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

Impact de l'affinage de modèles génératifs pour l'inférence en langue naturelle appliquée aux essais cliniques : comparaison avec des approches de *few-shot learning*

Lounès Kebdi * Lubin Longuépée* Mathilde Aguiar Pierre Zweigenbaum
Nona Naderi
Université-Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
prenom.nom@universite-paris-saclay.fr

RÉSUMÉ

Les grands modèles de langue (LLM) ont atteint des résultats compétitifs dans de nombreuses applications, y compris dans des domaines de spécialité tels que le biomédical. Dans les essais cliniques, l'inférence en langue naturelle (ILN) permet de modéliser certaines tâches telles que l'appariement des patients aux essais cliniques. Dans cette étude, nous comparons l'affinage et l'apprentissage en contexte avec peu d'exemples afin d'améliorer les performances des LLM pour l'ILN appliquée aux essais cliniques. Nous utilisons les jeux de données NLI4CT et NLI4PR portant sur l'ILN dans le domaine clinique, tous deux en anglais. Nos résultats démontrent que l'affinage des LLM surpasse les autres approches pour les deux jeux de données. Cependant, la différence de performance entre affinage et apprentissage avec peu d'exemples reste parfois faible, en particulier lorsque l'on optimise la sélection des exemples.

ABSTRACT

Impact of fine-tuning generative models for natural language inference for clinical trials : comparison with few-shot learning approaches

Large language models (LLMs) have demonstrated competitive results in many applications, including the biomedical domain. In clinical trials, natural language inference (NLI) can be used to model tasks, such as patient-to-clinical-trial matching. In this study, we compare fine-tuning and few-shot in-context learning to improve LLM performance for NLI in clinical trials. We use the NLI4CT and NLI4PR datasets, which focus on NLI in the clinical domain. Our results show that fine-tuned LLMs outperform the other approaches for both datasets. However, the performance gap between fine-tuning and few-shot learning sometimes remains small, especially when using an optimised demonstration selection.

MOTS-CLÉS : Inférence en langue naturelle, grands modèles de langue, essais cliniques, affinage, apprentissage avec peu d'exemples.

KEYWORDS: Natural language inference, Large Language Models, clinical trials, fine-tuning, few-shot learning.

∗. Ces auteurs ont contribué à parts égales à ce travail.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 303–321.

1 Introduction

Les essais cliniques sont essentiels au développement de nouveaux traitements. Cependant, ils nécessitent un processus long et fastidieux, pouvant durer de quelques années à une décennie [1]. Sertkaya *et al.* (2016) estiment le coût moyen d'un essai entre 1,4 et 52,9 millions de dollars américains. Plusieurs études ont proposé d'utiliser des modèles pré-entraînés pour faciliter certaines tâches comme vérifier qu'un patient donné est éligible à un essai clinique (Nievas *et al.*, 2024; Jin *et al.*, 2024; Wornow *et al.*, 2025). Cette tâche consiste à s'assurer que les caractéristiques du patient sont en accord avec les critères d'inclusion et d'exclusion de l'essai clinique. Si tel est le cas, l'essai considéré peut être proposé au patient. La tâche d'inférence en langue naturelle (ILN) (Dagan *et al.*, 2006) consiste à déterminer, à partir d'une prémisse donnée, ici le rapport d'essai clinique, s'il est possible de déduire l'affirmation ou l'hypothèse considérée. Le modèle doit alors prédire s'il existe une implication ou une contradiction. NLI4CT (Jullien *et al.*, 2023a) et NLI4PR (Aguiar *et al.*, 2025) proposent d'utiliser l'ILN pour l'appariement de patients aux essais cliniques, l'analyse de résultats ou encore la compréhension des interventions menées. Ces tâches restent difficiles en raison du nombre d'interactions entre les différents types de connaissances qu'elles requièrent. Lors des précédentes campagnes d'évaluation NLI4CT (Jullien *et al.*, 2023b, 2024), les publications sur les grands modèles de langue (LLM) ont démontré une performance compétitive (jusqu'à 80 % de macro F1) mais n'ont pas apporté de comparaison claire entre les approches basées sur l'affinage des LLM et différentes techniques d'amorces (*prompts*). Dans cet article, nous comparons, optimisons et évaluons l'impact de ces deux approches en termes de performance et de coût computationnel.

Nos contributions sont : (1) l'affinage paramétrique des modèles Qwen2.5-7B-Instruct (Yang *et al.*, 2024) et Phi-4 (Abdin *et al.*, 2024) évalués via des amorces directes (*zero-shot*) ; (2) l'évaluation de stratégies d'apprentissage en contexte avec peu d'exemples (*few-shot*) sans modification des poids sur les mêmes modèles ; ainsi que (3) la comparaison systématique de ces deux paradigmes sur les deux jeux de données d'essais cliniques NLI4CT et NLI4PR. Tous nos scripts sont disponibles sur notre GitHub [2].

2 Travaux connexes

2.1 Affinage de grands modèles de langue

L'affinage de modèles pré-entraînés a démontré son efficacité pour améliorer les performances d'un modèle sur une tâche donnée. Malgré leur vaste corpus d'entraînement initial, les LLM confirment ce constat, tout particulièrement dans les domaines de spécialité. Plusieurs approches sont détaillées dans la littérature, telles que l'apprentissage supervisé, l'apprentissage continu ou encore l'apprentissage par renforcement (Lu *et al.*, 2025). Dans le domaine biomédical, des approches fondées sur l'apprentissage supervisé ont permis d'améliorer les performances par rapport à un LLM utilisé uniquement en inférence (Savage *et al.*, 2025; Anisuzzaman *et al.*, 2025). Comme les modèles utilisés ont une taille conséquente, l'adaptation de bas rang (*Low-Rank Adaptation (LoRA)*) (Hu *et al.*, 2021) est utilisée pour réduire la puissance de calcul nécessaire à l'affinage.

1. https://barometredelascienceouverte.esr.gouv.fr/sante/essais-cliniques/caracteristiques?id=caracteristiques.duree
2. https://github.com/LubinLgp/NLI_Finetuning

2.2 Amorçage (*Prompting*)

2.2.1 Techniques d'amorçage

Les méthodes à base d'amorces sont particulièrement répandues dans l'utilisation de modèles génératifs. Bien qu'elles soient peu coûteuses en développement, car elles ne nécessitent pas d'apprentissage supplémentaire, il convient néanmoins de porter une attention particulière à la manière de créer l'amorce (Fagbohun *et al.*, 2024). En effet, Jullien *et al.* (2025) démontrent que la structure de l'amorce explique à elle seule jusqu'à 44 % de la variance des scores de performance (macro F1) dans le cas de NLI4CT.

Amorces directes et formulations cliniques Les approches standard de type *Zero-shot* (sans démonstration) reposent traditionnellement sur un formatage strict. L'amorce est explicitement structurée par des mots-clés (par exemple : « Prémisse : [texte] », « Hypothèse : [texte] »), incitant le modèle à générer uniquement l'étiquette cible (*Entailment* ou *Contradiction*). Cette méthodologie a notamment été reprise pour l'ILN clinique par Jullien *et al.* (2025). Récemment, pour s'affranchir de cette rigidité, Aguiar *et al.* (2025) ont proposé de reformuler ces amorces sous la forme d'une phrase interrogative complète et conversationnelle, telle qu'un professionnel de la santé la formulerait dans la pratique. Cependant, des évaluations fines ont montré que, même avec ces formulations explicites, les LLM interrogés sans apprentissage préalable (*Zero-shot*) peinent à résoudre les inférences impliquant des négations ou des comparaisons numériques (Aguiar *et al.*, 2026).

Raisonnement par chaîne de pensée Des amorces basées sur le raisonnement par chaîne de pensée (*Chain-of-Thought (CoT)*) (Wei *et al.*, 2022) permettent au modèle d'expliciter son raisonnement, étape par étape, avant de conclure pour la question posée. Si cette approche reste performante pour l'appariement de patients (Jin *et al.*, 2024), son efficacité pour l'ILN de NLI4CT est contrastée. Brutti-Mairesse & Verlingue (2024) ont montré qu'une approche *Zero-shot CoT* améliore la fidélité et la cohérence des prédictions, alors que dans d'autres cas (Aguiar *et al.*, 2024), on observe une baisse du score F1. Jullien *et al.* (2025) proposent l'utilisation du *CoT* en combinaison avec *LoRA*, ce qui permet à de petits modèles d'atteindre des performances similaires à celles de modèles plus grands.

2.2.2 Optimisation du choix des démonstrations

La performance de l'apprentissage avec peu d'exemples (*few-shot learning*) repose fortement sur le choix des démonstrations intégrées à l'amorce. Il existe plusieurs stratégies basées sur la sélection d'exemples similaires à la requête test considérée (Rubin *et al.*, 2022; Lepagnol *et al.*, 2025) ou bien d'exemples diversifiés (Levy *et al.*, 2023; Agrawal *et al.*, 2023; Ye *et al.*, 2023). Liu *et al.* (2022) proposent la méthode KATE (*KNN-Augmented in-context Example selection*) qui consiste à encoder la requête test et les exemples d'entraînement (à l'aide d'encodeurs tels que RoBERTa (Liu *et al.*, 2019)), puis à sélectionner les k exemples d'entraînement dont les vecteurs sont les plus proches de la requête test.

2.3 Jeux de données pour l'ILN appliquée aux essais cliniques

***Natural Language Inference for Clinical Trials* (NLI4CT)** Le jeu de données NLI4CT (Jullien *et al.*, 2023a) vise à automatiser l'interprétation des rapports d'essais cliniques (REC) face au volume croissant de la littérature médicale. Chaque exemple est une paire (prémisse, hypothèse) exprimée en anglais (voir un exemple en annexe 1). Dans ce corpus, la prémisse est tirée d'une des quatre sections (Éligibilité, Intervention, Résultats, Effets secondaires) d'un rapport sur le cancer du sein, et l'hypothèse est une ou deux phrases affirmatives. Cette classification binaire (étiquettes *Entailment* ou *Contradiction*, fournies par des experts) comprend deux types d'exemples : *Single* (concerne un seul rapport) et *Comparison* (compare deux rapports). Le jeu d'entraînement comporte 1 700 instances et celui de test 500.

***Natural Language Inference for Patient Recruitment* (NLI4PR)** NLI4PR (Aguiar *et al.*, 2025), également en anglais, vise à étudier la faisabilité d'un recrutement automatisé, en laissant le patient utiliser ses propres mots pour décrire son profil médical (voir un exemple en annexe 2). Les auteurs ont reformulé manuellement en langage courant des profils médicaux issus de TREC-CT 2022 (Roberts *et al.*, 2022). Ici, la prémisse correspond aux critères d'éligibilité de l'essai clinique et l'hypothèse correspond au profil du patient. L'étiquette est *Entailment* si le patient est éligible et *Contradiction* sinon. NLI4PR permet de comparer l'usage d'une description en jargon médical à celui d'une description en langage patient pour un même exemple. Le jeu de données est plus conséquent que NLI4CT, avec un jeu d'entraînement d'environ 5 000 instances et un jeu de test de plus de 1 500 instances (voir tab. 7).

2.4 Bilan et positionnement de notre étude

Bien que l'ingénierie d'amorces et l'affinage aient été explorés dans le domaine médical, la littérature actuelle manque de comparaisons entre ces deux paradigmes pour des tâches d'inférence (Zaghir *et al.*, 2024). Certains travaux essayent de cartographier les faiblesses des LLM face aux différents types d'inférences (Aguiar *et al.*, 2026), mais peu d'études explorent les mécanismes d'échec de chaque approche. Nous proposons une comparaison entre l'apprentissage avec peu d'exemples et l'affinage sur deux jeux de données complémentaires, le tout en mettant en évidence la complémentarité de ces méthodes.

3 Méthodes

Pour toutes les expériences suivantes, l'évaluation porte sur l'intégralité des jeux de test, avec un modèle *baseline* (sans affinage) et un modèle affiné.

3.1 Stratégies d'amorçage

Nous utilisons différentes amorces inspirées de la littérature. Contrairement aux amorces P2 à P4 qui regroupent le contexte dans un unique message utilisateur, P1 et les configurations P5 (*few-shot*) exploitent le format conversationnel des modèles *Instruct* en séparant le message système (rôle du

modèle et tâche) du message utilisateur (prémisse et hypothèse). Les textes exacts sont fournis en annexe C.

— $P1_{\text{NLI4CT-NLI4PR}}$ (*Standard*) : Demande de classer la relation de la prémisse vers l'hypothèse par un seul mot (`Entailment` ou `Contradiction`).
— $P2_{\text{NLI4CT}}$ (*Explicit question*) : Même structure, mais formulée comme une interrogation directe (*« Is this premise in agreement with the following hypothesis ? »*).
— $P3_{\text{NLI4PR}}$ (*Clinical matching*) : Interroge directement le modèle sur l'éligibilité du profil patient face aux critères d'éligibilité.
— $P4_{\text{NLI4PR}}$ (*CoT*) : Ajoute à $P3$ une instruction de raisonnement explicite étape par étape avant de conclure.
— $P5_{\text{NLI4CT-NLI4PR}}$ (*Few-shot*) : Précède la requête de deux démonstrations (une pour chaque étiquette) pour guider le modèle.

Dans le cas de NLI4CT *Few-shot*, plusieurs stratégies de choix des exemples ont été expérimentées :
— Sélection par *type* et *section* : les exemples sont choisis dans le jeu d'entraînement en conservant le *type* de tâche et l'identifiant de *section* de l'instance de test. Pour chaque couple (type, section), on retient un exemple de chaque label.
— KATE-R : pour chaque instance de test, des exemples similaires selon leur représentation RoBERTa sont sélectionnés dans le jeu d'entraînement sans tenir compte du type ou de la section.
— KATE-S : variante de KATE exploitant *sentence-transformers* (Reimers & Gurevych, 2019) afin de générer des représentations vectorielles nativement optimisées pour la similarité sémantique, palliant ainsi les limites du RoBERTa standard sur cette tâche.

Le tableau 8 de l'annexe D.1 résume les hyperparamètres pour l'inférence.

3.2 Affinage

Nous utilisons Qwen2.5-7B-Instruct (Yang *et al.*, 2024) ainsi que Phi-4 (Abdin *et al.*, 2024), des modèles de taille moyenne (respectivement sept et 14 milliards de paramètres) dont la pertinence pour l'ILN clinique a été récemment démontrée par Jullien *et al.* (2025) et Aguiar *et al.* (2025). Dans le cadre de nos expériences, le terme affinage désigne un affinage supervisé (*Supervised Fine-tuning*, SFT) sur les données d'entraînement, dédié spécifiquement à la tâche de prédiction en ILN. Pour rendre l'entraînement réalisable efficacement, nous avons appliqué une méthode d'optimisation paramétrique efficace (*PEFT*) basée sur *LoRA*. Le détail des hyperparamètres est disponible dans l'annexe D.2. Dans le cas de NLI4PR, pour des raisons de ressources matérielles, la phase d'affinage n'a pas été réalisée sur l'intégralité des données d'entraînement mais sur une combinaison de 50 % des instances en langage patient et 50 % en langage médical. La phase d'évaluation a été menée sur tout le jeu de test.

4 Résultats

Les tableaux 1a et 1b résument les performances obtenues sur NLI4CT et NLI4PR respectivement. Chaque expérience est réalisée trois fois avec différentes graines aléatoires (13, 42, 1234) et évaluée sur le jeu de test. Nous nous comparons aux performances de la littérature : Mixtral-8x7B en *Zero-shot*

CoT (Brutti-Mairesse & Verlingue, 2024) et l'état de l'art basé sur un ensemble de modèles (Jullien *et al.*, 2024) pour NLI4CT ; ainsi que Qwen-14B en *zero-shot* (Aguiar *et al.*, 2025) pour NLI4PR.

4.1 Résultats sur NLI4CT

Nous privilégions la macro-F1 (non pondérée), plus représentative des performances réelles en cas de déséquilibre entre les classes.

Approche	**Exact. (± ET)**	**F1 (± ET)**
Littérature		
Mixtral-8x7B (*Zero-shot*)	-	70,0
État de l'art (Ensemble)	-	80,0
Zero-shot (Baselines)		
Qwen P1	65,3 ± 0,5	62,9 ± 0,6
Qwen P2	66,5 ± 0,2	64,6 ± 0,3
Phi-4 P1	73,9 ± 0,6	73,3 ± 0,6
Phi-4 P2	69,1 ± 1,4	71,6 ± 1,0
Few-Shot (P5)		
Qwen (Type + Section)	71,3 ± 0,4	71,0 ± 0,5
Qwen (KATE-R)	71,0 ± 0,5	70,9 ± 0,5
Qwen (KATE-S / Cosine)	70,7 ± 0,4	70,5 ± 0,4
Phi-4 (Type + Section)	77,3 ± 1,7	77,3 ± 1,7
Phi-4 (KATE-R)	75,1 ± 0,3	75,3 ± 0,2
Phi-4 (KATE-S / Cosine)	75,3 ± 1,0	75,3 ± 1,0
Affinage (LoRA)		
Qwen P1	76,1 ± 0,3	76,0 ± 0,3
Qwen P2	74,8 ± 0,9	74,8 ± 0,9
Phi-4 P1	77,5 ± 0,6	77,5 ± 0,6
Phi-4 P2	**79,2 ± 0,7**	**79,2 ± 0,7**

(a) NLI4CT : Exactitude et Macro F1 (%) moyennés sur 3 seeds

Approche	**Global (± ET)**	**PAT (± ET)**	**MED (± ET)**
Littérature			
Qwen-14B (*Zero-shot*)	-	71,8	73,1
Zero-shot (Baselines)			
Qwen P1	51,8 ± 0,2	48,2 ± 0,3	55,1 ± 0,2
Qwen P3	66,5 ± 0,4	65,0 ± 0,4	67,9 ± 0,5
Qwen P4	71,3 ± 0,6	69,5 ± 1,3	73,1 ± 0,1
Phi-4 P1	62,9 ± 0,3	66,1 ± 0,6	59,6 ± 0,1
Phi-4 P3	**76,1 ± 0,2**	**74,7 ± 0,3**	**77,6 ± 0,4**
Phi-4 P4	66,0 ± 0,6	64,2 ± 0,7	67,8 ± 0,6
Few-Shot (P5)			
Qwen (KATE-R)	65,9 ± 0,2	65,1 ± 0,4	66,6 ± 0,2
Phi-4 (KATE-R)	70,9 ± 0,3	68,9 ± 0,1	72,8 ± 0,5
Affinage (LoRA)			
Qwen P1	76,3 ± 1,0	76,1 ± 1,3	76,6 ± 0,8
Qwen P3	72,2 ± 0,3	72,2 ± 0,7	72,2 ± 0,2
Qwen P4	72,4 ± 0,4	71,6 ± 0,4	73,3 ± 0,7
Phi-4 P1	**76,9 ± 0,5**	**76,3 ± 0,9**	**77,4 ± 0,1**
Phi-4 P3	75,9 ± 0,3	75,8 ± 0,4	75,9 ± 0,4
Phi-4 P4	75,2 ± 0,4	74,7 ± 0,5	75,8 ± 0,3

(b) NLI4PR : Macro F1 (%) moyennés sur 3 seeds.

TABLE 1 – Comparaison des approches sur les deux jeux de données

En configuration *Zero-shot* sans affinage, Phi-4 surpasse Qwen pour les deux prompts (+10,4 points pour $P1$ et +7,0 points pour $P2$). Après affinage, l'écart de performance entre les deux modèles se réduit (+1,5 point pour $P1$ et +4,4 points pour $P2$). L'affinage avec le prompt $P1$ donne les meilleurs résultats pour Qwen, tandis que c'est le prompt $P2$ pour Phi-4. Le format de $P1$ permet manifestement aux poids de Qwen de se spécialiser plus efficacement sur notre tâche, alors que l'on observe le phénomène inverse pour Phi-4.

Les stratégies *Few-Shot* restent compétitives avec l'affinage. Sans aucune modification des poids, les sélections heuristiques (Type+Section) et sémantiques (KATE-R, KATE-S) surpassent largement les *baselines* et rivalisent presque avec les modèles affinés. Cet écart de performance est particulièrement faible pour Phi-4 (seulement 1,9 point de différence entre le meilleur *Few-shot* et l'affinage sur $P2$), bien qu'il soit plus élevé pour Qwen (environ 5 points d'écart avec l'affinage sur $P1$). Le *few-shot* offre ainsi un excellent compromis entre performance, faible besoin computationnel et empreinte carbone (voir la sec. 7). Cependant, peu importe la méthode employée, la section *Eligibility* reste l'une des plus difficiles à traiter (voir la sec. 5.1 ainsi que les fig. 4 et 5) ce qui motive une analyse supplémentaire sur NLI4PR qui se focalise sur cette section des REC. Nous observons également que l'affinage introduit des erreurs de prédiction et que les modèles sans affinage présentent des

tendances de prédiction où Qwen (particulièrement sur P1) tend à prédire plus souvent *Contradiction*, et à l'inverse, Phi-4 sans affinage sur-prédit *Entailment*. La fig. 6 en annexe illustre ces déséquilibres et l'amélioration apportée par l'affinage. Les trois méthodes *Few-shot* se comportent de manière similaire en fonction du type de tâche (*Single* ou *Comparison*) ou de la section du REC proposée (voir fig. 7).

Phi-4 affiné sur $P2$ est le modèle le plus performant en macro F1 (79,2 %), suivi de la stratégie *Few-shot* sur Phi-4 (77,3 %) et de Qwen affiné sur $P1$ (76,0 %). Nos modèles surpassent l'approche *Zero-shot CoT* de Brutti-Mairesse & Verlingue (2024) (70,0 %), tout en utilisant des architectures 3 à 6 fois plus petites. L'état de l'art de 80,0 % de macro F1 (Jullien *et al.*, 2024) repose sur des méthodes d'ensembles et d'auto-cohérence nécessitant une puissance computationnelle largement plus élevée que nos approches. Nous démontrons ici qu'un modèle de taille intermédiaire, affiné en LoRA ou simplement guidé par un *few-shot* judicieux, offre une alternative compétitive se situant à moins d'un point de l'état de l'art. Par ailleurs, bien que la sélection par *type* et *section* soit la stratégie *few-shot* la plus efficace sur NLI4CT, elle n'est pas compatible avec NLI4PR (où il n'y a pas de type ni de section de protocole). Nous utilisons donc la stratégie KATE-R sur NLI4PR qui, elle, est compatible.

4.2 Résultats sur NLI4PR

Le gain apporté par l'affinage varie selon l'architecture et l'amorce utilisée. Pour l'amorce standard $P1$, l'affinage permet une amélioration importante pour les deux modèles (+24,5 points pour Qwen et +14,0 points pour Phi-4). En revanche, pour l'amorce $P4$ (« chaîne de pensée ») le gain reste marginal pour Qwen (+1,1 point) contrairement à Phi-4 (+9,2 points). Les données d'entraînement utilisées pour l'affinage ne comportent aucune étape de raisonnement, imposant aux modèles d'apprendre à répondre directement. Ce décalage de format entre l'entraînement et l'inférence pénalise Qwen, dont le gain sur l'amorce P4 reste marginal (+1,1 point). À l'inverse, Phi-4 parvient à améliorer significativement son score sur cette même amorce (+9,2 points). Cela suggère que l'architecture de Phi-4 lui permet de tirer profit des connaissances cliniques acquises durant l'affinage, tout en conservant sa capacité pré-entraînée à générer et exploiter des chaînes de pensée.

Le format de l'instruction peut concurrencer l'affinage : la version *Zero-shot* de Phi-4 avec l'amorce $P3$ (76,1 %) surpasse ainsi la majorité des modèles affinés. Par ailleurs, l'affinage permet de réduire l'écart de performance entre les versions PAT et MED. Pour Qwen $P1$, cet écart passe de 6,9 à 0,5 point, indiquant une meilleure généralisation sur le langage patient. Les meilleures performances affinées, atteignant 76,3 % pour Qwen $P1$ et 76,9 % pour Phi-4 $P1$, surpassent les résultats de Aguiar *et al.* (2025), confirmant l'efficacité de l'approche LoRA face à des modèles *Zero-shot* de taille équivalente ou supérieure. Enfin, la faible variance entre les graines aléatoires confirme la robustesse de ces observations.

5 Analyse d'erreurs

5.1 NLI4CT

Nous évaluons la robustesse de nos modèles à travers trois axes : la significativité statistique des écarts de performance, la stabilité des prédictions face aux variations aléatoires et l'influence des

caractéristiques linguistiques.

Comparaison des approches Nous appliquons le test de McNemar[3] ($N = 500$) pour comparer les performances de l'affinage (*FT*) et de l'apprentissage avec peu d'exemples (*FS*) (tab. 2).

TABLE 2 – Tests de McNemar sur NLI4CT ($N = 500$). Δ *Exact.* désigne la différence d'exactitude entre les deux approches. La colonne *Disc.* détaille les paires discordantes : elle indique le nombre de cas où le premier modèle cité a faux alors que le deuxième a juste, par rapport au scénario inverse (le premier a juste mais le deuxième a faux).

	Paradigmes (FT vs FS)		**Modèles (Phi-4 vs Qwen)**		**Bilan**
Métrique	**(A) Phi-4 (FT P2 vs FS T+S)**	**(B) Qwen (FT P1 vs FS T+S)**	**(C) FT Phi4 P2 vs FT Qwen P1**	**(D) FS T+S**	**(E) Phi-4 ZS/FT P2**
Δ Exact.	+2,6%	+4,4%	-4,0%	-5,8%	+8,6%
Disc.	56/43	77/55	44/64	44/73	82/39
p-value	0,227	0,067	0,067	**0,009**	**<0,001**

La différence de performance entre affinage et *few-shot* optimisé n'est pas significative, malgré un coût computationnel plus important pour l'affinage. En revanche, de manière générale, Phi-4 est significativement plus performant que Qwen sur NLI4CT.

L'analyse de la stabilité de performance entre les trois graines aléatoires démontre que les approches *few-shot* sont plus stables que les modèles affinés. En observant le taux d'unanimité (la proportion d'instances (N=500) pour lesquelles un modèle prédit exactement la même classe lors de ses trois exécutions), on constate une grande constance pour les approches *few-shot* : 91,2% d'unanimité pour Qwen et 89,6% pour Phi-4. En revanche, les modèles affinés (LoRA) présentent des désaccords plus fréquents (prédiction différente selon la graine), ce qui fait chuter l'unanimité à 85,8% pour Qwen et à 82,0% pour Phi-4. Dans la partie suivante, les prédictions analysées sont issues d'un vote majoritaire sur les trois exécutions pour chaque instance.

5.1.1 Analyse des indicateurs linguistiques

Nous analysons l'influence de quatre indicateurs linguistiques : (1) chevauchement lexical entre prémisse et hypothèse (Jaccard : taille de l'intersection sur l'union des mots ; couverture : part des mots de l'hypothèse présents dans la prémisse) ; (2) densité numérique (somme du nombre de chiffres, de pourcentages, d'expressions avec unités et de nombres en toutes lettres dans prémisse et hypothèse) ; (3) négations (nombre d'occurrences de négations dans prémisse et hypothèse) ; (4) longueur en mots (nombre de tokens alphanumériques). Le détail du calcul de chaque indicateur est donné en annexe (tab. 10).

Analyse globale des paradigmes Le tab. 3 compare l'affinage (utilisant le prompt $P1$ pour Qwen et $P2$ pour Phi-4) et le *few-shot* (avec la stratégie de sélection par *type* et *section*). Pour chaque modèle, nous étudions les moyennes des indicateurs linguistiques, réparties selon quatre profils de prédiction croisée. Via le test de Kruskal-Wallis (K-W), nous déterminons quels indicateurs linguistiques influent sur la performance.

3. Dans tous les tests de significativité suivants nous choisissons un seuil $p < 0,05$ pour la p-valeur.

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
FS ✓ / FT ✓ (303)	0,106	0,344	43,14	0,82	131,7	21,9
FS × / FT × (65)	0,100	0,370	53,75	1,13	154,6	22,3
FS ✓ / FT × (55)	0,098	0,380	51,52	1,34	154,2	21,6
FS × / FT ✓ (77)	0,109	0,354	44,96	1,05	137,9	21,9
p-value (K-W)	0,482	0,152	0,130	0,257	0,093	0,902

(a) Modèle Qwen2.5-7B (Few-shot vs affiné P1)

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
FS ✓ / FT ✓ (344)	0,106	0,344	43,14	0,82	131,7	21,9
FS × / FT × (57)	0,100	0,370	53,75	1,13	154,6	22,3
FS ✓ / FT × (43)	0,107	0,378	43,15	1,17	146,2	22,4
FS × / FT ✓ (56)	0,118	0,373	45,63	1,37	159,6	27,6
p-value (K-W)	0,169	**0,045**	0,940	0,136	0,224	**<0,001**

(b) Modèle Phi-4 (Few-shot vs affiné P2)

TABLE 3 – Indicateurs linguistiques : comparaison entre Qwen2.5-7B et Phi-4.

La plupart des indicateurs ne montrent pas de différence significative pour Qwen, ce qui suggère que le modèle se comporte de manière similaire peu importe la configuration. Phi-4 est sensible au chevauchement lexical, les erreurs survenant plus fréquemment lorsque l'hypothèse partage une proportion importante de vocabulaire avec la prémisse (couverture $> 0,37$). D'autre part, la longueur de l'hypothèse a un impact hautement significatif : Phi-4 affiné parvient à corriger les erreurs du *few-shot* spécifiquement sur des hypothèses nettement plus longues (27,6 mots en moyenne contre environ 22 mots). L'affinage améliore donc la capacité de Phi-4 à traiter des affirmations complexes et longues.

Nous nous concentrons sur la section *Eligibility* car elle est la section la plus difficile à traiter même pour nos meilleures approches (voir fig. 3).

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
FS ✓ / FT ✓ (84)	0,119	0,388	15,65	1,94	153,2	23,8
FS × / FT × (15)	0,106	0,444	13,60	2,33	133,3	23,8
FS ✓ / FT × (14)	0,084	0,457	22,14	2,64	**256,4**	24,5
FS × / FT ✓ (19)	0,104	0,386	18,00	3,53	229,8	23,8
p-value (K-W)	0,972	0,098	0,175	0,230	**0,042**	0,972

(a) Modèle Qwen2.5-7B (Few-shot vs affiné P1)

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
FS ✓ / FT ✓ (80)	0,116	0,373	13,98	1,82	151,8	23,5
FS × / FT × (20)	0,113	0,470	19,35	2,35	179,1	24,3
FS ✓ / FT × (11)	0,126	0,375	28,54	**5,45**	255,3	25,1
FS × / FT ✓ (21)	0,118	0,441	17,52	2,43	198,4	25,1
p-value (K-W)	0,489	**0,042**	0,266	**0,014**	0,078	0,692

(b) Modèle Phi-4 (Few-shot vs affiné P2)

TABLE 4 – Section Eligibility : Indicateurs linguistiques pour le Few-shot et l'affinage.

Dans le cas de Qwen, le nombre de mots dans la prémisse influe sur la performance. Lorsque les deux approches échouent conjointement (FS × / FT ×) cela correspond aux prémisses les plus courtes (133,3 mots en moyenne), suggérant que ces échecs sont dus à une difficulté sémantique plutôt qu'à la longueur. À l'inverse, les textes très longs provoquent une forte instabilité et des désaccords entre les approches : l'affinage tend à échouer sur les prémisses très longues (256,4 mots en moyenne), tandis que le modèle *few-shot* échoue sur des prémisses légèrement moins longues (229,8 mots).

Dans le cas de Phi-4, en configuration *few-shot*, le modèle montre une sensibilité au chevauchement lexical. Le modèle se base sur la similarité de la prémisse et l'hypothèse, tandis qu'une fois affiné, il s'y montre moins sensible. Cependant, lorsque le modèle affiné échoue là où le *few-shot* réussissait, le texte contient en moyenne 5,45 négations. L'affinage compromet donc la capacité du modèle à traiter des critères d'exclusion multiples.

5.1.2 Comparaisons inter-modèles (Eligibility)

Le tab. 5 compare les performances de Qwen et Phi-4 à paradigme d'apprentissage équivalent (*few-shot* d'une part, puis modèles affinés d'autre part).

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
Qw ✓/ Ph ✓(93)	0,118	0,388	14,35	1,82	147,7	23,6
Qw × / Ph × (11)	0,094	0,474	14,09	3,72	171,0	24,1
Qw ✓/ Ph × (9)	0,111	0,404	28,33	4,33	253,3	25,3
Qw × / Ph ✓(19)	0,103	0,419	21,36	2,78	214,3	24,8
p-value (K-W)	0,816	0,230	0,172	**0,049**	0,086	0,918

(a) Finetuné (Qwen P1 vs Phi-4 P2)

Config. (n)	Jac.	Couv.	Num.	Nég.	Mots Prém.	Mots Hyp.
Qw ✓/ Ph ✓(88)	0,119	0,382	15,31	1,84	153,6	23,7
Qw × / Ph × (22)	0,094	0,453	15,50	2,36	152,7	23,9
Qw ✓/ Ph × (10)	0,086	0,551	24,70	2,40	265,3	25,6
Qw × / Ph ✓(12)	0,112	0,415	14,50	3,41	213,0	24,6
p-value (K-W)	0,154	**0,004**	0,332	0,188	**0,042**	0,593

(b) Few-shot T+S (Qwen vs Phi-4)

TABLE 5 – *Eligibility* : Comparaison des modèles Qwen et Phi-4 (affiné vs Few-shot Type+Section)

Après affinage, Qwen démontre une résistance supérieure aux structures syntaxiques complexes : il parvient à résoudre des cas avec de nombreuses négations (4,33 en moyenne) là où Phi-4 échoue, une différence statistiquement significative. À l'inverse, bien que Phi-4 obtienne de meilleures performances globales en *few-shot*, il s'avère beaucoup plus sensible que Qwen aux formes de surface. En effet, les échecs spécifiques de Phi-4 face aux succès de Qwen sont fortement corrélés à un chevauchement de vocabulaire élevé (couverture de 0,551). De plus, en *few-shot*, Qwen conserve une performance stable sur des longues prémisses (265,3 mots) contrairement à Phi-4.

5.2 NLI4PR

Dans cette section, la prédiction finale de chaque exemple est sélectionnée via un vote à la majorité sur les trois prédictions issues des différentes graines aléatoires. Pour les approches *Baseline* et l'affinage, nous ne sélectionnons que les instances où le modèle échoue simultanément sur les trois amorces testées ($P1$, $P3$, $P4$). Cela permet d'isoler les instances les plus complexes en s'affranchissant des erreurs de prédiction liées au format de l'amorce. Le tab. 6 rapporte les indicateurs linguistiques moyens pour ces erreurs (*Err.*) ainsi que pour leurs groupes de contrôle respectifs (*Ctrl.*) (instances réussies ou non systématiquement fausses).

Groupe (n)	Jac.	Couv.	Dens.	Nég.	Mots Prém.	Mots Hyp.
Baseline (Zero-shot)						
Err. PAT (234)	0,074	0,183	36,8	3,83	228,4	118,8
Ctrl. PAT (1344)	0,076	0,197	38,4	3,78	251,1	117,5
Err. MED (187)	0,092	0,221	37,0	4,29	211,5	100,9
Ctrl. MED (1391)	0,091	0,233	41,3	4,27	252,6	101,8
Few-shot						
Err. PAT (549)	0,074	0,188	36,5	3,76	241,3	119,6
Ctrl. PAT (1029)	0,077	0,198	39,0	3,81	251,2	116,7
Err. MED (525)	0,091	0,237	42,3	4,46	**268,3**	101,7
Ctrl. MED (1053)	0,091	0,229	40,0	4,17	**237,6**	101,7
Affinage (LoRA)						
Err. PAT (197)	0,076	0,191	37,4	3,54	232,3	119,2
Ctrl. PAT (1381)	0,076	0,195	38,2	3,83	250,0	117,5
Err. MED (172)	0,090	0,213	36,3	4,08	**210,0**	102,1
Ctrl. MED (1406)	0,091	0,234	41,3	4,29	**252,4**	101,7

(a) Modèle Qwen2.5-7B

Groupe (n)	Jac.	Couv.	Dens.	Nég.	Mots Prém.	Mots Hyp.
Baseline (Zero-shot)						
Err. PAT (219)	0,074	0,186	38,9	4,12	233,1	117,9
Ctrl. PAT (1359)	0,076	0,196	38,0	3,74	250,1	117,7
Err. MED (178)	0,090	0,228	39,4	4,57	238,1	100,9
Ctrl. MED (1400)	0,091	0,232	40,9	4,23	249,0	101,8
Few-shot						
Err. PAT (497)	0,074	0,197	41,2	3,85	**271,5**	120,4
Ctrl. PAT (1081)	0,077	0,194	36,7	3,76	**236,8**	116,5
Err. MED (429)	0,090	0,238	43,7	4,60	278,3	101,2
Ctrl. MED (1149)	0,092	0,229	39,7	4,15	236,4	101,9
Affinage (LoRA)						
Err. PAT (199)	0,075	0,182	**33,2**	3,61	**207,1**	118,4
Ctrl. PAT (1379)	0,076	0,197	**38,9**	3,82	**253,6**	117,6
Err. MED (173)	0,090	0,212	37,6	4,11	216,2	102,1
Ctrl. MED (1405)	0,091	0,234	41,2	4,29	251,7	101,6

(b) Modèle Phi-4-14B

TABLE 6 – NLI4PR : Indicateurs linguistiques moyens pour les erreurs persistantes et leurs groupes de contrôle respectifs. Les valeurs en gras signalent les écarts statistiquement significatifs discutés dans l'analyse.

En configuration *Zero-shot*, les erreurs ont tendance à survenir sur des prémisses légèrement plus courtes et moins denses numériquement que pour le groupe de contrôle (par exemple, 233 mots

en erreur contre 250 en contrôle pour Phi-4 PAT) sans toutefois être significatives. En l'absence d'affinage ou de démonstrations, les erreurs sont réparties de manière homogène par rapport à la structure du texte.

Malgré une diminution des erreurs, les modèles affinés échouent sur des prémisses plus courtes et présentant une plus faible densité numérique. Contrairement à la *baseline*, ces écarts deviennent statistiquement significatifs. C'est notamment le cas pour Phi-4 sur le domaine PAT (différence significative sur la longueur et la densité numérique) ainsi que pour Qwen sur le domaine MED (différence significative sur la longueur). Ces résultats suggèrent que les modèles affinés nécessitent une description clinique suffisamment riche et détaillée et des critères quantitatifs explicites pour appliquer correctement les règles apprises lors de l'entraînement.

Pour le *Few-shot*, la stratégie KATE-R présente un comportement inverse à celui de l'affinage. Les erreurs se concentrent sur les prémisses les plus longues et denses (ex. : 271 mots en moyenne pour les erreurs contre 236 pour le contrôle sur Phi-4 PAT). Ces différences de longueur sont significatives pour plusieurs sous-ensembles (notamment pour Qwen MED). En conclusion, l'apprentissage avec peu d'exemples voit ses performances se dégrader face à des textes trop longs, tandis que l'affinage perd en efficacité face à des descriptions cliniques trop concises manquant de critères explicites.

6 Conclusion et perspectives

Dans cette étude, nous avons comparé l'efficacité de l'affinage paramétrique (*LoRA*) et de l'apprentissage avec peu d'exemples (*few-shot*) pour l'ILN dans le cadre d'essais cliniques. Les évaluations sur les corpus NLI4CT et NLI4PR indiquent que l'affinage permet d'atteindre les meilleures performances globales, jusqu'à 25 points de F1 supplémentaires selon la tâche et le modèle. Néanmoins, les stratégies *few-shot* basées sur une sélection optimisée des démonstrations, comme KATE-R, restent compétitives dans le cas de NLI4CT, tout en ne nécessitant pas une grande puissance de calcul. En revanche, sur NLI4PR, la stratégie *few-shot* s'avère nettement moins efficace que l'affinage. Par ailleurs, l'analyse des erreurs révèle que les modèles réagissent très différemment selon la longueur des prémisses et la densité numérique. Ces constats motivent le développement d'architectures hybrides, capables de sélectionner dynamiquement la méthode d'inférence (affinage ou *few-shot*) selon les propriétés de surface du texte à traiter.

7 Considérations éthiques

En plus d'une comparaison de performance, nous avons estimé l'empreinte carbone de nos expériences avec *Green Algorithms* [4]. L'affinage s'avère coûteux, générant en moyenne pour Qwen2.5 (7B) 367 gCO_2e sur NLI4CT et 1350 gCO_2e sur NLI4PR, contre respectivement 697 gCO_2e et 2565 gCO_2e pour Phi-4 (14B). En comparaison, la phase d'inférence seule est bien plus légère : 6,38 et 27,35 gCO_2e pour Qwen, et 13,4 et 57,4 gCO_2e pour Phi-4. Bien que l'affinage offre les meilleurs résultats de classification, cette analyse souligne l'avantage écologique majeur du *few-shot*, qui permet d'atteindre des scores compétitifs tout en réduisant l'empreinte carbone de plus de 98 %.

4. https://calculator.green-algorithms.org/

Remerciements

Ces travaux ont bénéficié des financements du CNRS via l'allocation 80|PRIME et celle de l'Agence Nationale pour la Recherche via l'ANR-22-CPJ1-0087-01. Ces travaux ont bénéficié d'un accès aux moyens de calcul de LabIA de l'Université Paris-Saclay.

Références

ABDIN M., ANEJA J., BEHL H., BUBECK S., ELDAN R., GUNASEKAR S., HARRISON M., HEWETT R. J., JAVAHERIPI M., KAUFFMANN P., LEE J. R., LEE Y. T., LI Y., LIU W., MENDES C. C. T., NGUYEN A., PRICE E., DE ROSA G., SAARIKIVI O., SALIM A., SHAH S., WANG X., WARD R., WU Y., YU D., ZHANG C. & ZHANG Y. (2024). Phi-4 technical report.

AGRAWAL S., ZHOU C., LEWIS M., ZETTLEMOYER L. & GHAZVININEJAD M. (2023). In-context examples selection for machine translation. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 8857–8873, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.564.

AGUIAR M., ZWEIGENBAUM P. & NADERI N. (2024). SEME at SemEval-2024 task 2 : Comparing masked and generative language models on natural language inference for clinical trials. In A. K. OJHA, A. S. DOĞRUÖZ, H. TAYYAR MADABUSHI, G. DA SAN MARTINO, S. ROSENTHAL & A. ROSÁ, Éds., *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, p. 986–996, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.semeval-1.143.

AGUIAR M., ZWEIGENBAUM P. & NADERI N. (2025). Am I eligible ? natural language inference for clinical trial patient recruitment : the patient's point of view. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, p. 243–259, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.cl4health-1.21.

AGUIAR M., ZWEIGENBAUM P. & NADERI N. (2026). Assessing the difficulty of inference types in natural language inference for clinical trials. ⟨hal-05533706⟩.

ANISUZZAMAN D., MALINS J. G., FRIEDMAN P. A. & ATTIA Z. I. (2025). Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings : Digital Health*, **3**(1), 100184. DOI : https ://doi.org/10.1016/j.mcpdig.2024.11.005.

BRUTTI-MAIRESSE C. & VERLINGUE L. (2024). CRCL at SemEval-2024 task 2 : Simple prompt optimizations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, p. 437–442, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.semeval-1.67.

DAGAN I., GLICKMAN O. & MAGNINI B. (2006). The pascal recognising textual entailment challenge. In J. QUIÑONERO-CANDELA, I. DAGAN, B. MAGNINI & F. D'ALCHÉ BUC, Éds., *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, p. 177–190, Berlin, Heidelberg : Springer Berlin Heidelberg.

FAGBOHUN O., HARRISON R. M. & DEREVENTSOV A. (2024). An empirical categorization of prompting techniques for large language models : A practitioner's guide. *ArXiv*, **abs/2402.14837**.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

JIN Q., WANG Z., FLOUDAS C. S., CHEN F., GONG C., BRACKEN-CLARKE D., XUE E., YANG Y., SUN J. & LU Z. (2024). Matching patients to clinical trials with large language models. *Nature Communications*, **15**(1), 9074.

JULLIEN M., VALENTINO M. & FREITAS A. (2024). SemEval-2024 task 2 : Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, p. 1947–1962, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.semeval-1.271.

JULLIEN M., VALENTINO M., FROST H., O'REGAN P., LANDERS D. & FREITAS A. (2023a). NLI4CT : Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 16745–16764, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.1041.

JULLIEN M., VALENTINO M., FROST H., O'REGAN P., LANDERS D. & FREITAS A. (2023b). SemEval-2023 task 7 : Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, p. 2216–2226, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.semeval-1.307.

JULLIEN M., VALENTINO M., RANALDI L. & FREITAS A. (2025). Dissecting clinical reasoning in language models : A comparative study of prompts and model adaptation strategies. *arXiv preprint*. DOI : 10.48550/arXiv.2507.04142.

LEPAGNOL P., GHANNAY S., GERALD T., SERVAN C. & ROSSET S. (2025). Leveraging Information Retrieval to Enhance Spoken Language Understanding Prompts in Few-Shot Learning. In *Interspeech 2025*, Rotterdam, Netherlands. DOI : 10.21437/Interspeech.2025-175, HAL : hal-05095796.

LEVY I., BOGIN B. & BERANT J. (2023). Diverse demonstrations improve in-context compositional generalization. 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. 1401–1422, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.78.

LIU J., SHEN D., ZHANG Y., DOLAN B., CARIN L. & CHEN W. (2022). What makes good in-context examples for GPT-3 ? In E. AGIRRE, M. APIDIANAKI & I. VULIĆ, Éds., *Proceedings of Deep Learning Inside Out (DeeLIO 2022) : The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, p. 100–114, Dublin, Ireland and Online : Association for Computational Linguistics. DOI : 10.18653/v1/2022.deelio-1.10.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LU W., LUU R. K. & BUEHLER M. J. (2025). Fine-tuning large language models for domain adaptation : Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, **11**(1), 84.

NIEVAS M., BASU A., WANG Y. & SINGH H. (2024). Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*, **31**(9), 1953–1963. DOI : 10.1093/jamia/ocae073.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-*

IJCNLP), p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

ROBERTS K., DEMNER-FUSHMAN D., VOORHEES E. M., BEDRICK S. & HERSH W. R. (2022). Overview of the trec 2022 clinical trials track. In *Text Retrieval Conference*.

RUBIN O., HERZIG J. & BERANT J. (2022). Learning to retrieve prompts for in-context learning. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2655–2671, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.191.

SAVAGE T., P MA S., BOUKIL A., RANGAN E., PATEL V., LOPEZ I. & CHEN J. (2025). Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization : Comparative evaluation. *J Med Internet Res*, **27**, e76048. DOI : 10.2196/76048.

SERTKAYA A., WONG H.-H., JESSUP A. & BELECHE T. (2016). Key cost drivers of pharmaceutical clinical trials in the united states. *Clinical Trials*, **13**, 117 – 126.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E. H., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA : Curran Associates Inc.

WORNOW M., LOZANO A., DASH D., JINDAL J. A., MAHAFFEY K. W. & SHAH N. H. (2025). Zero-shot clinical trial patient matching with llms. *NEJM AI*, **2**(1), AIcs2400360.

YANG A., YANG B., HUI B., ZHENG B., YU B., ZHOU C., LI C., LI C., LIU D., HUANG F., DONG G., WEI H., LIN H., TANG J., WANG J., YANG J., TU J., ZHANG J., MA J., XU J., ZHOU J., BAI J., HE J., LIN J., DANG K., LU K., CHEN K., YANG K., LI M., XUE M., NI N., ZHANG P., WANG P., PENG R., MEN R., GAO R., LIN R., WANG S., BAI S., TAN S., ZHU T., LI T., LIU T., GE W., DENG X., ZHOU X., REN X., ZHANG X., WEI X., REN X., FAN Y., YAO Y., ZHANG Y., WAN Y., CHU Y., LIU Y., CUI Z., ZHANG Z. & FAN Z. (2024). Qwen2 technical report. *arXiv preprint arXiv :2407.10671*.

YE X., IYER S., CELIKYILMAZ A., STOYANOV V., DURRETT G. & PASUNURU R. (2023). Complementary explanations for effective in-context learning. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 4469–4484, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.273.

ZAGHIR J., NAGUIB M., BJELOGRLIC M., NÉVÉOL A., TANNIER X. & LOVIS C. (2024). Prompt engineering paradigms for medical applications : Scoping review. *Journal of Medical Internet Research*, **26**, e60501. DOI : 10.2196/60501.

A Statistiques sur les jeux de données

Tâche	Partition	Valeur
NLI4CT	Entraînement	1700
	Développement	200
	Test	500
NLI4PR	Entraînement	4904
	Développement	525
	Test	1578

TABLE 7 – Volume de données pour chacune des tâches utilisées.

B Exemples tirés des jeux de données

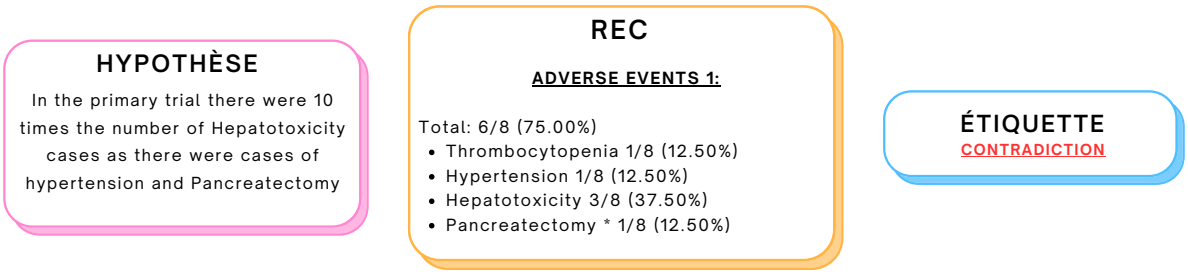

FIGURE 1 – Exemple tiré du jeu de données NLI4CT.

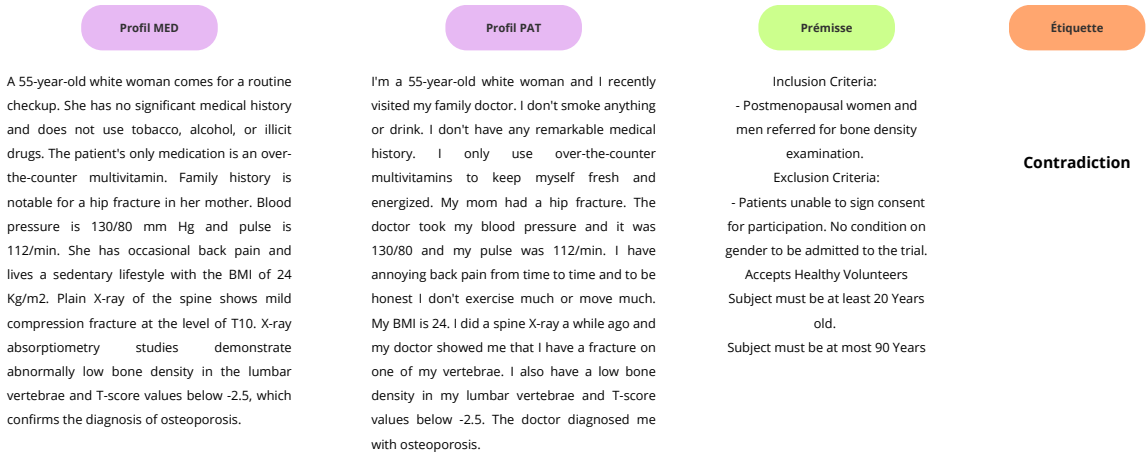

FIGURE 2 – Exemple tiré du jeu de données NLI4PR.

C Textes exacts des amorces

Les amorces respectent le format conversationnel des modèles *Instruct*. Les variables textuelles injectées sont notées entre crochets (ex. : `[Prémisse]`).

P1 (Standard) • **[System]** : *Classify the relationship between the premise and the hypothesis. Respond with only one word : 'Entailment' or 'Contradiction'.* • **[User]** : *PREMISE :* `[Prémisse]` | *HYPOTHESIS :* `[Hypothèse]`

P2 (Question explicite) • **[User]** : *PREMISE :* `[Prémisse]`
Is this premise in agreement with the following hypothesis ?
HYPOTHESIS : `[Hypothèse]`
Answer only with : Entailment or Contradiction.

P3 (Clinical matching) • **[User]** : *Does the patient with the statement* `"[Hypothèse]"` *satisfy the following clinical trial admission criteria ?*
`"[Prémisse]"`
Respond with only one word : 'Entailment' or 'Contradiction'.

P4 (CoT) • **[User]** : *Does the patient with the statement* `"[Hypothèse]"` *satisfy the following clinical trial admission criteria ?*
`"[Prémisse]"`
First, explain your reasoning step-by-step by comparing the patient's characteristics to the inclusion and exclusion criteria. Then, conclude on a new line with only one word : 'Entailment' or 'Contradiction'.

P5 (Few-shot) • **[System]** : *You will see several examples of premise-hypothesis pairs with their classification (Entailment or Contradiction). Use these examples to understand the task. Then classify the relationship between the last premise and hypothesis. Respond with only one word : 'Entailment' or 'Contradiction'.* • **[User]** : *PREMISE :* `[Prémisse 1]` | *HYPOTHESIS :* `[Hypothèse 1]` • **[Assistant]** : *Entailment* • **[User]** : *PREMISE :* `[Prémisse 2]` | *HYPOTHESIS :* `[Hypothèse 2]` • **[Assistant]** : *Contradiction* • **[User]** : *PREMISE :* `[Prémisse test]` | *HYPOTHESIS :* `[Hypothèse test]`

D Hyperparamètres d'inférence et d'affinage

D.1 Inférence

Jeu de données	Prompt	Température	Top-p	Max. tokens générés	Fenêtre de parsing
NLI4CT	Prompt 1 (NLI standard)	0,7	1,0	10	50 premiers caractères
NLI4CT	Prompt 2 (question explicite)	0,7	1,0	10	50 premiers caractères
NLI4CT	Few-shot (Prompt 1, 2 exemples)	0,7	1,0	10	50 premiers caractères
NLI4PR	Prompt 1 (NLI standard)	0,7	1,0	10	50 premiers caractères
NLI4PR	Prompt 2 (Clinical Matching)	0,7	1,0	10	50 premiers caractères
NLI4PR	Prompt 3 (CoT)	0,7	1,0	1024	Dernière ligne non vide
NLI4PR	Prompt 4 (few-shot KATE-R)	0,7	1,0	10	50 premiers caractères

TABLE 8 – Hyperparamètres d'inférence par jeu de données et prompt.

Le tableau 8 résume les réglages d'inférence par jeu de données et par *prompt*. La prédiction est issue du *parsing* de la sortie (on cherche les racines `entail` / `contradict` dans la fenêtre indiquée). Le message système *few-shot* précise que le modèle doit s'inspirer des exemples puis classer la dernière paire en un mot. L'évaluation repose sur le même principe de *parsing*

D.2 Affinage

L'affinage est réalisé avec la bibliothèque `trl` (*SFTTrainer*). Les hyperparamètres numériques (époques, *batch*, *learning rate*, LoRA) sont regroupés dans le tableau 9 ; la seule différence entre NLI4CT et NLI4PR est le nombre d'époques (5 vs 4). L'adaptation LoRA cible les modules

de projection du *Transformer* (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). Le modèle de base est quantifié en 4-bit (`bitsandbytes`, format *nf4*, double quantification) avec calculs en `bfloat16` ; le *gradient checkpointing* est activé pour limiter l'empreinte mémoire.

Jeu de données	Époques	Batch / GPU	Accumulation	LR	Longueur max.	LoRA r	LoRA α	LoRA dropout
NLI4CT	5	2	4	2×10^{-4}	4096	16	32	0,05
NLI4PR	4	2	4	2×10^{-4}	4096	16	32	0,05

TABLE 9 – Hyperparamètres d'affinage (affinage) pour NLI4CT et NLI4PR.

E Définition des indicateurs linguistiques (section 5.1)

Le tableau 10 précise le calcul de chaque indicateur utilisé dans l'analyse linguistique (complémentarité P1 / KATE-R et analyses univariées).

Indicateur	Nom interne	Méthode de calcul
Jaccard lexical	`lexical_jaccard`	Mots = tokens alphanumériques (minuscules). Jaccard = $\lvert P \cap H \rvert / \lvert P \cup H \rvert$ sur les ensembles de mots de la prémisse P et de l'hypothèse H.
Couverture	`lexical_coverage`	$\lvert P \cap H \rvert / \lvert H \rvert$ si $H \neq \emptyset$: part des mots de l'hypothèse présents dans la prémisse.
Densité numérique	`numeric_total`	Somme (prémisse + hypothèse) de : nombre de chiffres (`\d`), d'occurrences `nombre %`, d'expressions `nombre + unité` (mg, g, ml, years, months, etc.), et de nombres en toutes lettres (one, two, …, hundred, thousand, etc.).
Négations	`neg_total`	Somme prémisse + hypothèse du nombre d'occurrences de mots de négation (*not*, *no*, *never*, *none*, *cannot*, etc.) et de formes en *n't* (tokenisé).
Mots prémisse	`words_premise`	Nombre de *tokens* alphanumériques dans la prémisse.
Mots hypothèse	`words_hypothesis`	Nombre de *tokens* alphanumériques dans l'hypothèse.

TABLE 10 – Moyen de calcul des indicateurs linguistiques.

F Analyses quantitatives complémentaires

Cette annexe regroupe les évaluations graphiques détaillant les performances de nos modèles. La première partie décompose les résultats sur NLI4CT selon le type d'inférence, la section clinique ciblée et les matrices de confusion, tandis que la seconde illustre l'impact systématique de l'affinage pour chaque amorce sur NLI4PR (langages patient et médical).

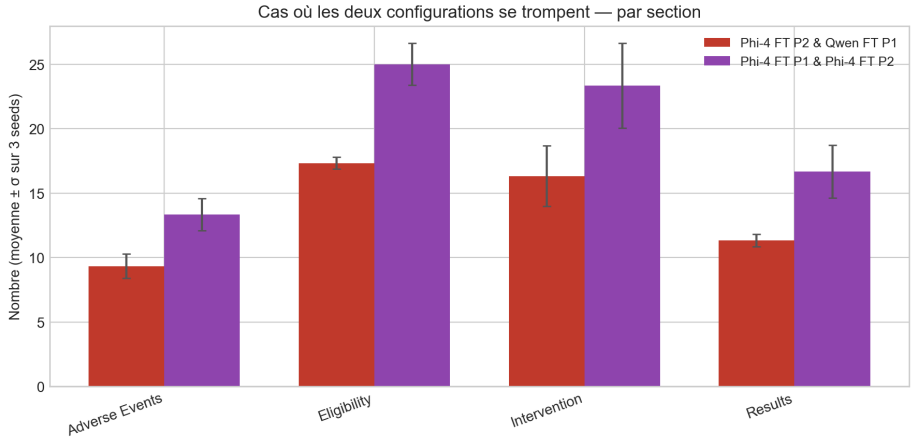

FIGURE 3 – Nombre d'erreurs communes au modèle affiné sur $P1$ et à celui affiné sur $P2$ en fonction de la section du REC. Les deux modèles sont nettement plus susceptibles de se tromper sur les sections *Eligibility* et *Intervention* que sur *Results* ou *Adverse Events*.

F.1 NLI4CT

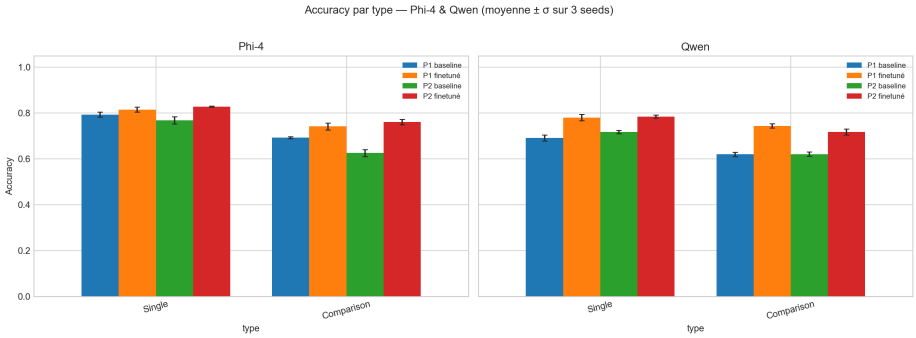

FIGURE 4 – Macro F1 par type de tâche (*Single / Comparison*) pour les modèles sans affinage et modèles affinés (*Prompt* 1 et *Prompt* 2).

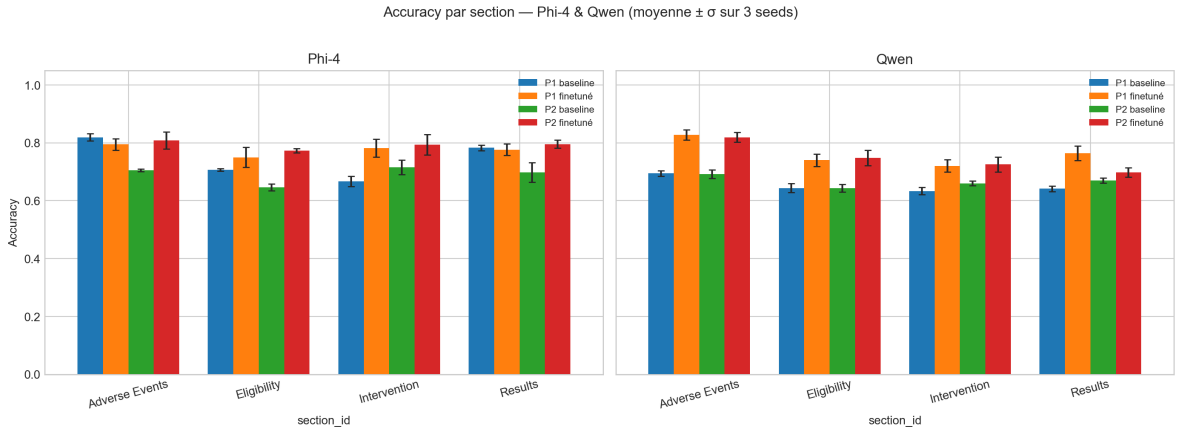

FIGURE 5 – Macro F1 par section de protocole pour les modèles sans affinage et modèles affinés.

FIGURE 6 – Matrices de confusion pour P1 *Baseline*, P1 affiné, P2 *Baseline* et P2 affiné. Les modèles sans affinage tendent à prédire majoritairement *Contradiction* ; l'affinage rééquilibre les prédictions, ce qui peut expliquer une partie des régressions (exemples où une prédiction correcte en contradiction du modèle sans affinage devient erronée après affinage).

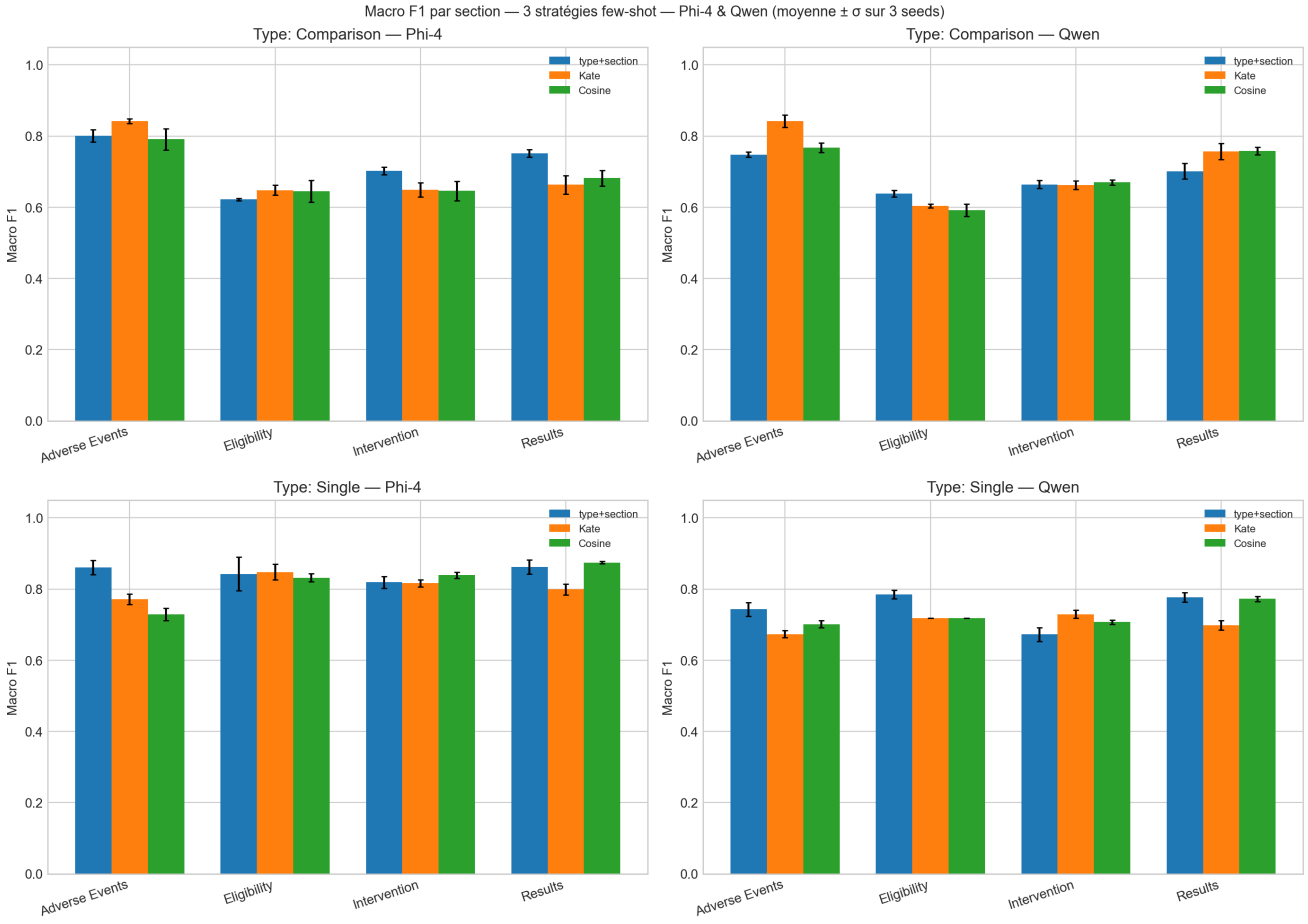

FIGURE 7 – Macro F1 par type de tâche (*Single / Comparison*) et par section de protocole pour les trois stratégies *few-shot* (type+section, KATE-R, KATE-S). Les trois stratégies montrent des profils très proches, justifiant un comportement similaire.