Agent Neuro-Symbolique pour l'Identification de Relations Causales

Baptiste Brunet de la Charie[1,2], Elöd Egyed-Zsigmond[2], Thomas Veran[1], Ludovic Moncla[2],

(1) Relyens, 18 rue Edouard Rochet, 69372 Lyon, France

(2) INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France

baptiste.brunetdelacharie@relyens.eu, elod.egyed-zsigmond@insa-lyon.fr, thomas.veran@relyens.eu, ludovic.moncla@insa-lyon.fr

RÉSUMÉ

L'extraction de relations causales entre évènements reste un défi en raison de la complexité sémantique, des dépendances à longue distance, de la complexité d'annotation des benchmarks existants et du fort déséquilibre des classes. Les LLMs appliqués à cette tâche souffrent par ailleurs d'asymétries directionnelles et d'hallucinations en l'absence de supervision. Nous présentons un cadre agentique *zero-shot* augmentant un LLM de deux outils légers : `coherence_check`, appliquant des règles logiques de cohérence directionnelle propres au jeu de données, et `counterfactual_pairs`, réévaluant les dépendances causales par génération contrefactuelle au niveau sémantique. Cette conception neuro-symbolique garantit une portabilité multilingue sans adaptation linguistique. Évalué sur les cinq langues de MECI, notre système obtient des résultats compétitifs avec les méthodes supervisées sans aucune donnée d'entraînement. Nous mettons également en évidence un biais directionnel systématique des LLMs pour l'ECI, stable au rééchantillonnage. L'ensemble du code est disponible en *open-source*.

ABSTRACT

Neuro-Symbolic Agent for Causal Relationship Identification

Extracting causal relations between events remains challenging due to semantic complexity, long-distance dependencies, the annotation complexity of existing benchmarks, and severe class imbalance. LLMs applied to this task further suffer from directional asymmetries and hallucinations in the absence of supervision. We present a *zero-shot* agentic framework augmenting an LLM with two lightweight tools : `coherence_check`, enforcing dataset-specific logical consistency rules on the predicted relation graph, and `counterfactual_pairs`, reassessing causal dependencies through counterfactual generation at the semantic level. This neuro-symbolic design provides natural multilingual portability without linguistic adaptation. Evaluated on the five languages of MECI, our system achieves competitive results with supervised methods without any training data. We further expose a systematic directional bias of LLMs for ECI, stable under resampling. All code is released as *open-source*.

MOTS-CLÉS : Extraction de relations causales, Agents LLM, Apprentissage zero-shot.

KEYWORDS: Causal Relation Extraction, LLM Agents, Zero-shot Learning.

1 Introduction

En traitement automatique des langues (TAL), l'extraction de relations structurées à partir de textes non structurés est cruciale pour des applications telles que la réponse aux questions, la construction de graphes de connaissances et la compréhension de récits événementiels. Il convient ici de distinguer l'identification de la causalité événementielle (*Event Causality Identification*, ECI), qui se limite à détecter la présence d'un lien causal sans en préciser le sens ni la classification, de l'extraction de relations d'évènements (*Event Relation Extraction*, ERE). L'ERE, au contraire, vise à identifier, classifier et orienter les relations entre les mentions d'évènements au sein d'un document. Notre approche exploite la directionnalité des relations uniquement pour améliorer la prédiction de leur existence, et se compare à des méthodes ECI seulement (celles qui se limitent à la détection de l'existence sans orientation ni étiquette). Cependant, ces tâches se heurtent à des défis majeurs : la complexité de l'annotation limite souvent la taille des jeux de données (Wang *et al.*, 2022), les dépendances sémantiques entre évènements sont complexes (Guan *et al.*, 2024), et la rareté des instances causales positives entraîne un fort déséquilibre des classes (Caselli & Vossen, 2017).

Les avancées récentes des grands modèles de langage (*Large Language Models*, LLM) ont orienté l'ERE vers des approches génératives et basées sur des agents (Guan *et al.*, 2025; Cai *et al.*, 2025). Bien que ces modèles s'appuient sur des instructions (*prompts*) pour inférer des relations, ils peinent encore face aux relations implicites, aux hallucinations et aux asymétries directionnelles. Cela a été étudié pour des relations beaucoup plus simples, par exemple un LLM entraîné sur le motif « A est B » peut échouer à déduire que « B est A », un phénomène connu sous le nom de *Reversal Curse* (Berglund *et al.*, 2023). De plus, les approches supervisées traditionnelles s'appuient souvent sur des marqueurs lexicaux ou discursifs propres à chaque langue. Cette dépendance limite fortement leur portabilité multilingue et empêche leur application à de nouveaux domaines sans données d'entraînement spécifiques.

Afin de pallier ces limites, nous proposons un cadre agentique dédié à l'extraction et la classification orientée de ces relations (ERE au niveau du document). Notre approche enrichit un LLM avec deux outils spécialisés dans le raisonnement causal : `coherence_check`, qui évalue le respect de règles logiques de cohérence directionnelle, et `counterfactual_pairs`, qui évalue les dépendances causales par la génération de paires contrefactuelles. Contrairement à l'utilisation de marqueurs lexicaux de surface, ce test contrefactuel opère directement sur la sémantique des évènements. Cela confère à notre méthode une portabilité multilingue inhérente. L'ensemble constitue une architecture neuro-symbolique *zero-shot* : le LLM génère des relations candidates, et les outils en assurent la validation logique via des règles déclaratives, le tout sans requérir de données d'entraînement. Cette conception s'inspire des systèmes de débat multi-agents (Guan *et al.*, 2025; Wang & Huang, 2024) et des cadres de raisonnement décomposés (Cai *et al.*, 2025), offrant un raffinement itératif. Évalué sur le jeu de données MECI (Lai *et al.*, 2022), notre agent obtient des résultats compétitifs par rapport aux méthodes supervisées, tout en opérant dans un contexte strictement *zero-shot*.

Nos contributions incluent : (1) l'identification et l'évaluation empirique d'outils spécifiquement adaptés au raisonnement causal (le test contrefactuel et la vérification de cohérence des relations), intégrés à une boucle agentique de type délibération-vérification-réparation ; (2) le développement d'une architecture neuro-symbolique *zero-shot* qui combine raisonnement génératif et vérification logique, aisément portable à de nouveaux jeux de données par la simple mise à jour de ses règles ; (3) des évaluations démontrant des performances supérieures à celles des LLMs non augmentés et compétitives avec les modèles supervisés ; (4) la mise à disposition de ressources *open-source* pour

garantir la reproductibilité ; (5) la mise en évidence expérimentale des difficultés des LLMs face à l'antisymétrie directionnelle des relations causales ; (6) la démonstration par rééchantillonnage que ces erreurs directionnelles persistantes relèvent d'un biais systématique inhérent aux modèles, plutôt que d'une simple variation stochastique.

2 Travaux connexes

Les développements récents dans le domaine de l'ECI ont conduit à une grande diversité de méthodes, résumées par Cheng *et al.* (2026). Celles-ci incluent entre autres les modèles d'encodage, le raisonnement sur graphes d'évènements et le *fine-tuning* basé sur les prompts.

2.1 Extraction de relations d'évènements et identification causale

L'ERE englobe l'identification des relations de coréférence, temporelles, causales et de sous-évènements entre les mentions d'évènements (Wang *et al.*, 2022). Les approches traditionnelles reposent sur des classificateurs supervisés ou des modèles de graphes, mais souffrent de la rareté des données et des dépendances à longue distance (Guan *et al.*, 2024). Des benchmarks comme MAVEN-ERE (Wang *et al.*, 2022) fournissent des annotations à grande échelle, tandis qu'EventStoryLine (Caselli & Vossen, 2017) et MECI (Lai *et al.*, 2022) se concentrent sur les relations causales et temporelles dans les récits. Pour l'ECI, la tâche se simplifie à la détection de l'existence d'un lien causal, là où l'ERE impose une classification et une orientation explicites.

Les techniques d'encodage utilisent des *encoders* pour extraire des représentations contextuelles riches. mBERT (Devlin *et al.*, 2019) et XLMR (Conneau *et al.*, 2020), variantes multilingues de BERT, peinent avec les relations implicites et restent dépendantes de ressources annotées par langue. KnowMMR (Liu *et al.*, 2020) améliore les représentations via des graphes de connaissances, SemSIn (Hu *et al.*, 2023) via des structures sémantiques centrées sur les évènements, et DiffusECI (Man *et al.*, 2024a) par débruitage itératif des signaux causaux. Ces méthodes atteignent de hautes performances mais requièrent des données d'entraînement annotées.

Les approches fondées sur le raisonnement par graphes modélisent les dépendances évènementielles pour permettre une inférence globale. ERGO (Chen *et al.*, 2022) construit des graphes relationnels pour traiter l'ECI au niveau du document, tandis que CHEER (Chen *et al.*, 2023) exploite un graphe d'interaction où la centralité des évènements (évaluée par leur position et leur connectivité) pondère directement le processus de raisonnement causal. RichGCN (Tran Phu & Nguyen, 2021) exploite, quant à lui, des structures documentaires riches (entités, syntaxe) via des réseaux de convolution sur graphes (GCN). Bien que ces méthodes capturent efficacement les dépendances structurelles, elles restent tributaires d'un processus d'annotation complexe et coûteux.

2.2 Méthodes basées sur les prompts et les LLMs

Le *fine-tuning* basé sur les prompts adapte les LLMs pour l'ECI avec un minimum de données. HOTECI (Man *et al.*, 2024b) utilise le transport optimal pour la sélection du contexte utile, et KEPT (Liu *et al.*, 2023) enrichit les prompts avec des connaissances externes. Ces méthodes s'appuient

toutefois souvent sur des marqueurs lexicaux (ex : « parce que »), ce qui limite leur portabilité à de nouveaux domaines sans adaptation dédiée.

Les méthodes s'appuyant sur les LLMs exploitent des modèles de grande taille en mode *zero-shot* ou *few-shot*. GPT-3.5-turbo et GPT-4o-mini (OpenAI *et al.*, 2024) effectuent des inférences directes mais souffrent d'un rappel excessif et d'une précision faible, révélant une tendance à l'hallucination en l'absence de validation. Les modèles multilingues tels que Meta-MK (Chen *et al.*, 2024) et GIMC (He *et al.*, 2024) permettent de traiter plusieurs langues de façon générale, mais nécessitent toujours une phase d'entraînement supervisé.

2.3 Raisonnement basé sur les agents et les débats

Les cadres agentiques enrichissent les LLMs par l'utilisation d'outils et le raisonnement multi-tours (Wang & Huang, 2024). Le débat multi-agents (MAD) (Liang *et al.*, 2024) favorise la coopération pour réduire les biais. Dans l'ERE, MMD-ERE (Guan *et al.*, 2025) utilise des débats multipartites. Le défi de la complexité quadratique $O(n^2)$ des paires d'évènements dans les longs documents est parfois mitigé par compression ou clustering (Guan *et al.*, 2024), alors que notre approche privilégie une phase de *planning* et des appels d'outils ciblés. Les modèles de diffusion (Man *et al.*, 2024a) opèrent par débruitage itératif des représentations, tandis que DAPrompt (Xiang *et al.*, 2025) contraint mathématiquement l'apprentissage par prompt. Parallèlement, des LLMs en cascade décomposent l'extraction relationnelle en sous-tâches spécialisées (Tan *et al.*, 2025). Face à ces approches génératives, le cadre GenRES (Jiang *et al.*, 2024) souligne l'incapacité des métriques exactes traditionnelles à évaluer correctement des prédictions syntaxiquement différentes mais sémantiquement valides.

Le *Reversal Curse* (Berglund *et al.*, 2023) illustre une limite fondamentale : les LLMs peinent à généraliser une relation dans sa direction inverse. Notre travail s'appuie sur ce constat en intégrant, dans une boucle agentique, deux outils spécifiques : `coherence_check` pour évaluer la cohérence (ici l'asymétrie des relations `EffectCause` et `CauseEffect`), et `counterfactual_pairs` pour imposer une évaluation sémantique plus avancée par le modèle. Cette architecture neuro-symbolique *zero-shot* ne dépend d'aucune donnée d'entraînement, assurant une portabilité multilingue illustrée par les bons résultats obtenus sur les cinq langues du corpus MECI.

3 Méthodologie

Nous introduisons un cadre agentique pour l'ECI reposant sur un agent conversationnel augmenté de deux outils spécialisés pour le raisonnement causal. Le framework vise à extraire des relations causales entre les mentions d'évènements annotées dans un texte (avec la mention entre les symboles « < > » et un id pour chaque mention) et où les relations sont orientées, capturant aussi bien des indices explicites (ex. « parce que ») que des liens implicites fortement suggérés par le contexte, tels que facilitation, prévention, ou dépendance contrefactuelle. L'objectif est d'assigner à chaque paire de mentions d'évènements un label orienté appartenant à un ensemble prédéfini, en gérant les dépendances intra- et inter-phrastiques au sein de documents complets. L'agent invoque selon les besoins deux outils sans état, COHERENCE_CHECK et COUNTERFACTUAL_PAIRS, tout en conservant l'intégralité de la conversation et du contexte du document d'un tour à l'autre (Figure 1). Les modèles sont accessibles via une interface standard de chat-completions (ex. OpenRouter), permettant de

substituer un modèle à un autre sans modifier la logique. Cette section détaille la tâche, les contraintes, les outils, la politique de décision, la boucle de l'agent et la stratégie de rééchantillonnage.

3.1 Description de la tâche

L'agent classifie des relations causales orientées pour un ensemble donné de paires de mentions d'évènements, en utilisant des labels prédéfinis : CauseEffect, EffectCause, ou NoRel. Il opère de manière stateful, préservant l'intégralité de l'historique conversationnel pour maintenir le contexte global, tandis que les outils reçoivent uniquement des entrées minimales (identifiants de paires, fenêtres contextuelles de texte, ou graphes candidats) et renvoient des sorties structurées. Les prédictions initiales sont générées en bloc sur l'ensemble des paires demandées (phase de *seeding*), après quoi des appels d'outils ciblés permettent de valider, corriger et finaliser le graphe de relations. Pour assurer l'observabilité et la reproductibilité, tous les appels (modèle et outils) sont tracés, capturant la latence, les entrées/sorties et l'arborescence des sous-tâches.

3.2 Contraintes

Dans MECI, les liens causaux positifs sont encodés de façon bidirectionnelle : toute relation CauseEffect de A vers B implique une relation EffectCause de B vers A, et réciproquement. Nous encodons cette propriété comme une contrainte de dualité directionnelle : les labels CauseEffect et EffectCause sont les relations converses l'un de l'autre, et toute prédiction doit respecter cette dualité pour être cohérente. Concrètement, dès que l'agent propose une arête orientée positive pour une paire (A, B), l'outil COHERENCE_CHECK vérifie que l'arête inverse (B, A) porte le label complémentaire attendu. Tout manquement, tel qu'une arête inverse absente ou incorrectement labellisée, est signalé comme une violation devant être réparée, par ajout, inversion ou retrait du label si les preuves sont insuffisantes. La cohérence globale du graphe est prioritaire sur son exhaustivité : une paire incohérente est préférablement ramenée à NoRel/NoRel plutôt que conservée avec une directionnalité non étayée.

3.3 Contrats d'outils et politiques d'utilisation

L'outil COUNTERFACTUAL_PAIRS exécute un test de substitution (*but-for*) pour une paire (T_i, T_j). Il exploite un sous-appel LLM pour évaluer si T_j aurait eu lieu sans T_i, forçant ainsi le modèle à explorer une sémantique orthogonale au *seeding* initial. Ce test opère au niveau sémantique des évènements, garantissant une portabilité aux cinq langues de MECI sans adaptation linguistique. Il retourne un objet JSON (label, score, justification), le label `NoRel` étant affecté par défaut en cas d'échec de parsing.

L'outil COHERENCE_CHECK valide la dualité directionnelle du graphe prédit. Il prend en entrée une liste de paires labellisées, traite les paires absentes comme NoRel, et renvoie : le nombre de paires considérées, le nombre de paires symétriquement valides, les conflits directionnels, les inverses manquants, le taux de cohérence global, ainsi que les ajouts ou corrections suggérés. Il normalise les labels de manière insensible à la casse et opère sur des prédictions partielles sans nécessiter l'ensemble complet des paires.

Par convention encodée dans le prompt de l'agent, COHERENCE_CHECK est obligatoirement invoqué

après chaque complétion ou édition substantielle du graphe. COUNTERFACTUAL_PAIRS est invoqué en priorité pour les paires inter-phrastiques ou à longue distance, les liens sans marqueur explicite, et les paires tentativement positives nécessitant une confirmation directionnelle.

3.4 Politique de décision (intégrée dans les prompts)

Les prompts intègrent une directive stricte (*precision-first doctrine*) avec des critères clairs justifiant un test « et si » ("*what if*"). Par défaut, l'agent affectera « NoRel » (Neutre/Négatif) sauf dans le cas de preuves textuelles directionnelles écrasantes. Une règle dite des « deux signaux » exige pour tout lien non formel un minimum de deux preuves : asymétrie contrefactuelle, mécanique décrite dans le texte, marqueurs robustes, ou indices contextuels. Une autre règle favorise toujours la sélection de la « cause la plus proche », évitant de sauter des intermédiaires flagrants. En cas d'asymétrie directionnelle non résolue, le label NoRel est privilégié. Les prompts complets sont fournis en annexe B.

3.5 Boucle de l'agent

L'agent suit une boucle exécutive itérative et autonome, dont l'architecture générale est illustrée par la Figure 1. Plutôt que de s'appuyer sur des étapes prédéfinies ou rigides, le modèle de langage orchestre librement de son propre raisonnement : il décide d'invoquer l'outil COUNTERFACTUAL_PAIRS (pour évaluer la dépendance sémantique d'arêtes incertaines) ou l'outil COHERENCE_CHECK (pour repérer et corriger d'éventuels conflits directionnels) quand il l'estime nécessaire et dans l'ordre de son choix. Ce cycle d'interaction avec les outils se poursuit tant que l'agent en ressent le besoin pour affiner ses prédictions, et s'arrête de lui-même lorsqu'il décide que le graphe est achevé. Une limite de sécurité matérielle, plafonnée à 100 appels d'outils par document, permet d'éviter les boucles infinies et d'en maîtriser le coût. Le processus s'achève par la génération d'un objet JSON final.

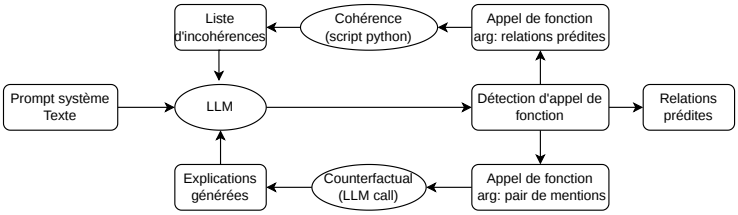

FIGURE 1 – Architecture de la boucle agentique

La complexité inhérente de la tâche est quadratique en le nombre d'évènements — $O(n^2)$ paires à classifier pour n évènements dans un document. La phase de planning limite les appels d'outils aux paires les plus incertaines, tandis que le seeding en bloc réduit la charge de tokens par rapport à une classification séquentielle paire-à-paire. Un cap de 100 appels d'outils par document constitue une borne opérationnelle contrôlant le coût computationnel, complété par jusqu'à quatre redémarrages complets en cas de défaut d'intégrité du parsing.

3.6 Re-échantillonnage

Comme détaillé dans la Figure 2, nous générons trois runs indépendants hébergés sur des threads découplés, agrégés par vote majoritaire par paire. En cas d'absence de relation majoritaire, le label NoRel est retenu pour privilégier la précision et éviter les faux positifs. Le traçage des runs via LangSmith permet de quantifier le taux d'accord inter-runs et d'évaluer la nature des erreurs observées : une faible variance inter-runs associée à des erreurs persistantes indique un biais systématique stable, non réductible par rééchantillonnage supplémentaire et nécessitant des interventions ciblées au niveau des prompts ou des outils. De plus le rééchantillonnage permet une amélioration des performances et de mieux évaluer le modèle en s'affranchissant de l'aspect stochastique. Enfin, certains modèles ne proposent plus d'argument de température.

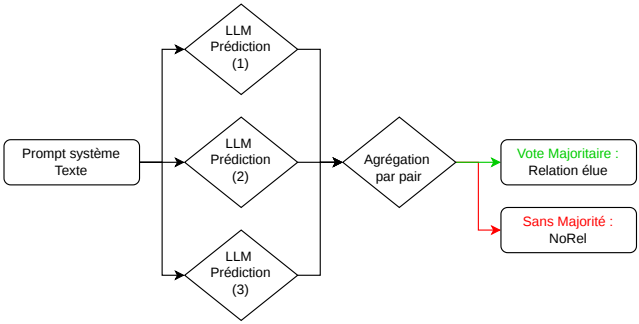

FIGURE 2 – Illustration de la stratégie de rééchantillonnage

4 Expérimentation

4.1 Jeu de données

MECI (Lai *et al.*, 2022) est un benchmark multilingue pour l'ECI évènement-à-évènement, construit à partir d'articles Wikipédia en cinq langues : anglais, danois, espagnol, turc et ourdou. Les évènements sont représentés par des spans de déclencheurs, et les systèmes doivent assigner un label causal orienté à chaque paire de mentions demandée. Le partitionnement utilisé provient de la diffusion publique [1] version 0.1, et les évaluations sont conduites sur l'intégralité du split de test.

4.2 Protocole d'évaluation

Le micro-F1 est retenu comme métrique principale, conformément à la pratique standard sur MECI, pour sa robustesse face à la variation du nombre de relations par document et au fort déséquilibre entre classes. Les instances NoRel sont exclues du calcul du micro-F1 afin de concentrer l'évaluation

1. https://github.com/nlp-uoregon/meci-dataset

sur les identifications causales positives. La précision, le rappel et le F1 sont rapportés en traitant les labels orientés comme des arêtes dirigées.

Les relations EffectCause et CauseEffect sont traitées séparément. Une extraction complète à l'aide de la relation EffectCause par exemple réduit à 'Causal' la relation entre une mention A et une mention B s'il existe une prédiction EffectCause entre les deux quelle que soit sa direction. Ceci afin de se comparer à des méthodes ECI.

Afin d'évaluer la contribution, nous conduisons des études d'ablation comparant : (i) le modèle seul sans outils (*Baseline*), (ii) le modèle avec outils activés (COUNTERFACTUAL_PAIRS + COHERENCE_CHECK) sans rééchantillonnage (*Outils*), et (iii) le modèle avec outils et rééchantillonnage (*Outils + Rééchantillonnage*). Chaque configuration est évaluée dans les deux orientations de prédiction (CauseEffect et EffectCause). Le traçage LangSmith capture le comportement par condition et facilite l'analyse d'erreurs détaillée.

Les ressources logicielles complètes pour reproduire nos expériences sont accessibles depuis notre dépôt Github[2].

4.3 Résultats

Les résultats sont présentés dans le Tableau 1. Nos configurations avec outils atteignent des performances compétitives avec les méthodes supervisées sur l'ensemble des cinq langues, et surpassent systématiquement les LLMs non augmentés (GPT-3.5-turbo, GPT-4o-mini). La configuration *Outils + Rééchantillonnage / EffectCause* obtient le meilleur F1 en danois (68.8), en espagnol (67.9), et en ourdou (67.4), et le deuxième meilleur en anglais (76.4) et en turc (71.2), sans recourir à aucune donnée d'entraînement — contrairement à DiffusECI, HOTECI et aux modèles à `fine-tuning`.

L'activation des outils améliore systématiquement l'équilibre précision/rappel par rapport aux baselines. Le rééchantillonnage apporte un gain marginal supplémentaire en précision.

4.4 Analyse

Asymétrie directionnelle et biais systématique. La configuration EffectCause surpasse systématiquement CauseEffect sur toutes les langues et toutes les conditions, avec un écart moyen de 2 à 6 points de F1. Cet écart persiste malgré COHERENCE_CHECK qui ne fait que rapporter au LLM les incohérences dans la prédiction, mais ne le contraint pas à adopter une prédiction finale cohérente. Cet écart constitue une manifestation directe du *Reversal Curse* (Berglund *et al.*, 2023) : les LLMs, entraînés sur des patterns directionnels dominants dans leurs données d'entraînement, produisent des prédictions asymétriques stables indépendamment de la formulation de la tâche.

Nature systématique des erreurs. La persistance de l'écart CauseEffect/EffectCause malgré le rééchantillonnage indique un biais structurel stable et non un artefact du non-déterminisme des LLM. La faible variation inter-runs mesurée indique que ces erreurs sont systématiques et ne sont pas résolues par une simple augmentation du nombre d'essais. Ce constat valide la nécessité d'interventions ciblées via des prompts ou des outils.

2. https://github.com/Alasdey/agenteci.git

Type	Modèle	Anglais			Danois			Espagnol			Turc			Ourdou		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Encodage Sémantique Profond	mBERT	38.4	46.0	41.9	25.2	26.6	25.9	43.9	41.5	42.7	36.2	48.7	41.6	31.9	34.3	33.0
	XLMR	48.7	59.9	53.7	35.9	36.2	36.0	50.6	49.1	49.9	44.0	59.4	50.5	40.4	43.2	41.8
	KnowMMR	42.1	45.2	43.6	43.2	32.5	37.1	39.2	49.8	43.9	34.1	14.9	20.7	44.6	25.8	32.7
	SemSIn*	56.6	69.1	62.2	57.9	58.6	58.2	54.9	65.8	59.9	37.5	49.3	42.6	50.8	59.4	54.8
	GCKAN*	52.8	70.5	60.4	53.5	47.3	50.2	53.8	55.4	54.6	43.2	45.8	44.5	56.1	45.9	50.5
	DiffusECI	70.1	68.3	69.2	42.7	53.3	47.4	62.9	50.2	55.8	52.6	66.5	58.7	58.1	52.5	55.2
Raisonnement sur Graphe d'Évènements	ERGO*	55.1	73.8	63.1	56.9	57.8	57.3	56.8	65.9	61.0	35.5	43.2	39.0	46.9	55.6	50.9
	CHEER*	59.9	73.9	66.2	61.3	62.6	61.9	60.2	68.8	64.2	42.9	53.2	47.5	53.8	62.3	57.7
	RichGCN	53.6	71.8	61.4	53.6	57.2	55.3	54.8	67.2	60.4	36.8	53.7	43.7	41.0	60.5	48.9
Fine-tuning basé sur Prompts	HOTECI	66.6	67.1	66.8	50.5	63.7	56.3	60.7	60.7	60.7	72.5	**76.6**	**74.5**	59.1	71.0	64.5
	KEPT*	49.5	72.6	58.9	49.3	44.6	46.8	51.2	53.8	52.5	41.6	43.2	42.4	52.8	45.1	48.6
Basé sur les LLMs	GPT-3.5-turbo	24.6	**79.4**	37.6	10.0	66.5	17.4	7.3	74.2	13.3	27.0	69.3	38.8	15.7	63.8	25.2
	GPT-4o-mini*	22.3	75.9	33.9	11.3	65.4	18.7	8.6	69.7	12.5	29.7	71.6	40.5	16.2	66.0	27.9
Modèles Multilingues	Meta-MK	55.3	71.4	62.3	36.6	47.8	41.5	57.7	61.0	59.3	58.1	66.7	62.1	39.2	61.5	47.9
	GIMC	63.4	54.8	58.8	60.2	45.2	51.6	**77.5**	55.7	64.8	70.1	60.1	64.7	62.1	42.4	50.4
Baseline	CauseEffect	71.6	71.5	71.6	55.2	67.3	60.6	44.7	72.7	55.3	78.4	55.1	64.7	62.4	66.7	64.5
	EffectCause	84.7	71.5	**77.6**	64.4	66.5	65.4	50.3	72.3	59.3	**84.1**	54.9	66.4	**66.7**	66.7	66.7
Outils	CauseEffect	70.3	72.6	71.4	63.4	**68.1**	65.7	47.9	**76.3**	58.9	74.8	64.3	69.1	59.4	**73.5**	65.7
	EffectCause	77.4	72.6	74.9	69.9	67.7	**68.8**	52.8	**76.3**	62.4	76.7	64.1	69.8	62.2	**73.5**	**67.4**
Outils + Rééchantillonnage	CauseEffect	77.5	68.6	72.8	65.9	66.5	66.2	56.7	75.9	64.9	81.8	62.3	70.7	59.6	72.5	65.4
	EffectCause	**86.2**	68.6	76.4	**71.3**	66.5	**68.8**	61.3	75.9	**67.9**	83.1	62.3	71.2	63.0	72.5	**67.4**

TABLE 1 – Précision (P), Rappel (R) et F1 pour les cinq langues de MECI. Le texte en gras indique le meilleur résultat pour chaque langue, celui surligné le deuxième meilleur. Le tableau, à l'exception de nos résultats, est issu de Cheng *et al.* (2026).

Contribution des outils. L'ablation confirme l'apport spécifique de chaque composant. Sans outils, les baselines souffrent d'un rappel élevé mais d'une précision dégradée par les hallucinations. L'outil COHERENCE_CHECK améliore la précision en filtrant les incohérences directionnelles, tandis que COUNTERFACTUAL_PAIRS renforce le F1 sur les cas complexes (liens implicites ou distants) par une évaluation sémantique orthogonale. Leur combinaison offre le meilleur compromis précision/rappel, alliant contrainte globale de cohérence et raffinement local des paires incertaines.

5 Limites et perspectives

L'évaluation présentée dans ce travail est restreinte au benchmark MECI, ce qui limite la généralisation du framework à d'autres schémas d'annotation ou domaines textuels. Bien que le registre de règles symboliques soit conçu pour être adaptable à de nouveaux jeux de données, cette adaptation reste manuelle et nécessite une connaissance préalable du schéma cible.

Plusieurs directions de recherche permettraient de dépasser ces limites. Une première consiste à étendre l'évaluation à des benchmarks additionnels tels que MAVEN-ERE (Wang *et al.*, 2022), EventStoryLine (Caselli & Vossen, 2017) et Causal-TimeBank, afin de tester la robustesse du framework dans des contextes d'annotation différents et d'évaluer sa portabilité à d'autres domaines.

Par ailleurs, les biais systématiques observés, tels que l'asymétrie directionnelle persistante, semblent en partie ancrés dans le modèle sous-jacent et ne peuvent être qu'imparfaitement atténués par les outils proposés. L'ambiguïté inhérente à la notion de causalité entre benchmarks (par exemple facilitation, empêchement ou simple corrélation temporelle) introduit également un bruit conceptuel que ni les prompts ni les outils ne peuvent entièrement résoudre.

Enfin, l'ensemble des expériences repose sur un unique backbone (`gpt-5-mini`). Une comparaison systématique avec d'autres LLMs, incluant des modèles open-source ou des architectures orientées raisonnement explicite, permettrait d'évaluer dans quelle mesure les gains observés sont spécifiques au modèle utilisé ou généralisables à d'autres architectures.

6 Conclusion

Nous avons présenté un cadre agentique neuro-symbolique *zero-shot* augmentant un LLM de deux outils spécialisés pour le raisonnement causal. Évalué sur les cinq langues de MECI, notre système atteint des performances compétitives avec les méthodes supervisées sans aucune donnée d'entraînement, et surpasse systématiquement les LLMs non augmentés. Le rééchantillonnage révèle que l'asymétrie directionnelle persistante entre relations constitue un biais structurel ancré dans le modèle sous-jacent et non un artefact stochastique. Ce constat appelle des interventions ciblées au niveau des prompts ou des outils et ouvre des perspectives vers l'évaluation sur des benchmarks complémentaires et la comparaison multi-modèles.

Références

BERGLUND L., TONG M., KAUFMANN M., BALESNI M., STICKLAND A. C., KORBAK T. &

EVANS O. (2023). The Reversal Curse : LLMs trained on “A is B” fail to learn “B is A”. In *The Twelfth International Conference on Learning Representations*.

CAI R., YU S., ZHANG J., CHEN W., XU B. & ZHANG K. (2025). Dr.ECI : Infusing Large Language Models with Causal Knowledge for Decomposed Reasoning in Event Causality Identification. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference on Computational Linguistics*, p. 9346–9375, Abu Dhabi, UAE : Association for Computational Linguistics.

CASELLI T. & VOSSEN P. (2017). The Event StoryLine Corpus : A New Benchmark for Causal and Temporal Relation Extraction. In T. CASELLI, B. MILLER, M. VAN ERP, P. VOSSEN, M. PALMER, E. HOVY, T. MITAMURA & D. CASWELL, Éds., *Proceedings of the Events and Stories in the News Workshop*, p. 77–86, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2711.

CHEN M., CAO Y., DENG K., LI M., WANG K., SHAO J. & ZHANG Y. (2022). ERGO : Event Relational Graph Transformer for Document-level Event Causality Identification. In N. CALZOLARI, C.-R. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K.-S. CHOI, P.-M. RYU, H.-H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S.-H. NA, Éds., *Proceedings of the 29th International Conference on Computational Linguistics*, p. 2118–2128, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

CHEN M., CAO Y., ZHANG Y. & LIU Z. (2023). CHEER : Centrality-aware High-order Event Reasoning Network for Document-level Event Causality Identification. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 10804–10816, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.604.

CHEN M., MA Y., SONG K., CAO Y., ZHANG Y. & LI D. (2024). Improving Large Language Models in Event Relation Logical Prediction. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 9451–9478, Bangkok, Thailand : Association for Computational Linguistics.

CHENG Q., ZENG Z., HU X., SI Y. & LIU Z. (2026). A Survey of Event Causality Identification : Taxonomy, Challenges, Assessment, and Prospects. *ACM Computing Surveys*, **58**(3), 1–37. DOI : 10.1145/3756009.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GUAN Y., PENG H., HOU L. & LI J. (2025). MMD-ERE : Multi-Agent Multi-Sided Debate for Event Relation Extraction. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference*

on Computational Linguistics, p. 6889–6896, Abu Dhabi, UAE : Association for Computational Linguistics.

GUAN Y., WANG X., HOU L., LI J., PAN J. Z., CHEN J. & LECUE F. (2024). TacoERE : Cluster-aware Compression for Event Relation Extraction. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 15511–15521, Torino, Italia : ELRA and ICCL.

HE Z., CAO P., JIN Z., CHEN Y., LIU K., ZHANG Z., SUN M. & ZHAO J. (2024). Zero-Shot Cross-Lingual Document-Level Event Causality Identification with Heterogeneous Graph Contrastive Transfer Learning.

HU Z., LI Z., JIN X., BAI L., GUAN S., GUO J. & CHENG X. (2023). Semantic Structure Enhanced Event Causality Identification.

JIANG P., LIN J., WANG Z., SUN J. & HAN J. (2024). GenRES : Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2820–2837, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.155.

LAI V. D., VEYSEH A. P. B., NGUYEN M. L., DERNONCOURT F. & NGUYEN T. H. (2022). MECI : A Multilingual Dataset for Event Causality Identification. In *International Conference on Computational Linguistics*.

LIANG T., HE Z., JIAO W., WANG X., WANG Y., WANG R., YANG Y., SHI S. & TU Z. (2024). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 17889–17904, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.992.

LIU J., CHEN Y. & ZHAO J. (2020). Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, p. 3608–3614, Yokohama, Japan : International Joint Conferences on Artificial Intelligence Organization. DOI : 10.24963/ijcai.2020/499.

LIU J., ZHANG Z., GUO Z., JIN L., LI X., WEI K. & SUN X. (2023). KEPT : Knowledge Enhanced Prompt Tuning for event causality identification. *Knowledge-Based Systems*, **259**, 110064. DOI : 10.1016/j.knosys.2022.110064.

MAN H., DERNONCOURT F. & NGUYEN T. H. (2024a). Mastering Context-to-Label Representation Transformation for Event Causality Identification with Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**(17), 18760–18768. DOI : 10.1609/aaai.v38i17.29840.

MAN H., NGUYEN C. V., NGO N. T., NGO L., DERNONCOURT F. & NGUYEN T. H. (2024b). Hierarchical Selection of Important Context for Generative Event Causality Identification with Optimal Transports. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 8122–8132, Torino, Italia : ELRA and ICCL.

OPENAI, ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S., AVILA R., BABUSCHKIN I., BALAJI S., BALCOM V., BALTESCU P., BAO H., BAVARIAN M., BELGUM J., BELLO I., BERDINE J., BERNADETT-SHAPIRO G., BERNER C., BOGDONOFF L., BOIKO O., BOYD M., BRAKMAN

A.-L., Brockman G., Brooks T., Brundage M., Button K., Cai T., Campbell R., Cann A., Carey B., Carlson C., Carmichael R., Chan B., Chang C., Chantzis F., Chen D., Chen S., Chen R., Chen J., Chen M., Chess B., Cho C., Chu C., Chung H. W., Cummings D., Currier J., Dai Y., Decareaux C., Degry T., Deutsch N., Deville D., Dhar A., Dohan D., Dowling S., Dunning S., Ecoffet A., Eleti A., Eloundou T., Farhi D., Fedus L., Felix N., Fishman S. P., Forte J., Fulford I., Gao L., Georges E., Gibson C., Goel V., Gogineni T., Goh G., Gontijo-Lopes R., Gordon J., Grafstein M., Gray S., Greene R., Gross J., Gu S. S., Guo Y., Hallacy C., Han J., Harris J., He Y., Heaton M., Heidecke J., Hesse C., Hickey A., Hickey W., Hoeschele P., Houghton B., Hsu K., Hu S., Hu X., Huizinga J., Jain S., Jain S., Jang J., Jiang A., Jiang R., Jin H., Jin D., Jomoto S., Jonn B., Jun H., Kaftan T., Kaiser Ł., Kamali A., Kanitscheider I., Keskar N. S., Khan T., Kilpatrick L., Kim J. W., Kim C., Kim Y., Kirchner J. H., Kiros J., Knight M., Kokotajlo D., Kondraciuk Ł., Kondrich A., Konstantinidis A., Kosic K., Krueger G., Kuo V., Lampe M., Lan I., Lee T., Leike J., Leung J., Levy D., Li C. M., Lim R., Lin M., Lin S., Litwin M., Lopez T., Lowe R., Lue P., Makanju A., Malfacini K., Manning S., Markov T., Markovski Y., Martin B., Mayer K., Mayne A., McGrew B., McKinney S. M., McLeavey C., McMillan P., McNeil J., Medina D., Mehta A., Menick J., Metz L., Mishchenko A., Mishkin P., Monaco V., Morikawa E., Mossing D., Mu T., Murati M., Murk O., Mély D., Nair A., Nakano R., Nayak R., Neelakantan A., Ngo R., Noh H., Ouyang L., O'Keefe C., Pachocki J., Paino A., Palermo J., Pantuliano A., Parascandolo G., Parish J., Parparita E., Passos A., Pavlov M., Peng A., Perelman A., Peres F. d. A. B., Petrov M., Pinto H. P. d. O., Michael, Pokorny, Pokrass M., Pong V. H., Powell T., Power A., Power B., Proehl E., Puri R., Radford A., Rae J., Ramesh A., Raymond C., Real F., Rimbach K., Ross C., Rotsted B., Roussez H., Ryder N., Saltarelli M., Sanders T., Santurkar S., Sastry G., Schmidt H., Schnurr D., Schulman J., Selsam D., Sheppard K., Sherbakov T., Shieh J., Shoker S., Shyam P., Sidor S., Sigler E., Simens M., Sitkin J., Slama K., Sohl I., Sokolowsky B., Song Y., Staudacher N., Such F. P., Summers N., Sutskever I., Tang J., Tezak N., Thompson M. B., Tillet P., Tootoonchian A., Tseng E., Tuggle P., Turley N., Tworek J., Uribe J. F. C., Vallone A., Vijayvergiya A., Voss C., Wainwright C., Wang J. J., Wang A., Wang B., Ward J., Wei J., Weinmann C. J., Welihinda A., Welinder P., Weng J., Weng L., Wiethoff M., Willner D., Winter C., Wolrich S., Wong H., Workman L., Wu S., Wu J., Wu M., Xiao K., Xu T., Yoo S., Yu K., Yuan Q., Zaremba W., Zellers R., Zhang C., Zhang M., Zhao S., Zheng T., Zhuang J., Zhuk W. & Zoph B. (2024). GPT-4 Technical Report. DOI : 10.48550/arXiv.2303.08774.

Tan X., Zhou Y., Pergola G. & He Y. (2025). Cascading Large Language Models for Salient Event Graph Generation. In L. Chiruzzo, A. Ritter & L. Wang, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2223–2245, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.112.

Tran Phu M. & Nguyen T. H. (2021). Graph Convolutional Networks for Event Causality Identification with Rich Document-level Structures. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3480–3490, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.273.

WANG S. & HUANG L. (2024). Debate as Optimization : Adaptive Conformal Prediction and Diverse Retrieval for Event Extraction. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 16422–16435, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.958.

WANG X., CHEN Y., DING N., PENG H., WANG Z., LIN Y., HAN X., HOU L., LI J., LIU Z., LI P. & ZHOU J. (2022). MAVEN-ERE : A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 926–941, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.60.

XIANG W., ZHAN C., ZHANG Q. & WANG B. (2025). DAPrompt : Deterministic assumption prompt learning for event causality identification. *Neural Computing and Applications*, **37**(26), 21743–21759. DOI : 10.1007/s00521-025-11486-x.

A Détails d'implémentation

L'accès aux modèles est géré via OpenRouter, les outils et le graphe agentique sont implémentés avec LangChain/LangGraph, et le traçage est assuré par LangSmith. Le backbone utilisé est `gpt-5-mini`, sélectionné après des essais préliminaires avec `gpt-5` qui montraient des gains marginaux et des baisses occasionnelles de performances. La boucle agentique est réalisée comme un graphe LangGraph conditionnel à deux nœuds avec un saut d'outil (*tool hop*) : l'agent maintient le contexte complet tandis que les outils ne reçoivent que leurs entrées minimales, les appels d'outils étant exécutés par un `ToolNode` renvoyant des objets `ToolMessage` à l'état.

B Prompts

B.1 Prompts sans outils

Prompt système – contexte MECI

Context
MECI (Multilingual Event Causality Identification) follows ACE-style event mentions and EventStoryLine causal guidelines. It addresses cause, enable, prevent, and covers both explicit and implicit causal links – even across sentences. Data is drawn from Wikipedia articles in EN/DA/ES/TR/UR, with possibly long-range links (10–50 tokens apart). [Reference : COLING'22 MECI paper.]

Label Set (Use these exact strings)
- "CauseEffect" : Ti causes/enables/leads to or prevents the absence of Tj (Ti -> Tj).
- "EffectCause" : Tj causes/enables/leads to or prevents the absence of Ti (Tj -> Ti).
- "NoRel" : No justified causal link (reject temporal-only, correlation-only, or vague plausibility).

Precision-First Rules (Apply Silently)
1. Two-Signal Rule for any non-overt link (no clear connective like “because/so/therefore/lead to/due to”) : require at least two independent signals among : (a) counterfactual asymmetry (but-for), (b) a concrete mechanism in the text, (c) strong discourse evidence, (d) local lexical patterns of enablement/prevention. If fewer than two signals -> NoRel.
2. Distance/Scope Gate : For inter-sentential pairs or long-distance pairs, must include an explicit cue or spelled-out mechanism. Otherwise NoRel.
3. Default to NoRel when uncertain : ambiguous or correlation-only evidence -> NoRel.
4. Confounder & Chain Check : If another event more directly explains Tj, label Ti,Tj as NoRel.
5. Reporting/Attribution is Not Causal : “said”, “reported”, “claimed” are not causes unless the content event is enabled/prevented by the speech act itself.
6. Negation/Contrast is Not Causality : “although”, “however”, “despite” do not justify causal labels unless rules above are satisfied.
7. Direction Test : If neither direction is necessary, or both seem plausible, -> NoRel.

Decision Checklist (Apply Silently)
1. Does Ti materially change the likelihood or state of Tj in the text ?
2. Is there two-signal support (or explicit cue for long-range links) ?
3. Reject temporal-only, correlation-only, reporting, and contrastive links.
4. Prefer the nearest, most specific cause.
5. If a more direct cause is present, label NoRel for the current pair.

Output Format
- Follow the provided JSON structure exactly.
- Use only the approved label strings for causal relations.

Gabarit utilisateur

Rules :
- Labels : "CauseEffect" | "EffectCause" | "NoRel".
- Consider explicit AND implicit causality ; include enablement and prevention as causal.
- Ensure EVERY requested pair appears exactly once in your final JSON (any order).
- Return ONLY a JSON array like : `[{"pair":"T0,T1","label":"CauseEffect"}]`

Coherence Rules :
- Symmetry : "Ti,Tj" "CauseEffect" implies "Tj,Ti" "EffectCause" and inversely.
- Missing : Any relation left missing will be considered NoRel.
- "Tj,Ti" "NoRel" is not compatible with "Ti,Tj" "CauseEffect" or "Ti,Tj" "EffectCause".

Text : `{doc_text}`

Pairs to classify (use EXACT pair ids ; output order does NOT matter) :
`{pair_lines}`

B.2 Avec outils

Prompt système

Context
MECI (Multilingual Event Causality Identification) follows ACE-style event mentions and EventStoryLine causal guidelines. It addresses cause, enable, prevent, and covers both explicit and implicit causal links – even across sentences. Data is drawn from Wikipedia articles in EN/DA/ES/TR/UR, with possibly long-range links (10–50 tokens apart). [Reference : COLING'22 MECI paper.]

Label Set (Use these exact strings)
- "CauseEffect" : Ti causes/enables/leads to or prevents the absence of Tj (Ti -> Tj).
- "EffectCause" : Tj causes/enables/leads to or prevents the absence of Ti (Tj -> Ti).
- "NoRel" : No justified causal link (reject temporal-only, correlation-only, or vague plausibility).

Precision-First Rules (Apply Silently)
1. Two-Signal Rule for any non-overt link (no clear connective like "because/so/therefore/lead to/due to") : require at least two independent signals among : (a) counterfactual asymmetry (but-for), (b) a concrete mechanism in the text, (c) strong discourse evidence, (d) local lexical patterns of enablement/prevention. If fewer than two signals -> NoRel.
2. Distance/Scope Gate : For inter-sentential pairs or long-distance pairs, must pass counterfactual AND include an explicit cue or spelled-out mechanism. Otherwise NoRel.
3. Default to NoRel when uncertain : ambiguous or correlation-only evidence -> NoRel.
4. Confounder & Chain Check : If another event more directly explains Tj, label Ti,Tj as NoRel.
5. Reporting/Attribution is Not Causal : "said", "reported", "claimed" are not causes unless the content event is enabled/prevented by the speech act itself.
6. Negation/Contrast is Not Causality : "although", "however", "despite" do not justify causal labels unless rules above are satisfied.
7. Direction Test : Use counterfactuals to establish directionality. If neither direction is necessary -> NoRel.

Decision Checklist (Apply Silently)
1. Does Ti materially change the likelihood or state of Tj in the text ?
2. Is there two-signal support (or explicit cue + counterfactual for long-range links) ?
3. Reject temporal-only, correlation-only, reporting, and contrastive links.
4. Prefer the nearest, most specific cause.
5. If a more direct cause is present, label NoRel for the current pair.

Output Format
- Follow the provided JSON structure exactly.
- Use only the approved label strings for causal relations.

Tool Guidance
- Use `coherence_check` after any complete prediction of all pairs (mandatory).
- Use `counterfactual_pairs` for any inter-sentential pair, any pair lacking an explicit causal connective, or any tentatively positive pair (mandatory).
- If `counterfactual_pairs` returns unclear for either direction -> Prefer NoRel.

Gabarit utilisateur

Rules :
- Labels : "CauseEffect" | "EffectCause" | "NoRel".
- Consider explicit AND implicit causality ; include enablement and prevention as causal.
- Ensure EVERY requested pair appears exactly once in your final JSON (any order).
- Return ONLY a JSON array like : `[{"pair":"T0,T1","label":"CauseEffect"}]`

Coherence Rules :
- Symmetry : "Ti,Tj" "CauseEffect" implies "Tj,Ti" "EffectCause" and inversely.
- Missing : Any relation left missing will be considered NoRel.
- "Tj,Ti" "NoRel" is not compatible with "Ti,Tj" "CauseEffect" or "Ti,Tj" "EffectCause".

Text : `{doc_text}`

Pairs to classify (use EXACT pair ids ; output order does NOT matter) :
`{pair_lines}`