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

Alignement contrastif de séries temporelles et métadonnées via verbalisation par LLM

Alexandre Legrosse[1, 2]
(1) EDF R&D, 91120 Palaiseau, France
(2) Université Lumière Lyon 2, ERIC, 69007, Lyon, France
alexandre.legrosse@edf.fr

RÉSUMÉ

L'alignement entre séries temporelles et métadonnées tabulaires est limité par la pauvreté sémantique des encodages catégoriels, qu'il s'agisse des représentations *one-hot* ou d'autres méthodes d'encodage. Ces approches d'encodages, pauvres sémantiquement, ignorent les dynamiques et les interactions qui donnent pourtant tout leur sens aux données. Nous proposons une architecture d'alignement contrastif pour structurer un espace latent commun, reposant sur un pipeline de verbalisation par LLM qui transforme les métadonnées en descriptions textuelles contextualisées. Cette approche permet une recherche d'information bidirectionnelle entre séries temporelles et métadonnées. Évaluée sur 20 000 courbes de charge électrique via le Semantic Recall@k (une métrique adaptée aux contextes où plusieurs individus partagent des attributs identiques), elle surpasse les baselines, démontrant que l'enrichissement sémantique par verbalisation améliore significativement la structuration de l'espace latent.

ABSTRACT

Contrastive Alignment between Time Series and Metadata via LLM-based Verbalization

The alignment between time series and tabular metadata is hindered by the semantic limitations of categorical encodings—whether one-hot or vector-based—which fail to capture the dynamic and interactive dimensions that give data its full meaning. To address this, we introduce a contrastive alignment architecture that structures a shared latent space using an LLM-based verbalization pipeline, which converts metadata into contextualized textual descriptions. This approach enables bidirectional information retrieval between time series and metadata. Evaluated on 20,000 electrical load curves using Semantic Recall@k—a metric tailored to scenarios where multiple individuals exhibit identical attributes—our method outperforms baseline approaches, demonstrating that semantic enrichment through verbalization significantly enhances the organization of the latent space.

MOTS-CLÉS : Alignement multimodal, séries temporelles, métadonnées tabulaires, apprentissage contrastif, verbalisation.

KEYWORDS: Multimodal alignment, time series, tabular metadata, contrastive learning, verbalization.

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 100–116.

1 Introduction

1.1 Contexte

Les courbes de charge électriques résidentielles, mesurées à un pas de temps de 30 minutes, sont essentielles pour des acteurs comme EDF afin d'optimiser la gestion des réseaux et la prédiction de la demande. Ces séries temporelles peuvent être associées à des métadonnées tabulaires statiques (type d'équipement, caractéristiques du bâtiment, etc.), qui décrivent le système sous-jacent mais ne reflètent pas suffisamment l'impact de ces informations sur la dynamique des courbes qui leur sont associées. Par exemple, un attribut comme "type de chauffage = électrique" n'indique pas si ce chauffage est utilisé de manière intensive le matin ou le soir, ni comment il interagit avec d'autres facteurs comme l'isolation du logement. Ces métadonnées, encodées via des représentations *one-hot* ou d'autres méthodes d'encodage, sont intégrées comme variables exogènes dans les modèles. Cependant, ces encodages limitent la capacité à capturer les interactions entre attributs statiques et comportements dynamiques observés dans les séries temporelles. Dans le cadre de ces recherches, qui portent sur l'explicabilité des modèles multimodaux, nous nous intéressons dans un premier temps à la construction d'un espace latent, capable de structurer conjointement séries temporelles et métadonnées, en utilisant une verbalisation préalable par LLM pour enrichir leur représentation.

1.2 Problématique

L'alignement contrastif multimodal, illustré par des architectures comme CLIP (Radford *et al.*, 2021), permet d'aligner deux modalités distinctes dans un espace latent partagé. Cependant, contrairement aux domaines de l'image ou de l'audio, où les paires (image-texte ou audio-texte) exploitent des informations complémentaires, les séries temporelles que nous étudions sont généralement accompagnées de métadonnées tabulaires. Or, ces métadonnées, bien que descriptives, ne reflètent ni les dynamiques ni les comportements sous-jacents des données. Contrairement aux approches traitant le texte comme une modalité indépendante, nous utilisons un LLM pour verbaliser les métadonnées en descriptions contextualisées. Les encodeurs pré-entraînés comme BERT (Devlin *et al.*, 2019) structurent alors l'espace latent de manière sémantique, plutôt que purement statistique, capturant des dynamiques que les méthodes statistiques classiques ne révèlent pas.

Cette approche, encore émergente pour les séries temporelles, comble les limites des représentations tabulaires. Évaluée sur 20 000 courbes de charge électrique avec leurs métadonnées, elle améliore la structure de l'espace latent, notamment pour la recherche d'information bidirectionnelle. Par exemple, elle permet de retrouver une série temporelle à partir de ses métadonnées (ou inversement). Cette bidirectionnalité repose sur la construction d'un espace latent commun : les séries temporelles et les métadonnées tabulaires étant de nature différente, elles doivent être projetées dans une représentation partagée afin de rendre leur comparaison possible.

Notre contribution principale réside dans la **verbalisation sémantique des métadonnées tabulaires**. Nous proposons un pipeline par LLM qui transforme les métadonnées associées aux séries temporelles en descriptions textuelles contextualisées. L'objectif dépasse la simple reformulation linguistique : la verbalisation explicite les liens entre les attributs contextuels et les comportements attendus du signal. Cette représentation intermédiaire permet d'exploiter les encodeurs textuels pré-entraînés pour structurer l'espace latent, là où un encodage catégoriel classique ne le permettrait pas. Pour valider cette approche, et parce que nos attributs sont catégoriels (plusieurs individus partageant

des profils strictement identiques), nous évaluons notre espace latent via le *Semantic Recall@k* (SR@k). Cette métrique valide toute réponse sémantiquement compatible avec la requête, permettant de diagnostiquer précisément quels attributs sont les mieux structurés dans l'espace latent.

Nous présentons un état de l'art ciblé, détaillons notre méthodologie, puis validons l'approche sur un corpus industriel de courbes de charge électrique avec des performances supérieures aux méthodes de référence.

2 État de l'art

Représentation des métadonnées tabulaires : Les métadonnées associées aux séries temporelles (comme le type d'équipement ou les caractéristiques du système) sont généralement encodées sous forme numérique (one-hot, représentation vectorielle) pour être intégrées dans des modèles prédictifs. Pourtant, ces méthodes d'encodage se contentent de représenter l'appartenance à une catégorie sans expliquer comment ces données influencent concrètement les séries temporelles. Par exemple, un encodage one-hot pour "chauffage électrique" indique simplement la présence de cet équipement, mais ne précise pas les pics de consommation hivernaux, les variations saisonnières ou l'amplitude des signaux qu'il génère. Récemment, TabLLM (Hegselmann *et al.*, 2023) a montré que transformer des attributs tabulaires en phrases naturelles améliore les performances en classification précisément parce que cette verbalisation intègre des connaissances contextuelles absentes des données brutes. Nous étendons cette approche à l'alignement multimodal : la verbalisation ne cible pas une tâche supervisée, mais génère une représentation intermédiaire contextualisée, alignant métadonnées et séries temporelles dans un espace latent. L'apport clé réside dans l'injection de connaissances métiers (ex : les pics de consommation ont lieu entre 18h et 20h en semaine en raison des retours de travail) qui éclairent les relations entre attributs statiques et dynamiques observées.

Apprentissage de représentations pour séries temporelles : L'extraction de caractéristiques a évolué des approches déterministes vers l'apprentissage profond. Les méthodes traditionnelles reposent sur la projection des signaux dans des bases de fonctions — telles que les coefficients de Fourier ou les ondelettes (Mallat, 1999) — ou sur l'analyse de données fonctionnelles (Ramsay & Silverman, 2005), qui traite les séries comme des objets continus via des bases de splines. Par ailleurs, des approches comme ROCKET (Dempster *et al.*, 2020) ont montré l'efficacité de noyaux de convolution aléatoires pour générer des représentations robustes à faible coût computationnel. Plus récemment, l'apprentissage auto-supervisé (TS2Vec (Yue *et al.*, 2022), TNC (Tonekaboni *et al.*, 2021)) a permis d'apprendre des représentations vectorielles denses en exploitant la structure temporelle des données. Enfin, l'émergence des modèles de fondation pour séries temporelles, entraînés sur de larges corpus de données, a marqué un tournant. Bien que ces modèles — Chronos-2 (Ansari *et al.*, 2025), TimesFM (Das *et al.*, 2024) ou Moirai (Woo *et al.*, 2024) — soient principalement conçus pour la prévision, Auer *et al.* (2025) démontrent leur potentiel pour produire des représentations vectorielles transférables, notamment en classification zero-shot. Leurs performances dans ce contexte suggèrent une capacité à capturer les dynamiques sous-jacentes des séries, bien au-delà de la simple tâche de prédiction.

Alignement multimodal impliquant des séries temporelles : Des méthodes comme CLIP (Radford *et al.*, 2021) ou BLIP-2 (Li *et al.*, 2023) utilisent l'alignement contrastif pour relier deux types de données complémentaires, par exemple une image et sa légende. Son extension aux séries temporelles suppose l'existence de descriptions textuelles associées. Des travaux récents comme TS-CLIP (Chen

et al., 2025) proposent d'aligner signaux et textes pour des tâches de recherche d'information, tandis que ChatTS (Xie *et al.*, 2025) s'appuie sur la génération de descriptions synthétiques pour pallier l'absence de données textuelles lors de l'entraînement. Notre approche diffère : nous partons de données où seules des métadonnées tabulaires sont disponibles (par exemple, des tableaux avec le type de chauffage ou la puissance contractuelle), sans texte descriptif associé. Au lieu de chercher à aligner une série avec un texte existant, nous utilisons un LLM pour transformer ces métadonnées en descriptions textuelles, ce qui permet ensuite d'appliquer l'alignement contrastif même en l'absence de texte initial.

3 Modèle d'alignement multimodal : verbalisation et apprentissage contrastif

Notre approche verbalise les métadonnées tabulaires via un LLM pour construire une représentation textuelle intermédiaire, puis projette cette représentation et les séries temporelles dans un espace latent commun via apprentissage contrastif. Elle repose sur trois étapes :

1. un pipeline de verbalisation transformant les métadonnées tabulaires en descriptions textuelles enrichies ;
2. des encodeurs pré-entraînés et gelés de chaque représentation ;
3. un modèle d'alignement de type double encodeur entraîné par une perte contrastive symétrique.

3.1 Formalisation du problème

Soit $\mathcal{X}$ l'espace des séries temporelles de longueurs variables et $\mathcal{C}$ l'espace des métadonnées tabulaires (combinaison de variables catégorielles et numériques). Nous considérons un jeu de données $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$, où chaque individu est défini par une série $x_i \in \mathcal{X}$ et son contexte statique associé $c_i \in \mathcal{C}$.

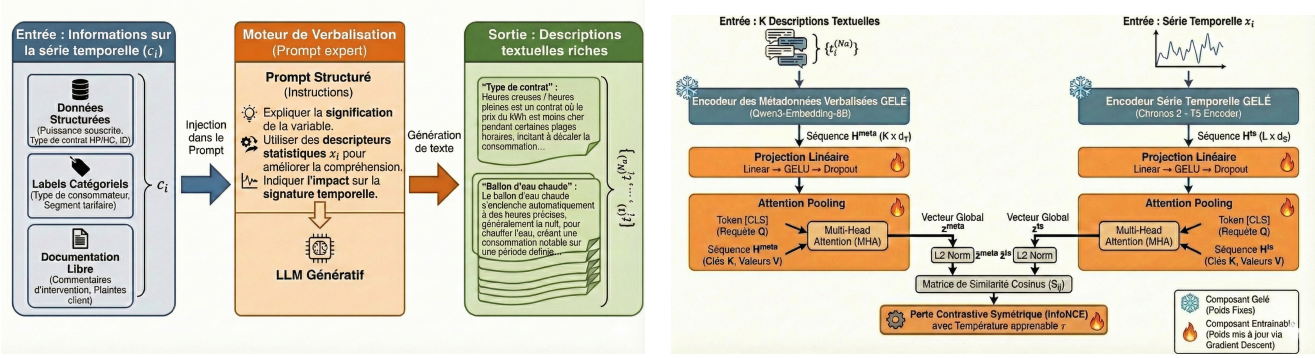
Nous introduisons une fonction de verbalisation $v : (\mathcal{C}, \mathcal{X}) \to \mathcal{T}^{N_a}$, où $\mathcal{T}^{N_a}$ représente l'ensemble des N_a séquences textuelles produites. Pour chaque individu, cette fonction transforme le vecteur de métadonnées $\mathbf{c}_i$ en une liste de descriptions enrichies $\mathbf{t}_i = \{t_i^{(1)}, \ldots, t_i^{(N_a)}\} = v(\mathbf{c}_i, \mathbf{x}_i)$. L'objectif est d'apprendre conjointement deux fonctions de projection dans un espace latent commun de dimension d : l'encodeur de séries temporelles $f_\theta : \mathcal{X} \to \mathbb{R}^d$, et l'encodeur de texte $g_\phi : \mathcal{T} \to \mathbb{R}^d$. Pour assurer l'alignement des deux modalités, nous utilisons la similarité cosinus comme mesure de proximité sémantique dans l'espace latent. Pour deux vecteurs $u, v \in \mathbb{R}^d$, elle est définie par :

$$s(u, v) = \frac{u^\top v}{\|u\|\|v\|}$$

Cette mesure est intégrée à une fonction de perte contrastive de type InfoNCE (van den Oord *et al.*, 2018). Ce choix permet de structurer l'espace latent : en maximisant l'information mutuelle entre les paires positives, elle force le modèle à apprendre des représentations discriminantes.

3.2 Génération de descriptions textuelles

Le pipeline de verbalisation (illustré par la Figure 1a) transforme les métadonnées tabulaires en descriptions textuelles enrichies via un LLM génératif, guidé par un prompt métier. Pour chaque vecteur de métadonnées $\mathbf{c}_i$ de l'individu i, nous transformons c_i en N_a descriptions textuelles $\{t_i^{(1)}, \ldots, t_i^{(N_a)}\}$ via Qwen2.5-32B-Instruct (Yang *et al.*, 2024), choisi pour son équilibre entre capacités de génération contextualisée et coût d'inférence. Chaque description correspond à un attribut de c_i, mais le processus intègre également une interprétation contextuelle de statistiques du signal x_i (ex. saisonnalité marquée, tendances) pour distinguer des séries qui partagent par ailleurs les mêmes métadonnées tabulaires. Concrètement, ces statistiques ne sont pas utilisées comme caractéristiques brutes, mais servent de base au sein du prompt pour générer des énoncés interprétables (ex. " un ratio de saisonnalité hiver/été de 700 % suggère une demande accrue de chauffage en période froide "). Cette approche, via le prompt métier, permet de produire pour chaque attribut une description combinant sa **signification métier** (ex. type de bâtiment), son **impact supposé** sur la dynamique temporelle (ex. " la présence d'une piscine augmente la consommation estivale "), et enfin une **interprétation des motifs temporels associés** (ex. " les pics nocturnes indiquent un usage résidentiel ").



(a) Pipeline de verbalisation.

(b) Architecture globale.

FIGURE 1 – Vue d'ensemble du pipeline de verbalisation et de l'architecture d'alignement.

Des exemples concrets de descriptions générées sont fournis en Annexe D.

3.3 Encodage des représentations

Notre architecture (Figure 1) suit une stratégie dual-encoder avec encodeurs gelés, suivant le paradigme de LiT (Zhai *et al.*, 2022) et BLIP-2 (Li *et al.*, 2023). Seuls les modules de projection et d'attention (en orange) sont entraînés.

3.3.1 Encodage des métadonnées verbalisées

Les N_a descriptions textuelles sont encodées par Qwen3-Embedding-8B (Zhang *et al.*, 2025). Le modèle projette les descriptions dans un espace de dimension $d_T = 4096$, préservant l'information sémantique nécessaire à l'alignement. Nous extrayons le dernier état caché de chaque description, produisant la matrice $H_i^{\text{meta}} = \left[\mathbf{h}_i^{(1)}, \ldots, \mathbf{h}_i^{(N_a)}\right] \in \mathbb{R}^{N_a \times d_T}$.

3.3.2 Encodage des séries temporelles

L'encodeur de séries temporelles (branche droite de la Figure 1b) s'appuie sur Chronos-2 (Ansari *et al.*, 2025). Chaque série x_i est d'abord normalisée par sa moyenne, puis convertie en une séquence discrète de jetons (*tokens*) via un processus de quantification. Cette séquence est ensuite encodée par le modèle gelé pour produire $H_i^{\text{ts}} = E_\theta(\text{Quantize}(x_i)) = \left[\mathbf{h}_i^{(1)}, \ldots, \mathbf{h}_i^{(L)}\right] \in \mathbb{R}^{L \times d_S}$, où L désigne la longueur de la séquence de jetons (*tokens*) et $d_S = 768$ la taille de l'espace de dimension de Chronos-2.

3.4 Architecture d'alignement

3.4.1 Projection

Les représentations H^{meta} et H^{ts} sont projetées vers un espace commun de dimension d_{proj} via une transformation non-linéaire : normalisation par couches, projection linéaire, activation GELU (Hendrycks & Gimpel, 2016), et Dropout :

$$\tilde{H}^{\text{meta}} = \text{Dropout}\Big(\text{GELU}\big(\text{LN}(H^{\text{meta}})W_{\text{meta}} + b_{\text{meta}}\big)\Big) \in \mathbb{R}^{N_a \times d_{\text{proj}}}, \tag{1}$$

$$\tilde{H}^{\text{ts}} = \text{Dropout}\Big(\text{GELU}\big(\text{LN}(H^{\text{ts}})W_{\text{ts}} + b_{\text{ts}}\big)\Big) \in \mathbb{R}^{L \times d_{\text{proj}}}. \tag{2}$$

où $W_{\text{meta}} \in \mathbb{R}^{d_T \times d_{\text{proj}}}$ et $W_{\text{ts}} \in \mathbb{R}^{d_S \times d_{\text{proj}}}$ sont les matrices de poids entraînables des projections linéaires, et $b_{\text{meta}}, b_{\text{ts}} \in \mathbb{R}^{d_{\text{proj}}}$ les biais associés.

3.4.2 Agrégation par Attention Pooling

Pour obtenir une représentation par individu, nous adoptons un mécanisme d'Attention Pooling (Lee *et al.*, 2019; Ilse *et al.*, 2018) : un token de requête $\mathbf{q} \in \mathbb{R}^{1 \times d_{\text{proj}}}$, optimisé avec le modèle, interroge la séquence via une attention croisée multi-têtes (Vaswani *et al.*, 2017) :

$$\mathbf{z} = \text{MHA}(\mathbf{Q} = \mathbf{q},\ \mathbf{K} = \tilde{H},\ \mathbf{V} = \tilde{H}) \in \mathbb{R}^{d_{\text{proj}}}. \tag{3}$$

Ce mécanisme fonctionne quelle que soit la longueur de la séquence, se concentre sur les positions informatives, et permet une interprétation via les poids d'attention. Les vecteurs sont normalisés sur la sphère unité par normalisation L2 afin de préparer le calcul de similarité :

$$\mathbf{z}^{\text{meta}} = \frac{\text{AttPool}(\tilde{H}^{\text{meta}})}{\|\text{AttPool}(\tilde{H}^{\text{meta}})\|_2}, \quad \mathbf{z}^{\text{ts}} = \frac{\text{AttPool}(\tilde{H}^{\text{ts}})}{\|\text{AttPool}(\tilde{H}^{\text{ts}})\|_2}. \tag{4}$$

3.5 Entraînement par perte contrastive symétrique

L'alignement est optimisé via une perte InfoNCE symétrique (van den Oord *et al.*, 2018). Pour un lot (*batch*) de B paires $\{(\mathbf{z}_i^{\text{meta}}, \mathbf{z}_i^{\text{ts}})\}_{i=1}^{B}$, nous calculons deux matrices de similarité :

$$S_{ij}^{\text{meta}\to\text{ts}} = \frac{1}{\tau} \cdot \mathbf{z}_i^{\text{meta}} \cdot (\mathbf{z}_j^{\text{ts}})^\top, \quad S_{ij}^{\text{ts}\to\text{meta}} = \frac{1}{\tau} \cdot \mathbf{z}_i^{\text{ts}} \cdot (\mathbf{z}_j^{\text{meta}})^\top, \tag{5}$$

où τ est un paramètre de température apprenable. La fonction de perte combine les contributions des deux directions :

$$\mathcal{L} = \frac{1}{2}\Big(\mathcal{L}_{\text{meta}\to\text{ts}} + \mathcal{L}_{\text{ts}\to\text{meta}}\Big), \tag{6}$$

avec

$$\mathcal{L}_{\text{meta}\to\text{ts}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}^{\text{meta}\to\text{ts}})}{\sum_{j=1}^{B}\exp(S_{ij}^{\text{meta}\to\text{ts}})}, \quad \mathcal{L}_{\text{ts}\to\text{meta}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii}^{\text{ts}\to\text{meta}})}{\sum_{j=1}^{B}\exp(S_{ij}^{\text{ts}\to\text{meta}})}. \tag{7}$$

Seuls τ, les projections et l'Attention Pooling sont entraînés (encodeurs gelés). Détails en Annexe E.

4 Protocole expérimental

4.1 Jeu de données

L'approche est évaluée sur un corpus propriétaire issu d'EDF, regroupant 20 000 courbes de charge électrique résidentielle. Chaque série temporelle couvre deux ans avec un pas de 30 minutes ($L \approx 35\,040$ points) et est associée à 10 attributs décrivant le logement et l'installation électrique (voir Annexe C pour la liste exhaustive et les définitions).

4.2 Baselines et ablations

Nous comparons notre approche à des méthodes de référence afin d'évaluer l'apport de la verbalisation et de l'utilisation de modèles de fondation.

MLP + CNN : Cette architecture représente une approche d'apprentissage profond classique mais aveugle à la sémantique. Les métadonnées y sont encodées en *one-hot* puis projetées par un MLP, tandis que les séries temporelles sont traitées par un encodeur CNN. L'objectif est d'évaluer si une architecture neuronale simple, sans connaissance métier préalable, suffit à structurer l'espace latent.

ACP + ACC : Cette baseline repose sur des méthodes statistiques linéaires. Une Analyse en Composantes Principales (ACP) est appliquée séparément sur les métadonnées brutes et sur les séries temporelles (représentées par leurs caractéristiques statistiques) pour réduire la dimensionnalité. Une Analyse Canonique des Corrélations (ACC) cherche ensuite à maximiser la corrélation entre ces composantes. Ce qui permet de mesurer l'apport des transformations non-linéaires et de l'encodage sémantique par rapport à une méthode statistique traditionnelle.

Ablations : Pour isoler les contributions respectives de chaque brique technologique, deux configurations intermédiaires sont testées :

1. **Chronos-2 + MLP**, qui conserve le modèle de fondation pour les séries mais remplace la verbalisation par un encodage one-hot classique. L'objectif est de vérifier si le gain de performance provient uniquement de la puissance de l'encodeur temporel.
2. **CNN + Verbalisation**, qui conserve l'enrichissement par LLM mais remplace Chronos-2 par un CNN entraîné de zéro. Cette configuration vise à prouver la nécessité d'utiliser un modèle de fondation pour capturer les dynamiques du signal.

4.3 Protocole d'évaluation

L'espace latent est évalué via une tâche de recherche d'information bidirectionnelle : retrouver les séries $\mathbf{x}$ à partir des métadonnées $\mathbf{c}$ (**Métadonnées** $\rightarrow$ **Série temporelle**) et identifier les métadonnées décrivant une courbe (**Série temporelle** $\rightarrow$ **Métadonnées**). Le processus consiste à classer les candidats de la modalité opposée par similarité cosinus (π_i). Prenons l'exemple d'une recherche **Série temporelle** $\rightarrow$ **Métadonnées** : on projette une courbe $\mathbf{x}_i$ inconnue dans l'espace latent pour chercher, parmi les métadonnées disponibles, celles dont la description textuelle est la plus proche. L'objectif est de retrouver en tête de liste des profils métier compatibles (ex. : même type de chauffage ou de consommation). Pour gérer les cas où plusieurs individus partagent des caractéristiques identiques, nous utilisons des masques binaires de compatibilité : pour chaque attribut $a \in \mathcal{A}$, on pose $M_{i,j}^{(a)} = 1$ si les individus i et j partagent la même valeur pour a.

Le masque global, $M_{i,j}^{\text{global}}$, vaut 1 si leurs métadonnées sont strictement identiques. La performance est mesurée par le *Semantic Recall@k* (SR@k), soit la probabilité de trouver au moins un voisin compatible dans le top k :

$$\text{SR@k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left[\sum_{j \in \text{top } k} M_{i,\pi_i(j)}^{\text{global}} \geq 1 \right]. \tag{8}$$

L'analyse est menée à deux niveaux : le SR@k global pour la recherche bidirectionnelle, et le SR@1 par attribut pour vérifier si la valeur prédite au premier rang pour un attribut spécifique est correcte.

5 Résultats

5.1 Recherche d'information bidirectionnelle

TABLEAU 1 – Recherche d'information bidirectionnelle. Moyenne du SR@k (%) sur le jeu de test.

	Notre approche		MLP+CNN		ACP+ACC	
Direction	**SR@1**	**SR@5**	SR@1	SR@5	SR@1	SR@5
Meta $\rightarrow$ TS	**24,91**	**55,28**	14,41	40,39	5,71	18,35
TS $\rightarrow$ Meta	**24,20**	**55,33**	12,34	34,63	5,41	17,24
Moyenne	**24,56**	**55,31**	13,37	37,51	5,56	17,80

Notre approche atteint 24,56% en SR@1 moyen, surpassant MLP+CNN ($\times$1,8) et ACP+ACC ($\times$4,4)(Tableau 1). La comparaison avec MLP+CNN isole la contribution de la verbalisation : à architecture comparable, le doublement des performances confirme que transformer les métadonnées en descriptions enrichies structure l'espace latent plus efficacement qu'un encodage catégoriel. L'écart avec PCA+CCA illustre les limites des approches linéaires.

TABLEAU 2 – Étude d'ablation. Moyenne du SR@k (%) sur le jeu de test.

Configuration	SR@1	SR@5
Notre approche (Chronos-2 + Verbalisation)	**24,56**	**55,31**
Chronos-2 + MLP (sans verbalisation)	19,30	42,80
CNN + Verbalisation (sans Chronos-2)	19,83	50,23

Les performances des deux configurations d'ablation sont systématiquement inférieures à celles de l'approche complète (Tableau 2). Ce constat suggère qu'aucun de ces composants, utilisé isolément, ne suffit à structurer l'espace latent avec la même efficacité. La progression observée semble ainsi résulter de la complémentarité entre l'enrichissement sémantique par le LLM et les capacités de capture des dynamiques temporelles de Chronos-2.

5.2 Analyse par attribut

TABLEAU 3 – Moyenne du SR@1 par attribut (%) sur le jeu de test. La colonne « Aléatoire » indique l'espérance mathématique du succès sous appariement aléatoire.

Attribut	Type	**Notre approche**	MLP+CNN	ACP+ACC	Aléatoire
Climatisation	binaire	**98,2%**	93,2%	93,5%	65,1%
Chauffage électrique	binaire	**98,3%**	95,5%	85,2%	52,1%
Véhicule électrique	binaire	**96,7%**	94,5%	89,4%	91,5%
Type ECS	3 modalités	**96,4%**	95,9%	76,8%	44,7%
Asservissement ECS	binaire	**91,3%**	89,3%	84,6%	86,2%
Option tarifaire	7 modalités	**84,0%**	73,6%	75,5%	46,8%
Plage heures creuses	68 modalités	**77,4%**	64,9%	70,3%	32,5%
Nombre d'occupants	6 modalités	**71,5%**	65,5%	49,0%	28,8%
Puissance souscrite	10 modalités	**60,8%**	47,7%	40,7%	37,1%
Superficie	37 modalités	**28,1%**	24,9%	15,0%	8,6%

On remarque que la qualité de l'alignement varie selon la manière dont chaque attribut marque concrètement la courbe de charge. Pour des équipements comme le chauffage ou la climatisation, qui laissent une signature importante sur le signal, les résultats sont encourageants. Pour la superficie, l'impact de cet attribut avec la consommation est moins évident, ce qui peut expliquer pourquoi les scores sont moins bons par rapport aux autres attributs.

6 Discussion

Ce travail s'inscrit dans la continuité des approches d'alignement contrastif, mais dans un cadre différent de la littérature image-texte : nous considérons ici l'alignement entre séries temporelles et métadonnées tabulaires. Contrairement aux couples multimodaux classiques, où le texte apporte une description riche et souvent indépendante du signal, les métadonnées sont sémantiquement pauvres lorsqu'elles sont encodées de manière standard (one-hot, représentations vectorielles apprises). L'enjeu n'est donc pas d'aligner une "modalité texte" disponible, mais de rendre ces métadonnées alignables avec la série en les transformant en une représentation textuelle exploitable par un encodeur sémantique pré-entraîné. Dans ce cadre, la verbalisation par LLM agit comme une représentation intermédiaire qui injecte de l'information métier : elle permet de relier des métadonnées (chauffage, option tarifaire, etc.) à des motifs attendus dans la forme des courbes, ce qui se traduit par une meilleure structuration de l'espace latent pour la recherche d'information bidirectionnelle. Nos résultats montrent que la structuration de l'espace latent est plus efficace pour les attributs qui ont un impact clair et direct sur la forme des courbes (comme le chauffage, la climatisation, le véhicule électrique ou la production d'eau chaude sanitaire). En revanche, les attributs moins directement liés

à des motifs temporels précis, par exemple la superficie du logement, sont plus difficiles à organiser de manière cohérente. Cela s'explique par la nature même des données : certains équipements génèrent des schémas répétitifs et faciles à identifier, comme les pics saisonniers ou les cycles de consommation, tandis que d'autres influencent plutôt le niveau général de la courbe ou dépendent de facteurs externes non mesurés. Une limite majeure de notre méthode réside dans sa dépendance au domaine d'application. Le prompt de verbalisation utilisé ici est conçu pour les courbes de consommation résidentielle, avec des termes spécifiques comme les "heures creuses" ou la "saisonnalité du chauffage". Pour l'adapter à d'autres secteurs comme la santé, l'industrie ou la finance, il ne suffirait pas de changer quelques mots : il faudrait repenser les liens de cause à effet et définir ce qui constitue une "signature" pertinente dans le nouveau contexte. Notre approche offre donc avant tout un cadre méthodologique centré sur la transformation de données tabulaires en descriptions textuelles exploitables, validé par un protocole d'évaluation (SR@k) adapté aux cas où plusieurs individus partagent les mêmes caractéristiques. Pour aller plus loin, une expérience complémentaire serait utile : comparer une verbalisation détaillée, expliquant les effets attendus sur la courbe, à une version minimaliste, se limitant à des descriptions basiques comme "le logement est équipé d'un chauffage électrique". Cela permettrait de déterminer si l'amélioration des performances vient simplement de l'utilisation d'un encodeur textuel pré-entraîné, ou si c'est bien l'enrichissement (explication des motifs, vocabulaire technique, relations causales) qui fait la différence.

7 Conclusion

Nous avons présenté une architecture d'alignement contrastif entre séries temporelles et métadonnées tabulaires, dans laquelle les métadonnées sont converties en descriptions textuelles afin d'être projetées dans un espace latent commun avec les courbes. Le point clé n'est pas l'usage d'un LLM "spécialisé", mais l'information métier injectée par le prompt de verbalisation, qui explicite le lien entre attributs statiques et motifs attendus dans la dynamique du signal. Évaluée sur 20 000 courbes de charge électrique, notre approche améliore la recherche d'information bidirectionnelle par rapport à un encodage catégoriel classique. L'analyse par attribut confirme que l'alignement est particulièrement robuste pour les équipements dont l'impact sur la courbe est marqué (chauffage, climatisation, véhicule électrique, ECS), tout en restant plus difficile pour des variables moins directement observables dans la forme du signal (par exemple la superficie). Ce travail montre que l'enrichissement des données brutes par un LLM, guidé par un prompt métier, permet d'optimiser la recherche d'information. La structure de l'espace latent ainsi obtenu ouvre également la voie à de nombreuses autres possibilités d'usage.

Remerciements

Ces travaux s'inscrivent dans le cadre d'une thèse CIFRE menée conjointement entre EDF R&D et le laboratoire ERIC de l'Université Lumière Lyon 2. Je tiens à remercier mes directeurs de thèse, Julien Jacques (Université Lumière Lyon 2) et Julien Velcin (École Centrale Lyon), pour leur encadrement scientifique et leurs conseils. Je remercie également mes tuteurs industriels, Laurent Bozzi et Alice Duquenne (EDF R&D), pour leur soutien et le cadre industriel qu'ils offrent à ces recherches.

Références

ANSARI A. F., SHCHUR O., KÜKEN J., AUER A., HAN B., MERCADO P., RANGAPURAM S. S., SHEN H., STELLA L., ZHANG X., GOSWAMI M., KAPOOR S., MADDIX D. C. *et al.* (2025). Chronos-2 : From univariate to universal forecasting. *arXiv preprint arXiv :2510.15821*.

AUER A., ANSARI A. F., SHCHUR O. & JANUSCHOWSKI T. (2025). Pre-trained forecasting models : Strong zero-shot feature extractors for time series classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

CHEN J. *et al.* (2025). Ts-clip : Time series understanding by clip. *arXiv preprint*.

DAS A., KONG W., LEBER A., MATHEWS R. & SEN S. (2024). A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv :2310.10688*.

DEMPSTER A., PETITJEAN F. & WEBB G. I. (2020). Rocket : Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, **34**(5), 1454–1495.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

HEGSELMANN S., BUENDIA A., LANG H., AGRAWAL M., JIANG X. & SONTAG D. (2023). Tabllm : Few-shot classification of tabular data with large language models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, p. 5549–5581 : PMLR.

HENDRYCKS D. & GIMPEL K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv :1606.08415*.

ILSE M., TOMCZAK J. M. & WELLING M. (2018). Attention-based deep multiple instance learning. In *ICML*.

LEE J., LEE Y., KIM J., KOSIOREK A. R., CHOI S. & TEH Y. W. (2019). Set transformer : A framework for attention-based permutation-invariant neural networks. In *ICML*.

LI J., LI D., SAVARESE S. & HOI S. C. H. (2023). Blip-2 : Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)* : PMLR.

MALLAT S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press, 2nd édition.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, p. 8748–8763 : PMLR.

RAMSAY J. O. & SILVERMAN B. W. (2005). *Functional Data Analysis*. Springer, 2nd édition.

TONEKABONI S., EYTAN D. & GOLDENBERG A. (2021). Unsupervised representation learning for time series with temporal neighborhood coding. In *Proceedings of ICLR*.

VAN DEN OORD A., LI Y. & VINYALS O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv :1807.03748*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, p. 5998–6008.

WOO G., LIU C., KUMAR A., XIONG C., SAVARESE S. & SAHOO D. (2024). Unified training of universal time series forecasting transformers. *arXiv preprint arXiv :2402.02592*.

XIE Z., LI Z., HE X., XU L., WEN X., ZHANG T., CHEN J., SHI R. & PEI D. (2025). Chatts : Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint*.

YANG A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., LI C., LIU D., HUANG F. *et al.* (2024). Qwen2.5 technical report. *arXiv preprint arXiv :2412.15115*.

YUE Z., WANG Y., DUAN J., YANG T., HUANG C., TONG Y. & XU B. (2022). Ts2vec : Towards universal representation of time series. In *Proceedings of AAAI*, p. 8980–8987.

ZHAI X., WANG X., MUSTAFA B., STEINER A., KEYSERS D., KOLESNIKOV A. & BEYER L. (2022). Lit : Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

ZHANG Y., LI M., LONG D., ZHANG X., LIN H., YANG B., XIE P., YANG A., LIU D., LIN J., HUANG F. & ZHOU J. (2025). Qwen3 embedding : Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv :2506.05176*.

Annexes

A Glossaire des termes

Afin d'assurer la précision de l'analyse et de l'évaluation, nous définissons la hiérarchie terminologique et conceptuelle utilisée dans cet article :

— **Individu (i)** L'unité d'observation élémentaire du jeu de données $\mathcal{D}$ (ici, un foyer résidentiel). Chaque individu est défini par le couple $(\mathbf{x}_i, \mathbf{c}_i)$ associant un signal dynamique et un contexte statique.

— **Série temporelle ($\mathbf{x}_i$)** La modalité dynamique représentant l'évolution de la consommation électrique de l'individu sur une période donnée. Elle constitue le signal brut à aligner.

— **Métadonnées ($\mathcal{C}$)** La modalité regroupant l'ensemble des informations contextuelles et statiques associées à l'individu. Dans ce travail, elles sont de nature tabulaire avant verbalisation.

— **Attribut (a)** Une variable ou caractéristique spécifique constitutive du vecteur de métadonnées (ex. : *Type de chauffage*). Le nombre total d'attributs par individu est noté N_a. Il correspond au nombre de colonnes de la table source.

— **Valeur (val)** L'instance concrète prise par un attribut pour un individu donné (ex. : « Électrique »).

— **Verbalisation** Le processus de transformation des N_a (Nombre d'Attributs) valeurs d'attributs en descriptions textuelles enrichies via un modèle de langage (LLM), agissant comme une représentation sémantique intermédiaire.

— **Espace latent commun** L'espace vectoriel de dimension d dans lequel les encodeurs projettent les deux modalités. C'est dans cet espace que la proximité sémantique est calculée via la similarité cosinus.

— **Alignement contrastif** L'objectif d'apprentissage (perte InfoNCE) visant à organiser l'espace latent en maximisant la similarité entre les paires $(\mathbf{z}_i^{ts}, \mathbf{z}_i^{meta})$ et en la minimisant pour les paires non associées.

— **Semantic Recall@k (SR@k)** Métrique d'évaluation proposée. Contrairement au rappel classique, elle valide une prédiction si le résultat appartient à la classe de compatibilité sémantique de la cible (partage d'attributs identiques), neutralisant ainsi l'effet des « doublons sémantiques ».

B Prompt de verbalisation

Le prompt suivant est utilisé pour générer les descriptions textuelles via Qwen2.5-32B-Instruct. Il est appliqué séparément à chaque métadonnée d'un individu. Le modèle étant optimisé pour l'anglais, le prompt a été rédigé dans cette langue. Par conséquent, l'ensemble des métadonnées (noms, définitions et valeurs), initialement en français dans notre corpus, a été traduit en anglais au préalable avant d'être injecté dans le prompt.

```
### ROLE & DOMAIN IDENTITY
You are a Senior Load Curve Analyst specializing in French
residential electricity profiles (Enedis/Linky data standards).

### OBJECTIVE
Generate a dense, physically-grounded text embedding for a
Contrastive Multimodal Model. The text must serve as a semantic
anchor, allowing the AI to "reconstruct" the curve's morphology
solely from your description.

### INPUT ATTRIBUTE
- Attribute Name: {name}
- Definition: {definition}
- Value: "{variable_value}"
- Causal Strength: {causal_strength}

EXPECTED LOAD CURVE SIGNATURE:
{expected_signature}

QUANTITATIVE IMPACT ON ELECTRICITY CONSUMPTION:
{impact_text}

### INSTRUCTIONS
Write a concise, expert-level paragraph describing the impact
of this variable on the load curve.
1. Translate the value into a physical grid phenomenon.
2. Describe the resulting changes in morphology, magnitude,
   and temporal dynamics.
3. Explain the causality using "because/due to" logic.

RULES:
- Use signal processing vocabulary (baseline, peak amplitude,
  load shifting, seasonality, intermittency, variance).
- Focus on causal relationships.
- Do NOT speculate beyond the data.
- Language: English.
```

Les placeholders `{name}`, `{definition}`, `{variable_value}`, etc. sont remplacés dynamiquement par les métadonnées de chaque individu. Les paramètres de génération sont : $\texttt{temperature} = 0.1$, $\texttt{max\_tokens} = 180$, $\texttt{top\_p} = 0.9$.

C Description des métadonnées du corpus

Le corpus est composé de 20 000 courbes de charge électrique résidentielle, chacune associée à 10 métadonnées tabulaires décrivant le logement et l'installation électrique. Le Tableau 4 détaille ces attributs.

Attribut	Type	Description
Climatisation	binaire	Présence d'un système de climatisation
Chauffage électrique	binaire	Présence d'un système de chauffage électrique (convecteurs, pompe à chaleur, etc.)
Véhicule électrique	binaire	Présence d'un véhicule électrique en recharge au domicile
Type ECS	3 modalités	Type de production d'eau chaude sanitaire : électrique, gaz, ou thermodynamique
Asservissement ECS	binaire	Pilotage de l'eau chaude sanitaire en heures creuses
Option tarifaire	7 modalités	Contrat tarifaire souscrit (Base, HP/HC, Tempo, etc.)
Plage heures creuses	68 modalités	Créneau horaire des heures creuses selon la zone de distribution
Nombre d'occupants	6 modalités	Nombre de personnes résidant dans le logement (1 à 6+)
Puissance souscrite	10 modalités	Puissance maximale du contrat (3 à 36 kVA)
Superficie	37 modalités	Surface habitable discrétisée

TABLEAU 4 – Description des 10 métadonnées tabulaires associées à chaque courbe de charge.

Ces métadonnées couvrent trois catégories : les équipements énergivores (chauffage, climatisation, véhicule électrique, eau chaude sanitaire), les caractéristiques contractuelles (option tarifaire, plage heures creuses, puissance souscrite), et les caractéristiques du logement (nombre d'occupants, superficie). Leur impact sur la dynamique de consommation varie : les équipements induisent des motifs temporels localisés (pics saisonniers, cycles de charge), tandis que les caractéristiques du logement modulent le niveau de consommation de façon plus diffuse.

D Exemple de verbalisation

Cette annexe illustre le pipeline de verbalisation avec deux exemples concrets issus du corpus.

Exemple 1 : Chauffage électrique

Entrée (variable tabulaire) :

```
Attribute: CHAUFF_ELEC (Electric heating)
Value: "Oui"
```

Sortie (description générée par Qwen2.5-32B-Instruct) :

> The presence of electric heating significantly elevates the annual consumption, consistent with households having this equipment. The load curve exhibits strong winter seasonality, with peak amplitudes 3-5x higher in December to February compared to June to August, due to the Joule effect causing elevated baseline and intermittent peaks from thermostatic cycling. This seasonal variance is unusually high, reflecting the substantial energy demand for heating during colder months.

Exemple 2 : Plage heures creuses

Entrée (variable tabulaire) :

```
Variable: PLAGE_HC (Off-peak hours time range)
Value: "HC2300_0700" (23h00-07h00)
```

Sortie (description générée) :

> The off-peak hours from 23 :00 to 07 :00 significantly influence the load curve, causing elevated consumption peaks during these times due to resistance heating and appliance charging. This results in a pronounced load shifting, with baseline consumption remaining relatively low outside these periods. The seasonal variance is unusually high, consistent with households having electric heating, creating a distinct peak amplitude during winter months because of the Joule effect and thermostatic cycling.

E Configuration d'entraînement

Pour la configuration du modèle, nous utilisons Qwen3-Embedding-8B ($d_T = 4096$) et Chronos-2 ($d_S = 768$) comme encodeurs dont les paramètres restent gelés. L'apprentissage se concentre ainsi exclusivement sur les projections linéaires ($W_{\text{meta}}, b_{\text{meta}}, W_{\text{ts}}, b_{\text{ts}}$), le token de requête $\mathbf{q}$ dédié au mécanisme d'Attention Pooling, ainsi que la température τ, initialisée à 0,07.

L'entraînement est réalisé sur un corpus de 20 000 courbes divisé en 16 000 exemples d'entraînement, 2 000 de validation et 2 000 de test. Le modèle est optimisé par Adam avec un taux d'apprentissage

de 10^{-3}, une taille de batch $B = 512$, sur 100 époques avec arrêt anticipé basé sur le SR@1 de validation. La dimension de projection est fixée à $d_{\text{proj}} = 512$. La température τ est initialisée à $0{,}07$ et apprise conjointement avec les projections et l'Attention Pooling.