylvain Meignier. 2021. End2End Acoustic to Semantic Transduction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada.

Valentin Pelloin, Franck Dary, Nicolas Hervé, Benoit Favre, Nathalie Camelin, Antoine Laurent, and Laurent Besacier. 2022. Asr-generated text for language model pre-training applied to speech tasks. In *Interspeech 2022*, pages 3453–3457.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*, pages 186–191. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1118–1127. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

S. Quarteroni, G. Riccardi, and M. Dinarelli. 2009. What's in an ontology for spoken language understanding. In *Interspeech*, Brighton, U.K.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI Technical Report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Vincent Segonne, Marie Candito, and Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 259–270, Gothenburg, Sweden. Association for Computational Linguistics.

Vincent Segonne, Aidan Mannion, Laura Cristina Alonzo Canul, Alexandre Daniel Audibert, Xingyu Liu, Cécile Macaire, et al. 2024. Jargon: A suite of language models and evaluation tasks for French specialized domains. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9463–9476, Torino, Italia. ELRA and ICCL.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5754–5758. IEEE Press.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Trans. Assoc. Comput. Linguistics*, 12:58–79.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. 2025. Audio-jepa: Joint-embedding predictive architecture for audio representation learning. *CoRR*, abs/2507.02915.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. 2025. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, et al. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *ACL (1)*, pages 2526–2547. Association for Computational Linguistics.

Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. 2025. Wavjepa: Semantic learning unlocks robust audio foundation models for raw waveforms. *CoRR*, abs/2509.23238.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling.

Language Resource References

2003. African Accented French. Type: dataset.

2019. Mpf. Https://hdl.handle.net/11403/mpf/v3, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

ATILF. 2020. TCOF : Traitement de corpus oraux en français. Https://hdl.handle.net/11403/tcof/v2.1, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estève. 2022. Speech resources in the Tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2066–2071, Marseille, France. European Language Resources Association.

Karim Boudahmane, Bianka Buschbeck, Eunah Cho, Josep Maria Crego, Markus Freitag, Thomas Lavergne, Hermann Ney, Jan Niehues, Stephan Peitz, Jean Senellart, Artem Sokolov, Alex Waibel, Tonio Wandmacher, Joern Wuebker, and François Yvon. 2011. Advances on spoken language translation in the quaero program. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 114–120, San Francisco, California.

S. Branca-Rosoff, S. Fleury, F. Lefeuvre, and M. Pires. 2012. Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000). Http://cfpp2000.univ-paris3.fr/CFPP2000.pdf.

Tanja Bänziger et al. 2012. Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric de la Clergerie. 2014. Deep syntax annotation of the sequoia French treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2298–2305, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jean Carrive, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Antoine Laurent, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Bénédicte Pincemin, Géraldine Poels, and Raphaël Troncy. 2021. Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quarterly*, 15(1). Editors: Taylor Arnold, Jasmijn van Gorp, Stefania Scagliola, and Lauren Tilton.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Iris Eshkol-Taravella et al. 2011. Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Ressources Linguistiques Libres - TAL*.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas.

2010. The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, Jaoe Alves, Ricardo Rei, Pedro Raphaël Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model.

Soline Felice, Solène Evain, Solange Rossato, and François Portet. 2024. Audiocite. net: A large spoken read dataset in french. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Gadet Françoise. 2017. Les parlers jeunes dans l'île-de-France multiculturelle. *Paris and Gap, Ophrys*.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech 2009*, pages 2583–2586.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The REPERE corpus : a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1102–1107, Istanbul, Turkey. European Language Resources Association (ELRA).

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A Canadian French emotional speech dataset. In *MMSys*.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. Cas: French corpus with clinical cases. In *LOUHI 2018-The Ninth International Workshop on Health Text Mining and Information Analysis*, pages 1–7.

G. Gravier, J-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. 2004. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Guillaume Gravier, Gilles Adda, Niklas Paulsson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 114–118, Istanbul, Turkey. European Language Resources Association (ELRA).

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2022. Clister: A corpus for semantic textual similarity in french clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4306–4315.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Beatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 295–301, Reykjavik, Iceland. European Language Resources Association (ELRA).

Clément Le Moine and Nicolas Obin. 2020. Att-HACK: An Expressive Speech Database with Social Attitudes.

Fabrice Lefèvre et al. 2012. Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : le projet PortMedia. In *JEP-TALN-RECITAL*.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao

Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Anne-Lyse Minard, Roberto Zanoli, Begoña Altuna, Manuela Speranza, Bernardo Magnini, and Alberto Lavelli. 2021. European clinical case corpus. Bruno Kessler Foundation.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, et al. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *INTERSPEECH*, Shanghai, China.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. In *Interspeech*, pages 3655–3659. ISCA.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. 2010. The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3):201.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online.

A. Pretraining Hyperparameters

Pre-training hyperparameters for the speech and text models are reported in Tables 10 and 11, respectively. All models are trained with the Adam optimizer (Kingma and Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.98$), using a cosine learning rate scheduler and a weight decay of 0.1.

For text models, the maximum sequence length is set to 512 tokens. We applied the the *complete* break mode in the fairseq library, which splits samples only at sentence boundaries but may include multiple sentences per sample. Layer drop is disabled for all text models.

B. Details on Benchmark experiments

This section presents text and speech downstream tasks used for evaluation. Unless otherwise noted, all results are averaged over five runs with different random seeds. We use ↑ (respectively ↓) to indicate that higher (respectively lower) values correspond to better performance.

B.1. Text models evaluation

Named Entity Recognition (NER) We perform NER on PxCorpus (Kocabiyikoglu et al., 2022), a dataset containing 1 981 French medical prescriptions in the form of speech recordings along with their transcriptions. Each word of the prescription is tagged in the IOB (Inside-Outside-Beginning) format in various chunks such as “drug name”, “frequency”, “dosage”, *etc*. In the case of the text modality, the task involves classifying tokens, aligning them with the corresponding words, and assigning an IOB label to each word by taking the label of the first token of the word.

A linear layer followed by a softmax are added on top of the encoder to predict one of the 37 different classes for each token. The encoder and the linear layer are fine-tuned together during training. The loss function used is the cross-entropy loss weighted to account for class imbalance. The metric used for evaluating the models is the F1-score from the SeqEval library (Nakayama, 2018) which is designed for IOB-tagging evaluation.

Automatic Coreference Resolution (CR) The task consists of identifying all mentions (*i.e.*, phrases referring to entities) in the text and grouping those that co-refer to the same entity into clusters. Due to the inherent complexity of coreference resolution, recent systems adopt diverse approaches, often tailored to specific datasets or domains. For benchmarking, we selected the WL-Coref model (Dobrovolskii, 2021; D’Oosterlinck et al., 2023; Liu et al., 2024), known for its strong

	B-1k / B-14k	L-14k	L-114K
Number GPUs	16	48	128
GPU type	H100	MI250x	H100
Batch size (seconds/GPU)	62.5	40	62.5
Learning rate	7.5×10^{-4}	4.0×10^{-4}	2.0×10^{-4}
N_{warmup} steps	8,000	5,000	20,000
N_{max}	400,000	300,000	1,000,000
Clip norm	-	1.0	1.0
Layerdrop	0.05	0.0	0.0
Multi-masks	8	12	12
EMA start	0.999	0.9997	0.9997
EMA end	0.99999	1.0	1.0
EMA anneal steps	75,000	300,000	600,000
Mask length	5	5	5
Mask ratio	0.5	0.55	0.55
Mask adjust	0.05	0.1	0.1
Top K target layers	8	16	16
Decoder layers	4	4	4
Decoder dimension	384	768	768
Decoder CNN groups	16	16	16
Decoder kernel	7	7	7

Table 10: Pre-training hyper-parameters for speech models.

	B-camtok-Wk / B-Wk	B-Wk-MLM	B-Osc-MLM	B-Crs-MLM
Number GPUs	16	16	32	48
Batch size (tokens/GPU)	4	32	28	28
Learning rate	2×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
N_{warmup} steps	4,000	8,000	16,000	16,000
N_{max}	500,000	250,000	400,000	400,000
λ_{start}	-	20.0	20.0	20.0
λ_{end}	-	1.0	2.0	2.0
N_{λ}	-	250,000	400,000	400,000
Clip norm	1.0	1.0	1.0	1.0
Multi-masks	8	8	8	8
EMA start	0.9999	0.9995	0.9995	0.9995
EMA end	0.99999	0.99995	0.99995	0.99995
EMA anneal steps	100,000	125,000	200,000	200,000
Mask length	3	3	3	3
Mask ratio	0.6	0.6	0.6	0.6
Mask adjust	0.0	0.0	0.0	0.0
Top K target layers	12	12	12	12
Decoder layers	5	5	5	5
Decoder dimension	768	768	768	768
Decoder CNN groups	1	1	1	1
Decoder kernel	9	9	9	9

Table 11: Pre-training hyper-parameters for text models. All text models are trained using Nvidia H100 GPU.

performance and computational efficiency during training and evaluation. All experiments were implemented using the Stanza library (Qi et al., 2020).

For evaluation, we used ANCOR, a corpus comprising manually transcribed interviews (Muzerelle et al., 2014). The corpus is freely available as part of CorefUD 1.3, a multilingual collection of coreference corpora (Novák et al., 2025). Since only train and validation splits are available, we used 10% of the train split for model validation during training and the validation split for reporting the performance. System performance was assessed using the official `corefud-scorer` (Novák et al., 2024), with head mention matching as the evaluation criterion. We report the CoNLL F1 score.

Extractive Question Answering (QA) The task consists in finding the shortest span of words in a context answering a given question. We used PIAF1.2 (Keraron et al., 2020), an extractive question answering dataset for French inspired by SQuAD1.1 (Rajpurkar et al., 2016) which contains 9 224 context/question/answer pairs extracted from French Wikipedia.

The context and the question are passed in the encoder separated by a <SEP> special token. A linear layer with two outputs is added on top of the encoder to predict, for each token, the logits that the token starts and ends the answer span. The predicted span is obtained by choosing the tuple $(\mathrm{token1}, \mathrm{token2})$ in the context that maximizes $\mathrm{token1}[\mathrm{start}] + \mathrm{token2}[\mathrm{end}]$ (the sum is used here instead of the product because the linear layer produces logits and not probabilities). At the end, the context span expressed in character numbers and not token numbers is obtained by taking the smallest word span that contains the predicted token span. The encoder and the linear layer are fine-tuned together during the training phase. The cross-entropy loss is used for computing the gradient.

The metric used for evaluation is a F1-score on words, where words are categorized in True/False Positives/Negatives depending on their inclusion in the predicted and ground-truth word spans. This F1-score is at the word-level and not the token-level as different models use different tokenizers that segment the same text into varying token counts. As a result using a token-level F1-score would lead to different evaluations for different models.

FLUE benchmark We evaluate the text models on part of the tasks included in the FLUE benchmark (Le et al., 2020). The tasks are as follows: text classification, paraphrase identification, natural language inference (NLI), verb sense disambiguation, and dependency parsing. The overall methodology and downstream model architecture are based on the FLUE framework described by Le et al. (2020).

For text classification, we use the Cross-Lingual Sentiment (CLS) dataset (Prettenhofer and Stein, 2010), formulated as a binary classification task that aims to determine whether a given review expresses a positive or negative sentiment. For paraphrase identification, we use the Cross-Lingual Adversarial Dataset for Paraphrase Identification (PAWS-X, Zhang et al., 2019), also a binary classification task where the goal is to assess whether two input sentences are semantically equivalent. The reported results include the macro F1 score and their standard deviation across five different seeds.

For natural language inference , we use the XNLI dataset (Conneau et al., 2018), a three-class classification task that determines the logical relationship between a premise and a hypothesis. Following Martin et al. (2020) and Le et al. (2020), we report the mean and standard deviation from 10 runs with different seeds. The output label indicates whether the hypothesis is entailed by, contradicts, or is neutral with respect to the premise.

For verb sense disambiguation, we evaluate on the FrenchSemEval dataset (Segonne et al., 2019), which tests sense disambiguation of French verbs. We report the macro F1 score as the evaluation metric.

For dependency parsing, we use a set of treebanks from the Universal Dependencies repository (de Marneffe et al., 2021): Sequoia (Candito et al., 2014), GSD (McDonald et al., 2013), as well as two treebanks for spoken French: Rhapsodie (Lacheret et al., 2014) and ParisStories (Kahane et al., 2021). We use the HOPS parser (Grobol and Crabbé, 2021) and report classic evaluation metrics for the task: Labelled Attachment Score (LAS) and part-of-speech accuracy (POS).

Jargon biomedical benchmark The Jargon biomedical benchmark (Segonne et al., 2024, Table 7) covers three types of downstream tasks. Five of the seven tasks are token-level classification tasks, which are accompanied by a sequence classification task and a semantic textual similarity task.

The token-level labeling tasks are built on CAS (Grabar et al., 2018) and ESSAI corpus (Dalloux et al., 2021), used for POS tagging and UMLS semantic group prediction (CAS-POS, ESSAI-POS, CAS-SG). The QUAERO FrenchMed corpus (Névéol et al., 2014) provides NER data from MEDLINE titles and EMEA drug descriptions. We also employ the E3C corpus (Minard et al., 2021; Magnini et al., 2020) for a 3-class BIO entity recognition task (E3C-NER), using Layers 1 (manual annotation) and Layer 2 (automatic labeling) for evaluation and fine-tuning respectively. We report macro-averaged F_1 scores for the POS-tagging tasks (**MacF**$_1$ $\uparrow$); for the others, a weighted average across the output classes (**WF**$_1$ $\uparrow$).

FrenchMedMCQA (Labrak et al., 2022) is a multiple-choice QA dataset of 3,105 questions from French medical specialization exams, each with five answer options. Since questions can have multiple correct answers, the Hamming score we use measure the overlap between the predicted answer combination and the ground truth. CLISTER (Hiebel et al., 2022), derived from CAS, comprises 1,000 manually annotated sentence pairs (scores 0–5) and is used to assess semantic similarity in clinical text, a task important for detecting redundant information (Wang et al., 2020). We use Spearman's correlation coefficient ρ for evaluation.

B.2. Speech models evaluation

Automatic Speech Recognition (ASR) We assess SSL models on ASR in two settings: with a high-resource pipeline where we train on multiple corpora, and dataset specific pipelines. For each pipeline, we use a character-based CTC architecture with an optimized number of linear layers and learning rate. Linear layers are placed on top of the SSL encoder, and the whole model is finetuned. Transcriptions are obtained with a greedy decoding and without the use of a language model.

High-resource setting use multiple audiovisuals training datasets, representing 258h of training data: Antract (Carrive et al., 2021), QUAERO (Boudahmane et al., 2011), EPAC (Estève et al., 2010), ESTER1 (Gravier et al., 2004) and REPERE (Giraudel et al., 2012).

Next, we assess SSL models on ASR on a dataset specific settings, where we train specific variations for each dataset with similar architectures. We first benchmark models on the French CommonVoice (CV) dataset. We use the CommonVoice 6.1 version to compare with results from Evain et al. (2021b); Parcollet et al. (2024). Similarly, we use ETAPE (Gravier et al., 2012) dataset for ASR on low-resource audiovisual settings.

Finally, we assessed the ability of ASR models to adapt to spontaneous child speech. For this purpose, we used data from the DyLNet project (Nardy et al., 2021), which ecologically collected conversational speech from children aged 3 to 6 years in a school context. The resulting corpus comprises approximately 35 000 hours of raw audio, of which 800 hours were manually annotated. Among these annotations, 60 hours correspond to speech segments. From this subset, 31 hours were selected for fine-tuning the SSL models. The validation and test sets consist of 4.09 and 4.16 hours of speech, respectively. We present in Table 9 the ASR results in Word Error Rate (WER $\downarrow$) for those pipelines.

Named Entity Recognition (NER) We benchmark our models on the French ETAPE corpus (Gravier et al., 2012), a 30-hour collection of TV and radio broadcasts covering diverse topics and speaking styles, with an emphasis on spontaneous and multi-speaker speech. ETAPE is a standard benchmark for French NER (Mdhaffar et al., 2022). We train end-to-end NER systems combining an SSL encoder with a three-layer linear probe. The PxCorpus dataset (Kocabiyikoglu et al., 2022) is also used for speech NER, with labels assigned directly to words or word groups, without the IOB scheme.

Speech Emotion Recognition (SER) *Speech Emotion Recognition (SER)* comprises (1) discrete-state single-label classification (such as happy, sad, or neutral) and (2) continuous estimation of a given affective dimension (*e.g.* satisfaction, valence, arousal), yielding an intensity time series. We focus on continuous prediction. We use AlloSat (Macary et al., 2020), a corpus of 303 spontaneous French telephone conversations (total 37h), with time-continuous annotations on a frustration$\rightarrow$satisfaction axis (a label every 250 ms). **Features and alignment.** We use frozen self-supervised encoders to extract frame-level embeddings produced approximately every 20 ms, while labels are provided at the annotation rate. To align these rates, we aggregate consecutive frames into 250 ms windows by alternating mean-pooling over 12 and 13 frames (average 12.5 $\approx$ 250 ms), so the feature and label timelines stay synchronized over long sequences. We drop the final partial window, compute the train-set mean and standard deviation per feature dimension, and apply this normalization to all splits. **Model and hyperparameters** The downstream model is a 5-BiLSTM (five stacked bidirectional LSTM layers with hidden sizes $[512, 200, 64, 32, 32]$) followed by a linear layer predicting one scalar per frame. Training minimizes negative CCC between predictions and labels using Adam (lr $= 1 \times 10^{-3}$), batch size 15, for 200 epochs. For development and test, we compute a single concatenated CCC per split over all valid . We then select the checkpoint with the best dev-CCC (best over epochs) and evaluate that checkpoint on the test set.

Results: analysis and discussion for SER Performances vary across LeBenchmark models (3K/7K/14K), while Pantagruel encoders are consistently strong in this frozen regression setup. Pantagruel-B-1k gets the best results on both the validation and test data; Pantagruel-L-114k is close on test. In contrast, LeBenchmark-14K-large is weak on test, indicating limited off-the-shelf generalization to telephone conversations. These findings suggest that, in this frozen-encoder regression setting, more pretraining hours or larger model size do not necessarily yield better SER performance.

Spoken Language Understanding (SLU) We benchmark our models on the SLU task with the French corpus MEDIA (Bonneau-Maynard et al., 2006). This corpus covers the topic of hotel information and reservations in France, and is made up of 1,250 human-machine dialogues transcribed and annotated with 76 semantic concepts.
MEDIA has been used extensively for benchmarking French models for SLU, both with symbolic and neural models, and both in pipeline systems, where an Automatic Speech Recognizer (ASR) feeds a

Natural Language Understanding (NLU) module (Quarteroni et al., 2009; Dinarelli et al., 2017), and *end-to-end* systems (Dinarelli et al., 2009a, 2010). In particular we can compare results presented here with those published in (Evain et al., 2021b; Parcollet et al., 2024). Downstream SLU models are based on the Transformer architecture and use either speech or text SSL encoders. All models are trained with a 3-stage learning rate (lr) scheduler, with 25 epochs of linear lr warm-up to reach the optimal learning rate, 65 epochs during which lr is unchanged, and 60 epochs where lr decays exponentially. The optimal lr for all text models and *Base* speech models is $5e^{-5}$, while we use $2.5e^{-5}$ for *Large* speech models. Like in (Parcollet et al., 2024) we compute the average of the 5 best models based on the loss on the development data and we perform greedy decoding. The evaluation metric is the Concept Error Rate (the lower the better $\downarrow$), and we report the average of 5 runs.

Results are reported in Tables 5 and 8 as *SLU* in terms of Concept Error Rate (CER $\downarrow$).

We note that when performing SLU from speech, and in order to extract both concepts (slots) and values (slot-values), the model predicts words, concept boundaries and concepts at the same time. For instance from a speech segment containing the text segment *"... in Majorca from the 11th to the 15th of May ..."*, the model should predict *"<Begin> in Majorca LOCALIZATION <End> <Begin> from the 11th START-DATE <End> <Begin> to the 15th of May END-DATA <End>"*. This output is then post-processed as *"LOCALIZATION[in Majorca] START-DATE[from the 11th] END-DATE[to the 15th of May]"*. Finally, the value extraction phase is performed, and the final output is *"LOCALIZATION[Majorca] START-DATE[11/05/2026] END-DATE[15/05/2026]"*.

Because of the presence of open-value concepts, such like dates and other amounts, value extraction has been performed for long with rule-based systems (Hahn et al., 2010). However this has been recently questioned since rules appeared to have a high oracle error rate (i.e. error rate over gold reference concepts) (Laperrière et al., 2022). Thus (Mdhaffar et al., 2025) evaluated SLU models with Concept Error Rate (CER in this work), that is error rate over sequences of concepts alone (i.e. *"LOCALIZATION START-DATE END-DATA"* in the example above), and on Concept-Value Error Rate (CVER) where values are actually unnormalized phrases with tokens concatenated to each other (i.e. *"LOCALIZATION[in-Majorca] START-DATE[from-the-11th] END-DATE[to-the-15th-of-May]"*).

While we are able to extract both concepts and concept-values, to keep this work comparable with to (Evain et al., 2021b; Parcollet et al., 2024), and for lack of space in the tables and in the paper, in this work we show only CER. We leave evaluation with CVER, and possibly more sophisticated approaches for normalized value extraction, to future work.

Category	Task	Datasets	Description	Metrics
General	NER	PxCorpus	Medical prescription entity extraction (IOB format).	SeqEval F1 ↑
	Coreference	ANCOR	Identifying and clustering mentions of the same entity.	CoNLL F1 ↑
	Extractive QA	PIAF 1.2	Finding answer spans within French Wikipedia contexts.	Word-level F1 ↑
	SLU	MEDIA	Concept extraction from manually transcribed hotel reservation dialogues.	CER ↓
FLUE	Classification	CLS	Binary sentiment analysis of Amazon reviews.	Macro F1 ↑
	Paraphrase	PAWS-X	Semantic equivalence detection between sentence pairs.	Macro F1 ↑
	NLI	XNLI	Logical relationship classification (Entailment, etc.).	Accuracy ↑
	VSD	FrenchSemEval	Disambiguating French verb meanings in context.	Macro F1 ↑
	Dependency	Sequoia, GSD, Rhapsodie, Paris-Stories	Syntactic dependency parsing and POS tagging.	LAS ↑ / POS ↑
Jargon Biomedical	POS Tagging	ESSAI-POS, CAS-POS	Token-level part-of-speech tagging for medical text.	Macro F1 ↑
	NER	MEDLINE, EMEA	Entity recognition on drug descriptions and titles.	Weighted F1 ↑
	Semantic Group.	CAS-SG	Predicting UMLS semantic groups for medical terms.	Weighted F1 ↑
	BIO Recognition	E3C	3-class entity recognition in clinical case reports.	Weighted F1 ↑
	Medical QA	FrMedMCQA	Multiple-choice French medical examination questions.	Hamming ↑
	Similarity	CLISTER	Semantic similarity (0–5 scale) for clinical sentences.	Spearman's ρ ↑

Table 12: Overview of French General and Biomedical Text Model Evaluation Benchmarks

Task	**Subtask**	**Datasets**	**Description**	**Metrics**
ASR	High-resource	Antract, QUAERO, ESTER1, REPERE	Broad-scale ASR on 258h of audiovisual data.	WER ↓
	Specific	CommonVoice, ETAPE	Benchmarking on crowdsourced and low-resource data.	WER ↓
	Child Speech	DyLNet	Conversational speech recognition for children (ages 3–6).	WER ↓
Speech NER	Broadcast	ETAPE	End-to-end entity recognition on TV/Radio broadcasts.	NEER ↓
	Medical	PxCorpus	End-to-end entity labeling on medical recordings. (direct to word groups, no IOB schema)	F1 ↑
SER	Continuous	AlloSat	Continuous frustration-satisfaction axis prediction.	CCC ↑
SLU	Concept Ext.	MEDIA	End-to-end spoken language understanding for hotel reservation dialogues.	CER ↓
ST	Translation	mTEDx	Multilingual speech-to-text translation (fr → en, es, pt).	BLEU ↑

Table 13: Overview of Speech Model Evaluation Benchmarks

C. Speech datasets

Corpus	License	Duration	Speech type
		1K dataset	
MLS French (Pratap et al., 2020)	CC BY 4.0	**1,096:43** 520:13 / 576:29 / –	Read
		14k dataset	
EPAC** (Estève et al., 2010)	ELRA NC	**1,626:02** 1,240:10 / 385:52 / –	Radio Broadcasts
African Accented French (noa, 2003)	Apache 2.0	**18:56** – / – / 18:56	Read
Att-Hack (Le Moine and Obin, 2020)	CC BY-NC-ND	**27:02** 12:07 / 14:54 / –	Acted Emotional
CaFE (Gournay et al., 2018)	CC NC	**1:09** 0:32 / 0:36 / –	Acted Emotional
CFPP2000* (Branca-Rosoff et al., 2012)	CC BY-NC-SA	**16:26** 0:14 / 1:56 / 14:16	Spontaneous
ESLO2 (Eshkol-Taravella et al., 2011)	CC BY-NC-SA	**34:12** 17:06 / 16:57 / 0:09	Spontaneous
GEMEP (Bänziger et al., 2012)	User agreement	**0:50** 0:24 / 0:26 / –	Acted Emotional
MPF (Françoise, 2017; MPF, 2019)	CC BY-NC-SA 4.0	**19:06** 5:26 / 4:36 / 9:03	Spontaneous
PORTMEDIA (French) (Lefèvre et al., 2012)	ELRA NC	**38:59** 19:08 / 19:50 / –	Acted telephone dialogue
TCOF (Adults) (ATILF, 2020)	CC BY-NC-SA	**53:59** 9:33 / 12:39 / 31:46	Spontaneous
NCCFr (Torreira et al., 2010)	User agreement	**26:35** 12:44 / 12:59 / 00:50	Spontaneous
Voxpopuli (*Unlabeled*) (Wang et al., 2021)	CC0	**4,532:17** – / – / 4,532:17	Professional speech
Voxpopuli (*Transcribed*) (Wang et al., 2021)	CC0	**211:57** – / – / 211:57	Professional speech
Audiocite.net (Felice et al., 2024)	CC BY + ND/NC/SA	**6698:35** 3477:24 / 1309:49 / 1911:21	Read
Niger-Mali Audio collection (Boito et al., 2022)	CC BY-NC-ND	**111:01** 52:15 / 58:46 / –	Radio broadcasts
14K dataset total		**14,529:18** 5,363:01 / 2,409:42 / 6,756:28	-
		114k dataset	
INA-100k**	private	**100,000:00** 52,958:00 / 22,642:00 / –	Radio/TV Broadcasts
114k dataset total		**114,529:18** 58,321:01 / 25,051:42 / 6,756:28	-

Table 14: Statistics for the speech corpora used to train SSL models according to gender information (male / female / unknown). The small dataset is from MLS only. Each dataset is composed of the previous one + additional data; duration: hour(s):minute(s).

*Composed of audio files not included in the CEFC corpus v2.1, 02/2021; **speakers are not uniquely identified.; Stats of CFPP2000, MPF and TCOF have changed a bit due to a change in data extraction; License: CC=Creative Commons; NC=non-commercial; BY= Attribution; SA= Share Alike; ND = No Derivative works; CC0 = No Rights Reserved; User agreement = open for research and NC.