Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Connaissances factuelles dans les modèles de langue : robustesse et anomalies face à des variations simples du contexte temporel

Hichem Ammar Khodja[1,2], Frédéric Béchet[2,3], Quentin Brabant[1], Alexis Nasr[2], Gwénolé Lecorvé[1]
[1]Orange - *Lannion, France*,
[2]Aix Marseille Université, CNRS, LIS, UMR 7020 - *Marseille, France*,
[3]International Laboratory on Learning Systems (ILLS - IRL2020 CNRS)
Correspondence: {hichem.ammarkhodja, quentin.brabant, gwenole.lecorve}@orange.com, {frederic.bechet, alexis.nasr}@lis-lab.fr

RÉSUMÉ

Ce papier explore la robustesse des modèles de langue (ML) face aux variations du contexte temporel dans les connaissances factuelles. Il examine si les ML peuvent associer correctement un contexte temporel à un fait passé valide sur une période de temps délimitée, en leur demandant de différencier les contextes corrects des contextes incorrects. La capacité de distinction des ML est analysée sur deux dimensions : la distance du contexte incorrect par rapport à la période de validité et la granularité du contexte. Pour cela, un jeu de données, TimeStress, est introduit, permettant de tester 18 ML variés. Les résultats révèlent que le meilleur ML n'atteint une distinction parfaite que pour 11% des faits étudiés, avec des erreurs critiques qu'un humain ne ferait pas. Ces travaux soulignent les limites des ML actuels en matière de représentation temporelle.

ABSTRACT

Factual Knowledge in Language Models : Robustness and Anomalies under Simple Temporal Context Variations

This paper explores the robustness of language models (LMs) to variations in temporal context within factual knowledge. It examines whether LMs can correctly associate a temporal context with a past fact valid over a defined period, by asking them to differentiate correct from incorrect contexts. The LMs' ability to distinguish is analyzed along two dimensions : the distance of the incorrect context from the validity period and the granularity of the context. To this end, a dataset called TimeStress is introduced, enabling the evaluation of 18 diverse LMs. Results reveal that the best LM achieves a perfect distinction for only 11% of the studied facts, with critical errors that humans would not make. This work highlights the limitations of current LMs in temporal representation.

MOTS-CLÉS : Modèles de langue, Factualité, Temporalité, Sondage des connaissances.

KEYWORDS: Language Models, Factuality, Temporality, Knowledge Probing.

ARTICLE : **Soumis à** CORIA-TALN 2025.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 167–195.

1 Introduction

Lorsqu'un Modèle de Langue (ML) complète l'amorce textuelle « La capitale de la France est » par « Paris », il démontre qu'il a stocké ce fait quelque part dans ses paramètres. Cependant, comme l'ont montré nombre de travaux (Elazar *et al.*, 2021; Dong *et al.*, 2023; Hagen *et al.*, 2024; Kassner & Schütze, 2020), ce genre de connaissance factuelle n'est pas nécessairement robuste à certaines variations dans l'amorce (emploi de paraphrases, alias, erreurs typographiques, négations…). Parmi ces facteurs de variabilité, la dimension temporelle des connaissances factuelles a été moins étudiée. Ainsi, dans cet article, nous étudions la robustesse des connaissances factuelles des ML face à des variations du contexte temporel.

Alors que l'état de l'art a démontré certains biais dans les ML liés à la distribution temporelle de leurs données d'entraînement ou leur faiblesse à raisonner avec des concepts temporels, notre travail vise à quantifier à quel point les ML parviennent à correctement associer un contexte temporel (par exemple, une année, une date, telles que « *En 2018, …* », « *Le 5 novembre 2022, …* ») à un fait passé, c'est-à-dire un fait ayant une certaine période de validité. Plus précisément, les questions de recherche traitées sont :

1. est-ce que les ML distinguent des contextes temporels corrects et incorrects pour des faits ?
2. est-ce qu'ils les différencient avec la même précision en fonction de la distance du contexte incorrect à la période de validité des faits ?
3. est-ce que les ML activent aussi bien leurs connaissances factuelles lorsque le contexte temporel est très précis ou grossier ?

Pour cela, comme illustré sur la figure 1, des matchs sont organisés entre des contextes temporels corrects et incorrects pour mesurer les préférences des modèles, dégager des tendances générales et relever des anomalies. Comme mentionné dans les questions de recherche, deux angles d'études précis sont adoptés pour faire varier les contextes temporels au sein de ces matchs : le positionnement des contextes sur l'axe du temps et en termes de granularité (de l'année jusqu'à une date précise).

Les contributions de l'article sont :

— la mise à disposition d'un jeu de données, *TimeStress*, constitué de connaissances factuelles populaires (selon un indice de popularité), annotées temporellement et verbalisées. Ces données permettent de répliquer nos expériences mais également de réfléchir à d'autres études sur la temporalité.

— la mise en évidence de la faible robustesse des ML actuels vis-à-vis de leurs connaissances factuelles lorsqu'il s'agit de les positionner dans le temps, ainsi que d'erreurs, certes rares,

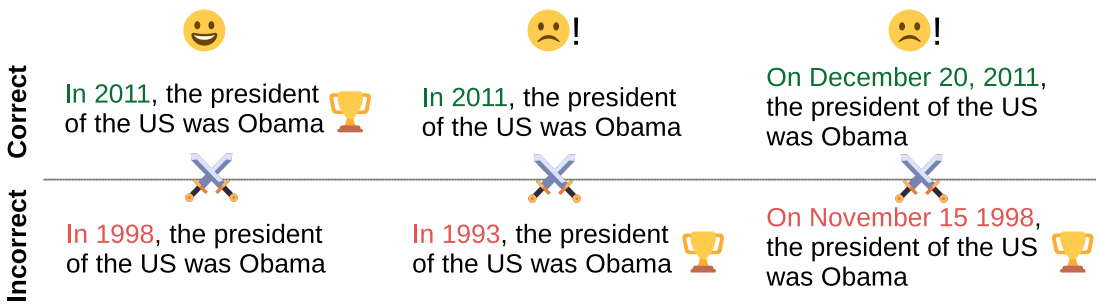


FIGURE 1 – La robustesse du ML sur un fait est évaluée en lui demandant de différencier un ensemble d'énoncés corrects et incorrectes. Le contexte temporel est varié sur deux dimensions : sa position sur la ligne du temps (colonnes 1 et 2) et sa granularité (colonnes 1 et 3).

mais critiques que ne ferait pas un humain. Ces résultats montrent les limitations des ML en termes de représentation interne de la temporalité, y compris pour les gros modèles (18 modèles testés de taille et famille variées).

Dans les sections suivantes, nous discutons d'abord des travaux connexes (section 2). Ensuite, nous développons les problématiques de l'article et présentons le jeu de données TimeStress (section 3). Enfin, nous décrivons nos expériences et analysons leurs résultats (section 4). Le code source permettant de reproduire nos résultats est publié sur GitHub [1] et TimeStress est distribué sur Hugging Face [2].

2 État de l'art

Cette section présente les travaux connexes au nôtre sous l'angle de l'étude des connaissances factuelles que possèdent des ML, de la prise en compte de l'aspect temporel de celles-ci puis de leur facultés de raisonnement temporel.

Robustesse des connaissances factuelles dans les ML. Il a été démontré que les ML stockent une quantité significative de connaissances factuelles (Petroni *et al.*, 2019; Jiang *et al.*, 2020; Sun *et al.*, 2024). Cependant, de nombreuses études indiquent que ces connaissances acquises manquent souvent de cohérence face aux perturbations du texte. Par exemple, Kassner & Schütze (2020) a mis en évidence les limites des ML pré-entraînés à s'adapter aux négations dans les questions, ce qui conduit à des réponses contradictoires. La robustesse à la paraphrase et aux fautes de frappe mineures a également été largement étudiée (Gan & Ng, 2019; von Geusau & Bloem, 2020; Matsuno & Tsuchiya, 2023; Mondal & Sancheti, 2024; Raj *et al.*, 2022a). Notamment, Elazar *et al.* (2021) et Raj *et al.* (2022b) ont découvert que les ML produisent des réponses différentes pour des requêtes factuelles sémantiquement équivalentes. De même, Hagen *et al.* (2024) ont découvert que les ML récents peuvent être impactés négativement par des fautes de frappe mineures qui préservent la sémantique d'origine.

Alignement temporel des connaissances dans les ML. L'état des connaissances factuelles étant en constante évolution, des études ont été menées pour comprendre comment adapter les ML face à cette évolution. Comme prévu, les ML se sont révélés incapables de prédire les faits futurs (Lazaridou *et al.*, 2021), ce qui implique la nécessité de les adapter pour maintenir leur alignement avec les connaissances actuelles. Pour pallier ce problème, des méthodes d'apprentissage continu (Liska *et al.*, 2022) ainsi que des techniques de pré-entraînement spécifiques ont été proposées, telles que la modélisation conjointe du texte et de son horodatage associé, facilitant l'acquisition de nouvelles connaissances temporelles (Dhingra *et al.*, 2022) ; des techniques d'édition de connaissances (Meng *et al.*, 2022; Hartvigsen *et al.*, 2023; Yu *et al.*, 2024; Zhang *et al.*, 2023) ; ou simplement l'externalisation des connaissances dans une base externe et accessible par le ML à travers la génération augmentée par récupération (Ram *et al.*, 2023). En parallèle, plusieurs jeux de données ont été proposés pour détecter les faits obsolètes dans les ML (Zhao *et al.*, 2024; Kim *et al.*, 2024; Margatina *et al.*, 2023; Kasai *et al.*, 2023; Mousavi *et al.*, 2024) ainsi que pour les mettre à jour (Ammar Khodja *et al.*, 2024; Yin *et al.*, 2024a; Thede *et al.*, 2025; Ge *et al.*, 2024).

1. github.com/Orange-OpenSource/TimeStress (Licence MIT)
2. huggingface.co/datasets/Orange/TimeStress (Licence CC BY-SA 4.0)

Raisonnement temporel dans les ML. Plusieurs études ont examiné les capacités de raisonnement temporel des ML (Zhang & Choi, 2021; Chu *et al.*, 2024; Wei *et al.*, 2023; Fatemi *et al.*, 2024; Dhingra *et al.*, 2022; Xiong *et al.*, 2024; Su *et al.*, 2024). Notamment, les travaux de Chen *et al.* (2021) et de Tan *et al.* (2023) ont chacun proposé un jeu de données dans lesquels les ML sont invités à répondre à des questions impliquant la compréhension de la temporalité des faits. Bien que ces travaux partagent des similitudes avec le nôtre en termes de données (des faits annotés temporellement), leurs objectifs et méthodologies diffèrent. Ces études testent la maîtrise de certains opérateurs de logique temporel (calculs de date, comparaison...) et évaluent une performance moyenne des ML selon un principe d'un test par fait. Pour notre part, nous nous intéressons non pas à la capacité de raisonnement mais à la robustesse des connaissances, c'est-à-dire à la capacité d'un ML de ne jamais commettre d'erreur lorsqu'un même fait lui est présenté avec des contextes temporels différents.

3 Problématique et jeu de données

L'objectif de l'article est de mesurer à quel point un modèle de langue est robuste au contexte temporel associé à un fait. Pour cela, le protocole expérimental proposé consiste à analyser les préférences du ML face à des contextes corrects ou incorrects pour un même fait. Cette section formalise tout d'abord cette problématique, puis présente le jeu de données TimeStress qui l'instancie.

3.1 Problématique

Faits et contextes temporels. De manière classique, nous considérons des *faits* comme des triplets RDF (sujet, relation, objet), notés (s, r, o), où les sujets et objets sont des entités ou littéraux et les relations émanent d'une ontologie (Petroni *et al.*, 2019; Elsahar *et al.*, 2018). Lorsqu'il s'agit de *faits temporels*, cette représentation est étendue pour inclure une période de validité $[a, b]$ (Yin *et al.*, 2024b; Jain *et al.*, 2020; Tan *et al.*, 2023). Pour un quintuplet (s, r, o, a, b), le sujet s est connecté à l'objet o via la relation r pendant la période allant de la date a à la date b. Par exemple, (Barack Obama, president, USA, 20 janvier 2009, 20 janvier 2017) est un fait temporel.

Nous définissons la notion de *contexte temporel* comme un intervalle de temps sur lequel nous souhaitons tester la validité d'un fait temporel. Pour réduire le nombre de possibilités et cadrer nos travaux, nous limitons ces intervalles de temps soit à des *années entières* (par ex., 1998, c.-à-d. tous les jours de l'année 1998), soit un *mois entier* d'une année donnée (novembre 1998), soit une *date précise* (15 novembre 1998). Par la suite, ces trois granularités distinctes seront notées A pour « Année », MA pour « Mois Année » et JMA pour « Jour Mois Année ».

Considérant un fait temporel $f = (s, r, o, a, b)$, un contexte temporel τ est dit *correct* pour f si τ est totalement inclus dans $[a, b]$ (c.à-d. $\tau \subseteq [a, b]$), *incorrect* s'il ne l'est pas du tout ($\tau \cap [a, b] = \varnothing$) ou *transitoire* sinon ($\tau \cap [a, b] \neq \varnothing$ et $\tau \not\subseteq [a, b]$).

Pour statuer sur la capacité d'un ML à distinguer un contexte correct τ^+ d'un contexte incorrect τ^- pour un fait temporel donné (s, r, o, a, b), deux énoncés textuels sont respectivement construits. La forme des énoncés que nous adoptons dans notre travail est celle d'une question portant sur le fait (s, r, o) suivie de sa réponse (« Quel est le r de s ? o ») et préfixée par une verbalisation du contexte temporel τ^+ ou τ^-. Sur l'exemple à propos de Barack Obama, deux contextes possibles sont par exemple $\tau^+ = 2011$ et $\tau^- = 1998$, produisant des énoncés « *In **2011**, who was the president of the*

USA? Barack Obama » et « *In **1998**, who was the president of the USA? Barack Obama* ».

Finalement, nous disons qu'un ML M distingue un contexte correct d'un contexte incorrect lorsqu'il associe une plus grande probabilité à la réponse o étant donné l'énoncé avec τ^+ en comparaison à un conditionnement sur τ^-, c'est-à-dire $\Pr_M(o|s,r,\tau^+) > \Pr_M(o|s,r,\tau^-)$. L'estimation globale de cette capacité consiste à considérer un grand ensemble de faits avec des entités, relations et période de validité variées, et à tester de nombreux couples (τ^+,τ^-) pour chaque fait. Pour rendre les résultats de ces matchs interprétables, nous imposons que les contextes d'un même couple soient de même granularité (A, MA, ou JMA).

Métriques. Nous introduisons deux métriques. Étant donné un fait f et un modèle M, nous exprimons d'une part les résultats *via* un *taux de victoire* $\mathcal{V}(M,f) \in [0,1]$ de M pour f, à savoir le rapport entre le nombre de fois où, pour le seul fait f, le modèle a préféré un contexte correct à un contexte incorrect sur le nombre de tests effectués. D'autre part, une autre métrique dite de *robustesse*, notée $\mathcal{R}(M,f)$ vérifie que les contextes corrects l'emportent systématiquement sur ceux incorrects, à savoir : $\mathcal{R}(M,f) = \mathbb{1}[\mathcal{V}(M,f) = 1]$ où $\mathbb{1}[]$ est la fonction indicatrice. Il est important de noter que les **contextes transitoires ne sont en aucun cas utilisées pour le calcul de ces métriques**, car leur validité est ambiguë. Étant donné un ensemble de faits, les taux de victoire moyens et la robustesse moyenne sont notées $\mathcal{V}(M)$ et $\mathcal{R}(M)$ respectivement.

Pour les besoins de segmentation des analyses, ces métriques globales peuvent être restreintes aux seuls tests effectués avec des contextes temporels d'une certaine granularité donnée (A, MA ou JMA).

Enfin, pour mesurer la distance d'un contexte τ par rapport à la période de validité $[a,b]$ d'un fait, nous calculons sa *position relative*, notée α, comme le nombre de jours écoulés entre le milieu de $[a,b]$ et le milieu de τ, divisé par le nombre de jours de $[a,b]$. Ainsi, on obtient $|\alpha| < \frac{1}{2}$ pour les contextes corrects, et $|\alpha| > \frac{1}{2}$ pour les contextes incorrects. Pour les contextes transitoires, la valeur $|\alpha|$ est explicitement fixée à $\frac{1}{2}$.

3.2 Le jeu de données TimeStress

Nous présentons le jeu de données TimeStress qui rend possible notre étude. Ce jeu comprend plus de 521 000 énoncés (sous forme de questions) générés à partir de 2 003 faits temporels, couvrant 1 883 entités uniques (1 385 sujets uniques et 1 113 objets uniques) et 86 relations. Un bref extrait en est donné dans la table 1.

Les faits sont issus d'une version post-traitée de Wikidata fournie par Ammar Khodja *et al.* (2025). Chaque fait (s,r,o,a,b) a une période de validité antérieure à 2021 ($b < 2021$) et est accompagné d'un indice de popularité calculé à partir du nombre de visites des articles Wikipédia des entités s et o sur une période récente de 12 mois. Bien que la popularité du sujet et de l'objet n'implique pas la popularité du fait, cet indice reste un outil intéressant pour trouver des faits « connus » par les ML, comme nous le verrons empiriquement dans les expériences. Seuls les 2 003 faits les plus populaires (selon cet indice) sont conservés afin de s'assurer que les ML les ont bien vus à l'apprentissage. Enfin, tous les faits ont une période de validité strictement supérieure à trois ans afin de garantir que le nombre de contextes corrects et incorrects est suffisamment grand pour qu'il soit quasiment impossible pour un modèle aléatoire de connaître un fait de manière robuste par hasard.

En moyenne, chaque fait est associé à 11 contextes temporels corrects et 74 incorrects, répartis sur

Fait temporel	Cont. temp.	Statut	Énoncé
(Betty Ford, spouse, Gerald Ford, 1948-10-15, 2006-12-26)	1983-03-21	Correct	On March 21, 1983, who was the spouse of Betty Ford? Gerald Ford
(Beirut, country, Ottoman Empire, 1520, 1918)	1759-05	Correct	In May 1759, to which sovereign state did Beirut belong? Ottoman Empire
(Jimmy Butler, member of sports team, Chicago Bulls, 2011, 2017-06-22)	1989-06-17	Incorrect	On June 17, 1989, which basketball team did Jimmy Butler belong to? Chicago Bulls
(Samarkand, country, Soviet Union, 1922-12-30, 1991-08-31)	1789-03-31	Incorrect	On March 31, 1789, what was the sovereign state of Samarkand? Soviet Union
(United States of America, head of government, Andrew Johnson, 1865-04-15, 1869-03-04)	1865	Transit.	In 1865, who served as the head of government for the United States of America? Andrew Johnson
(Chris Evans, unmarried partner, Minka Kelly, 2007-05, 2014-10)	2014	Transit.	In 2014, who was Chris Evans romantically involved with? Minka Kelly

TABLE 1 – Échantillon aléatoire d'énoncés produits à partir de divers faits et contextes temporels.

les trois granularités A, MA et JMA. Sur la base d'un intervalle de validité $[a, b]$ exprimé en années, centré sur $m = \frac{a+b}{2}$ et de durée $d = b - a$ [3], des contextes temporels sur la granularité de l'année sont tirés uniformément sur l'intervalle plus large $[m-5d, m+5d]$ avec un pas de $0.05 \times d$. À partir de ces contextes de granularité A, ceux de granularité MA sont générés en tirant un mois aléatoire. Puis, les contextes de granularité JMA sont déterminés en choisissant un jour aléatoire pour chaque contexte de granularité MA [4]. Cette opération engendre une hiérarchie entre les contextes issus d'une même année pour un même fait. Notons que, lorsqu'une date d_2 est choisie à partir d'une date d_1 de granularité supérieure, elle est nécessairement correcte (ou incorrecte) si d_1 l'est. Cependant, il est possible que d_1 soit transitoire, tandis que d_2 est correcte ou incorrecte. Dans ce cas, nous n'incluons pas d_2 dans l'ensemble des dates correctes ou incorrectes. De cette manière, il est **garantit** que le nombre de dates correctes et incorrectes ne varie pas selon la granularité considérée, ce qui introduirait un biais lors de la comparaison de la robustesse des modèles d'un niveau de granularité à l'autre. Les années des contextes temporels obtenues sont principalement situées dans la période contemporaine entre 1800 et 2020 (Annexe D), car l'indice de popularité utilisé pour sélectionner les faits dans TimeStress, tirent plus souvent des faits récents.

Les faits sont verbalisés en langage naturel sous forme de questions-réponses à l'aide de GPT-4o (une question par fait). Ces énoncés sont vérifiés manuellement sur un échantillon aléatoire. Il en ressort qu'ils sont de haute qualité, notamment exempts d'erreurs de syntaxe ou typographiques et systématiquement mis au passé. Pour produire les énoncés utilisés dans les matchs, chaque question est préfixée par un contexte temporel associé au fait considéré.

Une version détaillée du processus de création est disponible dans l'annexe A.

3. La médiane des dates (en précision jour) est utilisé pour effectuer les opérations arithmétiques entre dates.
4. Cet échantillonnage ne produit pas de dates erronées telles que le 29 février pour les années non-bissextiles, ou le 31 avril.

4 Expérimentation

Cette section détaille nos expériences sur le jeu de données TimeStress. Pour rappel, nos objectifs sont, dans l'ordre, de mesurer la capacité des modèles à distinguer des contextes temporels corrects et incorrects, d'analyser leur robustesse – à la recherche d'anomalies – pour cette tâche lorsque les contextes incorrects sont plus ou moins proche de l'intervalle de validité, puis que la granularité des contextes s'affine.

De nombreux modèles de familles et tailles différentes ont été testés :

— Mistral : `Mistral-Nemo-Base-2407`, `Mistral-7B-v0.3` (Jiang *et al.*, 2023) ;
— OpenEML : `OpenEML-{450M, 3B}` (Mehta *et al.*, 2024) ;
— Gemma2 : `gemma-2-{2b, 9b, 27b}` (Team *et al.*, 2024) ;
— Llama3.1 : `Llama-3.1-{8B, 70B}` (Grattafiori *et al.*, 2024).

Pour chacun, les versions simplement pré-entraînées et également instruites sont considérées, pour un total de 18 ML. Tous les modèles sont ceux fournis *via* huggingface.co.

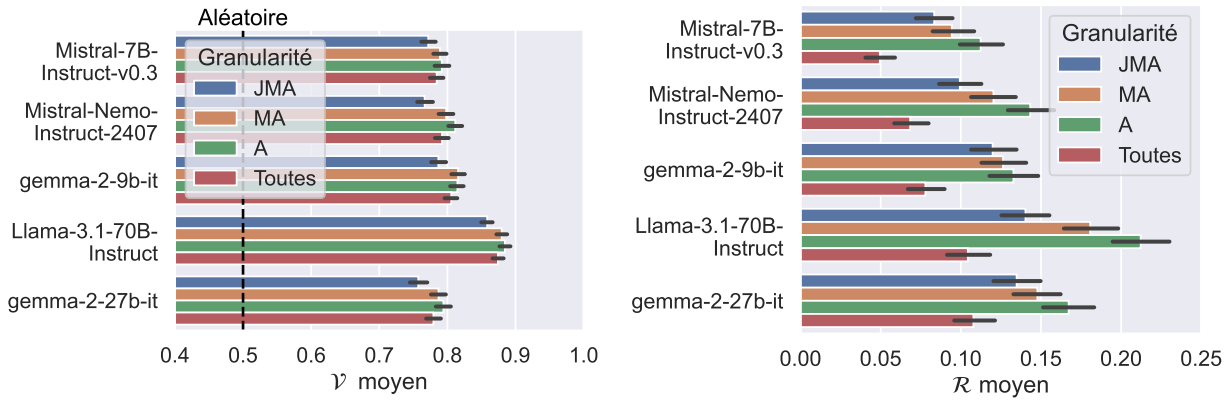
Dans une première séries d'expérience, les énoncés sont passés aux modèles sous la forme de texte brut et non de messages afin de mettre les modèles seulement pré-entrainés et ceux instruits sur un pied d'égalité. L'emploi d'un format « instructions/messages » est exploré dans un deuxième temps.

4.1 Maîtrise globale des contextes temporels

La Figure 2a présente le taux de victoire moyen pour les faits de TimeStress pour les 5 meilleurs modèles et pour chaque granularité temporelle A, MA et JMA, ainsi que pour leur union. Les résultats des autres modèles sont reportés dans l'annexe D. Dans l'ensemble, les résultats montrent que ces 5 meilleurs ML distinguent globalement bien les énoncés corrects des incorrects avec des taux de victoire entre 78% et 87%. Parmi nos autres résultats, nous avons constaté que même les modèles plus petits (<500M paramètres) font mieux que le hasard et que le taux de victoire s'améliore assez logiquement avec la taille du modèle, le meilleur modèle étant le plus gros, Llama-3.1-70B-Instruct.

La figure 3 permet une analyse plus fine, en rapportant la moyenne de $\log \Pr(o|f,\tau)$ en fonction de la valeur α, qui quantifie de manière normalisée la distance de τ à la période de validité de f (voir Section 3.1). La moyenne est calculée sur l'ensemble des faits, pour une granularité des contextes à l'année et sur l'ensemble des 18 ML étudiés. Nous observons bien que les probabilités les plus hautes sont celles des contextes dans l'intervalle de validité ($\alpha \in [-0.5, 0.5]$), tandis qu'à l'extérieur de cet intervalle, les probabilités décroissent graduellement à mesure qu'$|\alpha|$ augmente. Enfin, nous notons que la probabilité attribuée aux contextes transitoires (des années ni correctes ni incorrectes) est significativement plus élevée que pour les contextes corrects (en se basant sur les intervalles de confiance). Nous expliquons ce phénomène par l'hypothèse suivante : dans les données d'entraînement des ML, les années transitoires sont plus souvent associées au fait considéré que les autres années de la période validité, car elles correspondent aux faits marquant que sont le début et la fin du fait (par exemple, l'année du début ou de fin d'un mandat présidentiel).

Cet alignement élevé des ML avec la période de validité des faits temporels nous amène à conclure que les ML possèdent au moins une représentation de base de la temporalité.



(a) Taux de victoire moyen

(b) Robustesse moyenne

FIGURE 2 – Métriques moyennes sur l'ensemble de fait dans TimeStress pour les 5 meilleurs ML avec intervalles de confiance à 95% (bootstrap).

4.2 Robustesse et anomalies

La représentation temporelle des ML n'est pas robuste. La figure 2b présente la robustesse moyenne des 5 meilleurs modèles sur tous les faits de TimeStress. Pour rappel, cette métrique est plus stricte et ne tolère aucune défaite lors des matchs sur un fait donné. Les scores sont globalement faibles, ceci indique que les taux de victoire fait par fait atteignent rarement un score de 100%. Il est également intéressant de noter que le modèle le plus robuste n'est pas celui avec le plus haut taux de victoire Le modèle le plus robuste est en l'occurrence gemma-2-27b-it. Sa valeur $\mathcal{R}$ atteint seulement environ 17% pour la granularité la plus grossière A. Ce score chute à 11% lorsque toutes les granularités sont prises en compte. La plupart des autres modèles ne dépassent pas un score global de robustesse de 3%. Parmi nos autres résultats, nous avons également pu observer que les modèles instruits font la plupart du temps mieux que leurs homologues seulement pré-entraînés. Un cas notable est le modèle Llama-3.1-70B-Instruct ; bien qu'il ait été affiné sur des instructions, il est $3.6\times$ plus robuste que son homologue pré-entraîné Llama-3.1-70B. Cela suggère que les données d'entraînement et possiblement la procédure d'entraînement jouent un rôle important dans la robustesse temporelle. Enfin, des signes précoces d'échec dans le transfert de connaissances entre les granularités sont évidents en raison de l'écart substantiel entre les scores de robustesse individuels des granularités et le score global. Ce problème est exploré plus en détail plus loin dans cette section.

Les ML sont vulnérables aux contextes incorrects *faciles*. La table 2 investigue l'impact des positions relatives des contextes incorrects pour la granularité A en se concentrant sur les seuls cas de contextes incorrects qui mettent en défaut un ML dans un match pour des faits qui semblent pourtant bien « connus » des ML car avec un taux de victoire très élevé ($\mathcal{V} \geq 95\%$). Pour l'instant, seul la colonne « texte brut » nous intéresse. La table révèle que ces contextes incorrects ne sont pas entièrement concentrés autour de la période de validité comme il pourrait sembler raisonnable de l'imaginer. Au lieu de cela, une proportion non négligeable se situe loin de celle-ci. Plus précisément, les ML échouent à atteindre la robustesse en raison de dates avec une distance de $|\alpha| \geq 1$ dans 19% des cas. Cette proportion diminue à 6% pour $|\alpha| \geq 3$, ce qui reste important étant donné la proximité du taux de victoire à 100% pour les faits ici observés. Nous avons mené la même analyse en utilisant des seuils de victoire autres que 95% (voir l'annexe B). À mesure que le seuil tend vers 100%, la vulnérabilité aux dates incorrectes faciles diminue progressivement mais ne disparaît jamais

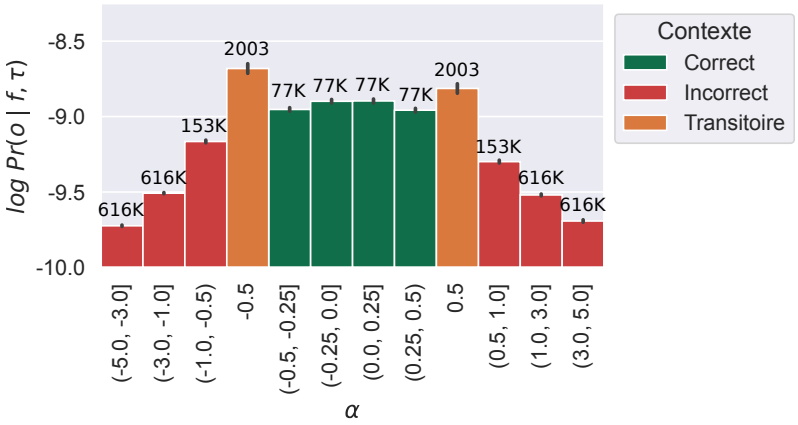


FIGURE 3 – Évolution de $\log \Pr(o|f,d)$ par rapport à la distance relative α, moyennée sur tous les faits de TimeStress et sur tous les ML, pour la granularité A avec les intervalles de confiance à 95% (bootstrap). Le nombre de points utilisés pour calculer chaque barre est indiqué au-dessus de celle-ci.

$\lvert\alpha\rvert$	Texte brut	Instruction
≥ 1	0.19±0.01	0.25±0.01
≥ 2	0.09±0.01	0.13±0.01
≥ 3	0.06±0.01	0.08±0.01
≥ 4	0.04±0.01	0.05±0.01

TABLE 2 – Proportion des dates incorrectes favorisées par rapport aux dates correctes situées au-delà d'une distance relative $|\alpha|$, lorsque le taux de victoire dépasse 95% (Intervalle de confiance de Wilson à 95%).

complètement. Même lorsque le seuil de taux de victoire est de 99%, des erreurs subsistent quand $|\alpha| \geq 4$. Nous pouvons en conclure que cette vulnérabilité est inhérente aux ML actuels. Bien qu'il puisse être argué de la nature probabiliste de ces modèles comme une explication tangible, il s'agit d'un comportement clairement non souhaitable car il s'agit typiquement d'erreurs qu'un humain ne commettrait pas lorsqu'il connaît la période de validité d'un fait.

Ces conclusions sont valables pour des instruction. Jusqu'à présent dans nos expérimentations, tous les modèles ont été alimentés par des énoncés sous la forme de texte brut et non d'instructions. Comme les performances de ML instruits pourraient s'en avoir été sous-estimées, les taux de victoire et les scores de robustesse ont été recalculés en utilisant un format « instructions/messages » [5]. La figure 4 compare les scores de robustesse calculés sur les 2 formats. En moyenne, la robustesse décroît avec l'emploi du format « instruction » (en particulier, les modèles gemma-2) et les scores de robustesse globaux restent très faibles. Cependant, aucune conclusion claire n'émergent quant à l'incidence positive ou négative de ce format car l'impact est très différent d'un modèle à un autre. Ensuite, la colonne « instruction » de la table 2 complète notre précédente analyse sur l'impact de la position relative des contextes incorrects pour les faits à taux de victoire élevé. Il ressort cette fois que le format « instruction » dégrade les performances avec davantage d'erreurs critiques (c.-à-d. éloignées de la période de validité). En se basant sur les intervalles de confiance, ces différences sont statistiquement significatives pour toutes les valeurs d'$|\alpha|$ étudiées. Des exemples de ces erreurs critiques sont montrés en annexe D.

Les ML échouent à propager parfaitement leurs connaissances entre granularités. Nous examinons la capacité des ML à propager la connaissance d'un fait entre différentes granularités temporelles. TimeStress permet des comparaisons entre deux granularités car les trois granularités étudiées ont le même nombre de dates correctes et incorrectes pour tous les faits temporels. La seule différence entre deux granularités est l'ajout d'un mois et/ou d'un jour aléatoire, ce qui n'affecte pas

5. Ceci passe par la construction de message et leur injection dans le gabarit de discussion (en anglais, *chat template*) de chaque ML. Les messages sont construits comme sur l'exemple suivant : `{user: "In 2011, who was the president of the USA?", assistant: "Barack Obama"}`.

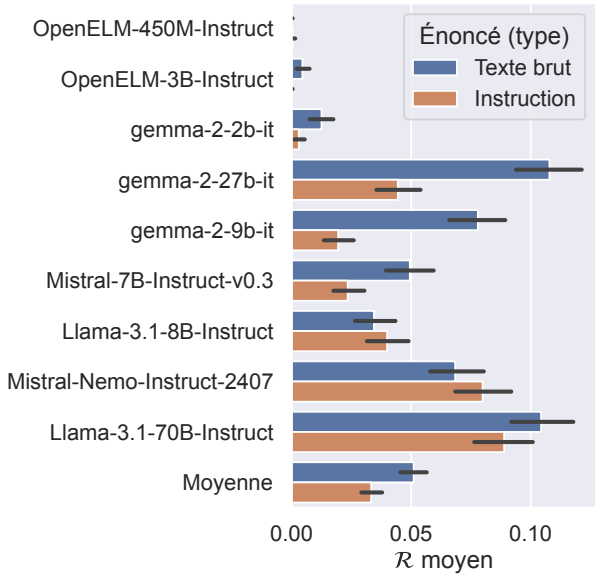


FIGURE 4 – Moyenne de $\mathcal{R}$ sur toutes les granularités des faits de TimeStress suivant le format des énoncés soumis aux modèles : texte brut (bleu) ou instruction (orange). Les intervalles de confiance à 95% (bootstrap) sont affichés.

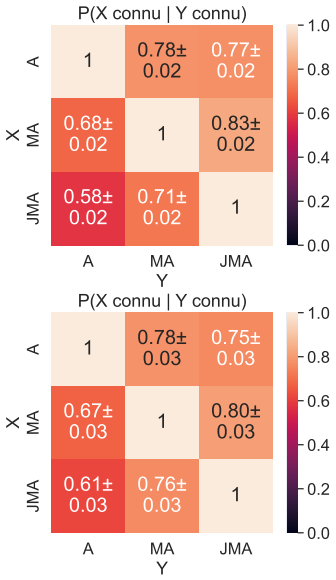


FIGURE 5 – Taux de succès moyen du transfert de connaissances entre paires de granularités sur les 5 ML les plus robustes avec des requêtes sous forme de texte brut (haut) ou d'instructions (bas). Les intervalles de confiance de Wilson à 95% sont indiqués.

la validité en passant d'une granularité inférieure à une granularité supérieure (par ex., de A vers MA). Par exemple, si un fait est incorrect pour toute une année, il l'est pour tout mois et date de cette année.

Nous considérons qu'un fait f est « *connu* » pour une granularité par un modèle M si $\mathcal{R}(M, f) = 1$. Cette définition peut être spécifique à une granularité donnée. Par exemple, un fait est connu à la granularité de l'année si tous les matchs avec des contextes temporels à la granularité de l'année ont été gagnés. Pour chacun des 5 ML les plus robustes et pour chaque paire de granularités (X, Y), nous calculons alors la proportion de faits « connus » à la granularité X, étant donné qu'ils sont « connus » à la granularité Y. La figure 5 rapporte cette proportion de transfert de la granularité Y vers X pour le format « texte brut » (en haut) et « instruction ». En moyenne pour le format « texte brut », les ML n'ont pas été capables de généraliser leurs connaissances à d'autres granularités dans 28% des cas (moyenne de toutes les cases, hors diagonale), ce qui est étonnamment élevé étant donné leur score parfait sur la granularité Y de départ. Les détails pour chaque modèle sont disponibles dans l'annexe C. Les performances varient entre les ML. Par exemple, pour le modèle le plus robuste gemma-2-27b-it, la transition de $Y = \text{A}$ vers $X = \text{MA}$ est réussie dans 74±5% des cas et les taux de succès des autres transitions varient entre 68±6% et 88±5%. La tendance générale est que les ML échouent plus dans les transitions allant de granularités grossières à fines. Aucun ML n'affiche de transition parfaite pour quelque paire de granularités que ce soit. Il y a de légères variations entre les formats « instruction » (Figure 5, en bas) et « brut », mais le taux de succès moyen est quasi-identique.

Vu la possibilité que la mauvaise propagation des connaissances entre granularités soit due à la méconnaissance des frontières de la période de validité par les ML [6], une analyse similaire a été menée qui prend en compte la position du contexte (Annexe C.1). La conclusion est que **l'incohérence globale est principalement due à la méconnaissance des frontières de la période de validité par**

6. Dans ce cas, la robustesse n'a été atteinte que par chance.

les ML ; en effet, la cohérence entre granularités s'approche sensiblement d'une cohérence parfaite à force que le contexte s'éloigne de la période de validité **mais ne l'atteint jamais** ; ce qui rappelle la vulnérabilité des ML face à des contextes incorrectes *faciles*.

À des fins d'exploration, nous avons étudié si l'inclusion dans l'amorce des ML des explications sur les concepts temporels pourrait les aider à mieux transférer les connaissances d'une granularité temporelle à une autre. Pour évaluer cela, deux amorces ont été préfixées à chaque énoncé de TimeStress. La première explique la nature hiérarchique des dates (i.e., une année est composée de mois et un mois est composé de jours), tandis que la seconde est plus directe et explique comment les connaissances sur un fait temporel peuvent être généralisées d'une granularité à une autre. Leur détail est donné dans l'annexe C.3. Nous avons recalculé les proportions de transfert entre granularités sur les mêmes 5 ML utilisés que deux de la figure 5. Les deux amorces explicatives ont amélioré la généralisation au format « texte brut » de 73% à 76%. Cependant, aucun gain substantiel par rapport à ne pas utiliser d'amorce explicative n'est observé lorsqu'on utilise le format « instruction ».

Autres observations. Il y a une corrélation positive entre la popularité d'un fait et la robustesse des ML sur celui-ci, avec un coefficient de Pearson de $+0.065$ et une p-value $< 10^{-51}$. D'autre part, les ML sont robustes sur des faits globalement différents. En effet, une paire de ML partagent 11% de faits sur lesquels ils sont robustes en moyenne. Cette proportion atteint 31% en se limitant aux 5 ML les plus robustes (c.f. figure 2b). Cependant, dans ce cas, il n'y a que 34 faits sur 384 (8.9%) qui soient robustes en même temps sur ces ML. Par ailleurs, plus un fait possède une période de validité longue, plus la robustesse est faible et plus le taux de victoire est haut (5 ML les plus robustes). Ces deux corrélations, statistiquement significatives [7], sont intrigantes, car il semble que la difficulté de situer un fait dans le temps est la même suivant qu'il ait une durée de 3 ans ou 30 ans. Bien qu'il a été noté que l'indice de popularité d'un fait est plus faible quand sa durée est grande, ceci n'explique pas l'orientation opposée des corrélations. Enfin, plus la période de validité d'un fait est loin du présent, moins les MLs sont robustes dessus avec des taux de victoire moindre aussi. Plus de détails sont en annexe D.

5 Discussion et conclusion

Cette étude a exploré la robustesse des ML face aux variations temporelles dans les connaissances factuelles. En analysant leur capacité à différencier des contextes temporels corrects et incorrects, les performances des ML ont été évaluées selon deux dimensions : la distance des contextes de la période de validité des faits et leur granularité. À cette fin, le jeu de données TimeStress a été introduit, permettant de tester 18 modèles de taille et familles variés. Les résultats montrent que le meilleur ML n'est robuste que pour 11% des faits étudiés, avec des erreurs, certes rares, mais critiques qu'un humain ne commettrait pas. Nous notons notamment la vulnérabilité des ML face à des contextes incorrects *faciles* et des lacunes au niveau de la généralisation des connaissances factuelles entre granularités. Ces résultats soulignent les limites des ML actuels dans leur représentation temporelle. Il convient de mentionner que, puisque les faits temporels étudiés sont relativement populaires, ces performances majorent probablement les performances des ML sur la population générale des faits, étant donné la forte relation entre la popularité des connaissances et la probabilité qu'elles soient apprises par les ML (Kandpal *et al.*, 2023; Kang & Choi, 2023). L'annexe E présente une discussion sur les limites de notre travail.

7. L'hypothèse nulle pour chaque métrique étant qu'une corrélation n'existe pas.

Références

AMMAR KHODJA H., AIT GUENI SSAID A., BECHET F., BRABANT Q., NASR A. & LECORVÉ G. (2025). Factual knowledge assessment of language models using distractors. 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. 8043–8056, Abu Dhabi, UAE : Association for Computational Linguistics.

AMMAR KHODJA H., BÉCHET F., BRABANT Q., NASR A. & LECORVÉ G. (2024). WikiFactDiff : A large, realistic, and temporally adaptable dataset for atomic factual knowledge update in causal language models. 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. 17614–17624, Torino, Italia : ELRA and ICCL.

CHEN W., WANG X. & WANG W. Y. (2021). A dataset for answering time-sensitive questions. In J. VANSCHOREN & S. YEUNG, Éds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

CHU Z., CHEN J., CHEN Q., YU W., WANG H., LIU M. & QIN B. (2024). TimeBench : A comprehensive evaluation of temporal reasoning abilities in large language models. 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. 1204–1228, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.66.

DHINGRA B., COLE J. R., EISENSCHLOS J. M., GILLICK D., EISENSTEIN J. & COHEN W. W. (2022). Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, **10**, 257–273. DOI : 10.1162/TACL_A_00459.

DONG Q., XU J., KONG L., SUI Z. & LI L. (2023). Statistical knowledge assessment for large language models. In A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT & S. LEVINE, Éds., *Advances in Neural Information Processing Systems 36 : Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

ELAZAR Y., KASSNER N., RAVFOGEL S., RAVICHANDER A., HOVY E. H., SCHÜTZE H. & GOLDBERG Y. (2021). Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguistics*, **9**, 1012–1031. DOI : 10.1162/TACL_A_00410.

ELSAHAR H., VOUGIOUKLIS P., REMACI A., GRAVIER C., HARE J., LAFOREST F. & SIMPERL E. (2018). T-REx : A large scale alignment of natural language with knowledge base triples. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

FATEMI B., KAZEMI M., TSITSULIN A., MALKAN K., YIM J., PALOWITCH J., SEO S., HALCROW J. & PEROZZI B. (2024). Test of time : A benchmark for evaluating llms on temporal reasoning.

GAN W. C. & NG H. T. (2019). Improving the robustness of question answering systems to question paraphrasing. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy,*

July 28- August 2, 2019, Volume 1 : Long Papers, p. 6065–6075 : Association for Computational Linguistics. DOI : 10.18653/V1/P19-1610.

GE X., MOUSAVI A., GRAVE E., JOULIN A., QIAN K., HAN B., AREFIYAN M. & LI Y. (2024). Time sensitive knowledge editing through efficient finetuning. In L. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, p. 583–593 : Association for Computational Linguistics.

GRATTAFIORI A., DUBEY A. & ET AL. A. J. (2024). The llama 3 herd of models.

HAGEN T., SCELLS H. & POTTHAST M. (2024). Revisiting query variation robustness of transformer models. In Y. AL-ONAIZAN, M. BANSAL & Y. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, p. 4283–4296 : Association for Computational Linguistics.

HARTVIGSEN T., SANKARANARAYANAN S., PALANGI H., KIM Y. & GHASSEMI M. (2023). Aging with GRACE : lifelong model editing with discrete key-value adaptors. In A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT & S. LEVINE, Éds., *Advances in Neural Information Processing Systems 36 : Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

JAIN P., RATHI S., MAUSAM & CHAKRABARTI S. (2020). Temporal Knowledge Base Completion : New Algorithms and Evaluation Protocols. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 3733–3747, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.305.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.

JIANG Z., XU F. F., ARAKI J. & NEUBIG G. (2020). How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, **8**, 423–438. DOI : 10.1162/TACL_A_00324.

KANDPAL N., DENG H., ROBERTS A., WALLACE E. & RAFFEL C. (2023). Large language models struggle to learn long-tail knowledge. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 de *Proceedings of Machine Learning Research*, p. 15696–15707 : PMLR.

KANG C. & CHOI J. (2023). Impact of co-occurrence on factual knowledge of large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 7721–7735, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.518.

KASAI J., SAKAGUCHI K., TAKAHASHI Y., BRAS R. L., ASAI A., YU X., RADEV D., SMITH N. A., CHOI Y. & INUI K. (2023). Realtime QA : what's the answer right now ? In A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT & S. LEVINE, Éds., *Advances in Neural Information Processing Systems 36 : Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

KASSNER N. & SCHÜTZE H. (2020). Negated and misprimed probes for pretrained language models : Birds can talk, but cannot fly. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7811–7818, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.698.

KIM Y., YOON J., YE S., BAE S., HO N., HWANG S. J. & YUN S. (2024). Carpe diem : On the evaluation of world knowledge in lifelong language models. In K. DUH, H. GÓMEZ-ADORNO & 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, p. 5401–5415 : Association for Computational Linguistics. DOI : 10.18653/V1/2024.NAACL-LONG.302.

LAZARIDOU A., KUNCORO A., GRIBOVSKAYA E., AGRAWAL D., LISKA A., TERZI T., GIMÉNEZ M., DE MASSON D'AUTUME C., KOCISKÝ T., RUDER S., YOGATAMA D., CAO K., YOUNG S. & BLUNSOM P. (2021). Mind the gap : Assessing temporal generalization in neural language models. In *Neural Information Processing Systems*.

LISKA A., KOCISKÝ T., GRIBOVSKAYA E., TERZI T., SEZENER E., AGRAWAL D., DE MASSON D'AUTUME C., SCHOLTES T., ZAHEER M., YOUNG S., GILSENAN-MCMAHON E., AUSTIN S., BLUNSOM P. & LAZARIDOU A. (2022). Streamingqa : A benchmark for adaptation to new knowledge over time in question answering models. In K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVÁRI, G. NIU & S. SABATO, Éds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 de *Proceedings of Machine Learning Research*, p. 13604–13622 : PMLR.

LYU C., WU M. & AJI A. (2024). Beyond probabilities : Unveiling the misalignment in evaluating large language models. In S. LI, M. LI, M. J. ZHANG, E. CHOI, M. GEVA, P. HASE & H. JI, Éds., *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, p. 109–131, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.knowllm-1.10.

MARGATINA K., WANG S., VYAS Y., JOHN N. A., BENAJIBA Y. & BALLESTEROS M. (2023). Dynamic benchmarking of masked language models on temporal concept drift with multiple views. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, p. 2873–2890 : Association for Computational Linguistics. DOI : 10.18653/V1/2023.EACL-MAIN.211.

MATSUNO T. & TSUCHIYA M. (2023). Evaluating the robustness of question answering model against context variations. In *2023 10th International Conference on Advanced Informatics : Concept, Theory and Application (ICAICTA)*, p. 1–6. DOI : 10.1109/ICAICTA59291.2023.10390252.

MEHTA S., SEKHAVAT M., CAO Q., HORTON M., JIN Y., SUN F., MIRZADEH I., NAJIBIKOHNEHSHAHRI M., BELENKO D., ZATLOUKAL P. & RASTEGARI M. (2024). Openelm : An efficient language model family with open training and inference framework. In *ICML Workshop*.

MENG K., BAU D., ANDONIAN A. & BELINKOV Y. (2022). Locating and editing factual associations in GPT. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems 35 : Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

MONDAL I. & SANCHETI A. (2024). On the robustness of chatgpt under input perturbations for named entity recognition task. In *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024* : OpenReview.net.

MOUSAVI S. M., ALGHISI S. & RICCARDI G. (2024). Dyknow : Dynamically verifying time-sensitive factual knowledge in llms. In Y. AL-ONAIZAN, M. BANSAL & Y. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, p. 8014–8029 : Association for Computational Linguistics.

PETRONI F., ROCKTÄSCHEL T., RIEDEL S., LEWIS P., BAKHTIN A., WU Y. & MILLER A. (2019). Language models as knowledge bases ? In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *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. 2463–2473, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1250.

RAJ H., ROSATI D. & MAJUMDAR S. (2022a). Measuring reliability of large language models through semantic consistency. In *NeurIPS ML Safety Workshop*.

RAJ H., ROSATI D. & MAJUMDAR S. (2022b). Measuring reliability of large language models through semantic consistency. *CoRR*, **abs/2211.05853**. DOI : 10.48550/ARXIV.2211.05853.

RAM O., LEVINE Y., DALMEDIGOS I., MUHLGAY D., SHASHUA A., LEYTON-BROWN K. & SHOHAM Y. (2023). In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, **11**, 1316–1331. DOI : 10.1162/TACL_A_00605.

SU Z., LI J., ZHANG J., ZHU T., QU X., ZHOU P., BOWEN Y., CHENG Y. & ZHANG M. (2024). Living in the moment : Can large language models grasp co-temporal reasoning ? 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. 13014–13033, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.703.

SUN K., XU Y. E., ZHA H., LIU Y. & DONG X. L. (2024). Head-to-tail : How knowledgeable are large language models (llms) ? A.K.A. will llms replace knowledge graphs ? In K. DUH, H. GÓMEZ-ADORNO & 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, p. 311–325 : Association for Computational Linguistics. DOI : 10.18653/V1/2024.NAACL-LONG.18.

TAN Q., NG H. T. & BING L. (2023). Towards benchmarking and improving the temporal reasoning capability of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

TEAM G., RIVIERE M. & ET EL. S. P. (2024). Gemma 2 : Improving open language models at a practical size.

THEDE L., ROTH K., BETHGE M., AKATA Z. & HARTVIGSEN T. (2025). Understanding the limits of lifelong knowledge editing in llms.

VON GEUSAU P. A. & BLOEM P. (2020). Evaluating the robustness of question-answering models to paraphrased questions. In M. BARATCHI, L. CAO, W. A. KOSTERS, J. LIJFFIJT, J. N. VAN RIJN & F. W. TAKES, Éds., *Artificial Intelligence and Machine Learning - 32nd Benelux Conference, BNAIC/Benelearn 2020, Leiden, The Netherlands, November 19-20, 2020, Revised Selected Papers*, volume 1398 de *Communications in Computer and Information Science*, p. 1–14 : Springer. DOI : 10.1007/978-3-030-76640-5_1.

WEI Y., SU Y., MA H., YU X., LEI F., ZHANG Y., ZHAO J. & LIU K. (2023). Menatqa : A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Conference on Empirical Methods in Natural Language Processing*.

XIONG S., PAYANI A., KOMPELLA R. & FEKRI F. (2024). Large language models can learn temporal reasoning. 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. 10452–10470, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.563.

YIN X., JIANG J., YANG L. & WAN X. (2024a). History matters : Temporal knowledge editing in large language model. In M. J. WOOLDRIDGE, J. G. DY & S. NATARAJAN, Éds., *Thirty-Eighth*

AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, p. 19413–19421 : AAAI Press. DOI : 10.1609/AAAI.V38I17.29912.

YIN X., JIANG J., YANG L. & WAN X. (2024b). History matters : Temporal knowledge editing in large language model. In M. J. WOOLDRIDGE, J. G. DY & S. NATARAJAN, Éds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, p. 19413–19421 : AAAI Press. DOI : 10.1609/AAAI.V38I17.29912.

YU L., CHEN Q., ZHOU J. & HE L. (2024). MELO : enhancing model editing with neuron-indexed dynamic lora. In M. J. WOOLDRIDGE, J. G. DY & S. NATARAJAN, Éds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, p. 19449–19457 : AAAI Press. DOI : 10.1609/AAAI.V38I17.29916.

ZHANG M. & CHOI E. (2021). SituatedQA : Incorporating extra-linguistic contexts into QA. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7371–7387, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.586.

ZHANG Z., FANG M., CHEN L., NAMAZI-RAD M.-R. & WANG J. (2023). How do large language models capture the ever-changing world knowledge ? a review of recent advances. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 8289–8311, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.516.

ZHAO B., BRUMBAUGH Z., WANG Y., HAJISHIRZI H. & SMITH N. A. (2024). Set the clock : Temporal alignment of pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

A TimeStress : Détails du processus de construction

Cette section fournit une description détaillée de la construction du jeu de données TimeStress. Avant d'aborder le processus de collecte, nous décrivons les principales caractéristiques de TimeStress. Tout d'abord, le jeu de données se concentre sur des faits passés valables strictement avant 2021, garantissant qu'ils sont des événements passés pour tous les ML récents. TimeStress inclut des énoncés de haute qualité, cohérentes avec les faits et présentant une diversité linguistique pour éviter les biais dus à une variété limitée de questions. Les énoncés sont soigneusement sélectionnés pour minimiser les fautes de frappe, les verbes sont systématiquement conjugués au passé et excluent les dates futures au-delà de 2020 afin d'éviter des questions absurdes, telles que « *In 2052, who was the president of USA?* ». Le jeu de données couvre un ensemble diversifié de 86 relations pour réduire les biais liés à un éventail restreint. Les faits ciblés sont populaires, ce qui est essentiel pour évaluer la généralisation des connaissances à différentes granularités, une tâche qui devient difficile si les ML ne connaissent pas les faits. Tous les faits ne sont valides que sur une unique période de validité, afin de pouvoir considérer tous les contextes en dehors de la période de validité sont comme étant

incorrectes. De plus, pour garantir l'équité, chaque granularité (A,MA,JMA) a un nombre égal de contextes temporels corrects et incorrectes pour tous les faits. Enfin, le nombre de contextes corrects et incorrects est suffisamment grand pour qu'il soit quasiment impossible pour un modèle aléatoire de connaître un fait de manière robuste par hasard.

Le processus de création du jeu de données TimeStress a été conçu avec soin pour répondre aux propriétés décrites précédemment, soutenant ainsi efficacement les affirmations de cet article. Ce processus comprend trois étapes principales. Premièrement, une collecte initiale de 2098 faits temporels est réalisée à partir de Wikidata pour inclusion dans TimeStress. Deuxièmement, des questions sont générées à partir de ces quintuplets en utilisant GPT-4o, accompagnées d'une évaluation de la qualité pour garantir des questions de haute qualité. Enfin, pour chaque fait, les dates correctes et incorrectes sont identifiées et intégrées aux questions pour produire des énoncés.

A.1 Processus de collecte des quintuplets

Le processus de collecte des quintuplets commence par la collecte de la version post-traitée de Wikidata fournie par (Ammar Khodja *et al.*, 2025).

Cette source fournit également une mesure de popularité d'une entité définie comme le nombre médian de visites humaines à l'article Wikipédia associé à cette entité au cours de l'année 2020. Cette mesure est utilisée pour définir la popularité d'un quintuplet, calculée comme la moyenne géométrique de la popularité de son objet et de son sujet. La figure 12 fournit une démonstration simple de l'efficacité de cette mesure de popularité pour trouver des faits sur lesquels les ML sont robustes, illustrant que la probabilité que les ML soient robuste sur un fait augmente avec sa popularité.

Initialement, tous les quintuplets ayant au moins une date de début ou de fin et dont les objets ne sont pas des littéraux, tels que des quantités et des dates, sont collectés, totalisant plus de 2.1 millions de quintuplets. Les quintuplets sont ensuite filtrés pour supprimer tout (s, r, o, a, b) où un autre quintuplet (s, r, o, a', b') existe avec une période de validité différente $[a', b']$, permettant de supposer que toutes les dates en dehors de $[a', b']$ sont incorrectes, ce qui simplifie l'analyse des résultats. Cette étape élimine une quantité négligeable de quintuplets (6.23%). De plus, les quintuplets sans date de début ou de fin sont supprimés car leur période de validité n'est pas bornée. Seuls les quintuplets ayant une mesure de popularité d'au moins 90 000 [8] et une période de validité strictement supérieure à trois ans sont conservés. Le résultat final est un jeu de données comprenant les 2 098 faits les plus populaires de Wikidata (selon l'indice de popularité utilisé), avec 1 910 entités uniques, 1 435 sujets uniques, 1 151 objets uniques et 86 relations, constituant un ensemble bien diversifié de faits temporels.

A.2 Verbalisation des quintuplets

Le processus de verbalisation des quintuplets en questions en langage naturel est effectué à l'aide de GPT-4o. L'amorce, adaptée de (Ammar Khodja *et al.*, 2024) (Annexe B), a été modifiée pour générer des questions plutôt que des phrases déclaratives. L'amorce *système* adaptée demande à GPT-4o de prendre un tuple (sujet, relation, objet, horodatage) et de générer quatre questions linguistiquement diverses. Par exemple, pour l'entrée `(British India, capital, Kolkata, 1929)`, une question possible pourrait être : « *In 1929, what was the capital of British India? Kolkata* ». Les

8. Ce seuil a été déterminé en diminuant progressivement le seuil à partir de 150 000 par pas de 10 000 jusqu'à ce que le nombre de faits récupérés dépasse 2 000.

questions doivent respecter des directives spécifiques : elles doivent être au passé, commencer par l'année suivie d'une virgule, et se terminer par la réponse. Les questions doivent se concentrer sur l'objet, être simples et concises, et éviter tout détail qui pourrait simplifier la réponse.

Voici l'amorce *système* utilisée :

```
You are an advanced knowledge verbalization system.
You take as input a knowledge quadruple (subject, relation, object, time) and generate
a list of 4 linguistically diverse questions on the quadruple.
For example, the input could be : (British India, capital, Kolkata, 1929) and one of
your questions may be : "In 1929, what was the capital of British India? Kolkata.".

All the questions you generate must be in past tense because the facts are not valid
anymore.
The questions must always start with the year, then a comma, then the question itself,
and then finally the answer.
The questions must always be asked on the object.
The questions must be straightforward and concise.
The questions must not contain details that could make them easier to answer.

Examples of questions:
- (Jimmy Butler, member of sports team, Chicago Bulls, 2014) -> "In 2014, which team
did Jimmy Butler play for? Chicago Bulls."
- (Philippines, head of state, Emilio Aguinaldo, 1900) -> "In 1900, who was the head of
state of Philippines? Emilio Aguinaldo."
- (Coretta Scott King, spouse, Martin Luther King Jr., 1960) -> "In 1960, who was
Coretta Scott King married to? Martin Luther King Jr."
- (European Union, currency, pound sterling, 2002) -> "In 2002, what was one of the
currencies of the European Union? Pound sterling."
```

Et voici l'amorce principale :

```
Here is the knowledge quadruple to verbalize: ([SUBJECT], [RELATION], [OBJECT],
[YEAR]).

Due to the ambiguity that could arise from the provided labels, here is their meaning:
- (subject) "[SUBJECT]" : "[SUBJECT_DESC]"
- (relation) "[RELATION]" : "[RELATION_DESC]"
- (object) "[OBJECT]" : "[OBJECT_DESC]"

Finally, here is an example where the relation "[RELATION]" is employed :
([EXAMPLE_SUBJECT], [RELATION], [EXAMPLE_OBJECT]).
```

Pour utiliser cette dernière amorce, il suffit de remplir les espaces réservés [SUBJECT], [RELATION], [OBJECT], [SUBJECT_DESC], [RELATION_DESC] et [OBJECT_DESC] avec les libellés et descriptions correspondants de Wikidata. Un exemple de relation est également récupéré de Wikidata en utilisant la relation *Wikidata property example (P1855)*. Si aucun exemple n'est disponible, la dernière ligne de l'amorce principale est omise. L'année [YEAR] est sélectionnée comme le milieu de la période de validité du quintuplet. GPT-4o génère ensuite quatre questions et réponses pour chaque quintuplet. Ensuite, le contexte temporel est supprimé de la question et il est vérifié que la réponse correspond à l'objet.

A.3 Qualité des questions générées

La qualité des verbalisations a été analysée pour identifier et éliminer les entrées incorrectes. Initialement, sur les 2 098 faits destinés à la verbalisation, 53 n'ont pas été verbalisés et 64 questions ont

utilisé à tort le sujet comme réponse au lieu de l'objet. Ces cas erronés ont été supprimés du jeu de données, ce qui a donné un total de 2 003 faits et $2003 \times 4 = 8012$ questions.

Un échantillon aléatoire de 50 questions a été évalué manuellement pour garantir la qualité générale des questions générées. L'évaluation a révélé que seulement 1 des 50 questions était incorrecte, tandis que les questions restantes étaient parfaitement construites (intervalle de confiance de Wilson à 95 % = [0.85, 0.99]) [9]. Ces résultats démontrent la haute qualité des questions de notre ensemble de données.

Enfin, chaque fait se voit attribuer de manière aléatoire l'une de ses quatre questions associées.

A.4 Génération de tests

L'arithmétique entre les contextes temporels est impliquée dans cette section. Il convient de noter que toutes les opérations entre les contextes sont effectués sur le milieu de contexte (car les contextes étudiés sont des intervalles). Par exemple, lorsque $a + b$ est calculé, le résultat est le milieu de a ajouté au milieu de b. La granularité la plus fine qu'un *milieu* peut avoir est la granularité JMA (c.-à-d, Jour-Mois-Année). Cela permet de contourner la nature d'intervalle des dates.

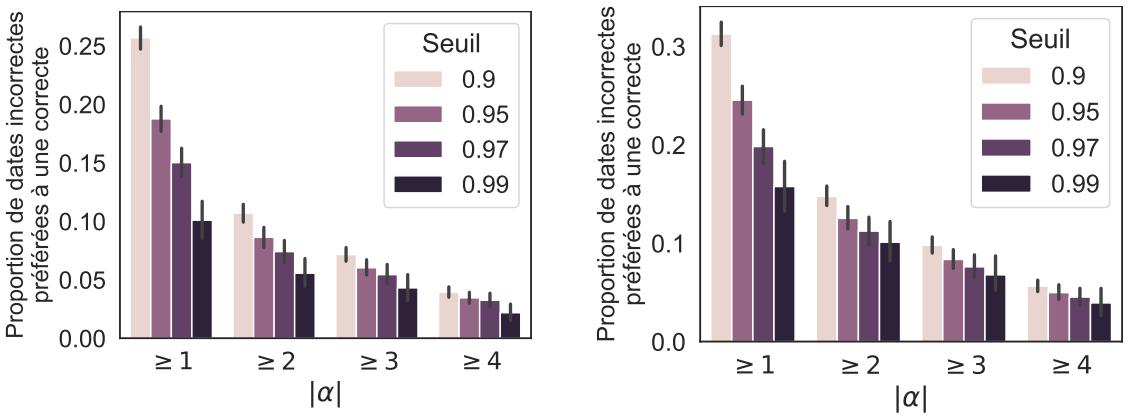
Pour chaque quintuplet, la plage de contextes testées est définie comme $m \pm 5l$, où m est le milieu de la période de validité $(a + b)/2$, et d est la durée de la période de validité $b - a$. Pour déterminer les dates de granularité A (c.-à-d, Année) à inclure dans TimeStress, nous effectuons une analyse à partir du milieu et en s'étendant jusqu'aux extrémités avec un pas de $0.05 \times d$. Cette taille de pas est choisie pour limiter le nombre maximal de contextes corrects et incorrects aux valeurs raisonnables de 21 et 180, respectivement. Pour chaque contexte de granularité A, un contexte de granularité MA est choisie en tirant aléatoirement un mois de l'année. Similairement, pour chaque contexte de granularité MA, un contexte de granularité JMA est choisie en tirant aléatoirement un jour de contexte de granularité MA précédent [10]. Cela crée une relation hiérarchique entre les différentes granularités (par exemple *2020*, *2020-03*, *2020-03-24*), permettant des comparaisons raisonnables en termes de taux de victoire et de robustesses, car elles partagent la même année et/ou le même mois. Tous les contextes sont maintenant classifiés comme correctes, incorrectes ou transitoires (cf. section 3.1).

Malgré cette configuration, un fait peut avoir un nombre variable de contextes corrects et incorrects par granularité en raison des contextes transitoires, qui peuvent être absentes dans les granularités fines si le pas de $0.05 \times d$ passent par-dessus. Cette différence pourrait biaiser les performances, en favorisant notamment la granularité A sur la métrique de robustesse, qui sont calculées sur moins de tests. Pour résoudre ce problème, les contextes de granularité MA et JMA associées aux contextes transitoires de granularité A sont supprimées des ensembles corrects et incorrects et affectées à une classe spéciale appelée **Écarté**.

Enfin, les contextes sont convertis en texte et préfixés aux questions pour créer des énoncés pour chaque contexte à chaque granularité pour chaque fait.

Le jeu de données résultant, nommé **TimeStress**, comprend 521 000 énoncés générées à partir de 2 003 faits temporels. En moyenne, il comprend 11 dates correctes et 74 dates incorrectes, comprenant 1 883 entités uniques, 1 385 sujets uniques, 1 113 objets uniques et 86 relations. Un échantillon

9. Cet intervalle de confiance a été calculé avec une correction de population finie.
10. Cet échantillonnage ne produit pas de dates erronées telles que le 29 février pour les années non-bissextiles, ou le 31 avril.



(a) Texte brut

(b) Format instruction

FIGURE 6 – Proportion de contextes incorrects privilégiés par rapport aux contextes corrects qui se situent au-delà d'une distance relative $|\alpha|$ par rapport à la période de validité, lorsque le taux de victoire dépasse le seuil, sur les 5 ML les plus robustes. Les expériences ont été réalisées avec la granularité A. Les intervalles de confiance à 95% ont été calculés à l'aide d'un *bootstrap*.

aléatoire de TimeStress est présenté dans le tableau 3.

B Vulnérabilité face aux contextes incorrects faciles : analyse des résultats sur différents seuils de taux de victoire

Dans la section 4.2, nous avons démontré que les ML, même lorsqu'ils sont presque parfaitement robustes sur un fait (c.-à-d, un taux de victoire très élevé mais inférieur à 100%), ne parviennent souvent pas à atteindre la robustesse en raison de leur vulnérabilité aux contextes faciles qui se situent bien en dehors de la période de validité (tableau 2). Dans cette section, nous étendons cette analyse en expérimentant différents seuils de taux de victoire pour observer comment la distribution des contextes incorrects favorisés par rapport aux contextes corrects évolue lorsque le seuil se rapproche des 100%.

Les résultats de la figure 6 indiquent que même lorsque le seuil se rapproche de 1, les ML restent vulnérables aux contextes incorrects faciles qui sont significativement éloignés de la période de validité. Nous nous attendons à ce que les ML excluent définitivement les contextes très éloignés une fois qu'ils ont acquis suffisamment d'informations sur la période de validité, comme le ferait un humain. Mais ce n'est pas le cas ici, car même lorsque le taux de victoire est très proche de 1, les ML continuent d'échouer sur ces contextes. Ces résultats suggèrent que les modèles linguistiques ne parviendront peut-être jamais à une véritable robustesse, car la proportion de faits incorrects converge vers zéro mais ne l'atteint jamais complètement, ce qui implique qu'il y aura toujours une possibilité pour un ML d'échouer sur un contexte incorrect lointain, un scénario qui ne se produirait pas avec une base de connaissances temporelle structurée. De plus, cela suggère que le faible pourcentage de faits robustes pourrait être encore plus faible si nous augmentions le nombre de contextes incorrects et corrects utilisés pour calculer la robustesse.

C Généralisation des connaissances entre granularités

Cette section présente des détails et des résultats supplémentaires concernant la généralisation des connaissances entre les granularités.

C.1 Cohérence entre granularités suivant la distance relative

Dans cette section, nous examinons la cohérence des prédictions des ML à travers différentes granularités (A, MA, JMA) à mesure que la distance entre le contexte testée et la période de validité augmente.

Pour évaluer cela et uniquement pour cette section, nous introduisons une métrique appelée robustesse locale. La robustesse locale pour un fait, un ML et un contexte incorrect donnés est définie comme égale à 1, si tous les contextes corrects sont préférés par le ML à ce contexte incorrect et 0 sinon.

Nous regroupons tous les énoncés dans TimeStress en fonction de la distance relative α de leur contexte temporel, en se restreignant aux 5 ML les plus robustes et au faits « connus » [11] en moins sur une granularité par ces ML. Ces énoncés sont catégorisés selon l'intervalle dont fait partie leur distance relative α. Les intervalles choisis sont de la forme $]s, s + 0.5]$, où s peut prendre des valeurs de $\{-5, -4.5, \ldots, 4.5\}$. Pour chaque intervalle, les contextes sont alignés par fait et par granularité de manière hiérarchique (par exemple, `2020`, `2020-04`, `2020-04-23`), ce qui est garanti possible grâce aux propriétés de TimeStress (cf. Section 3.2). La robustesse locale est ensuite calculée pour chaque contexte incorrecte, et l'exactitude [12] entre ces mesures est calculé pour toutes les paires de granularité (c.-à-d, A-MA, A-JMA et MA-JMA). Ces coefficients sont moyennés sur toutes les paires de granularité, tous les faits et les 5 ML les plus robustes, les résultats étant présentés dans la figure 7.

Les résultats indiquent que l'incohérence entre granularité est causée principalement par les contextes incorrectes se situant aux frontières de la période de validité. À force que le contexte s'éloigne de la période de validité, la cohérence s'approche sensiblement d'une cohérence parfaite mais ne l'atteint jamais quelque soit le ML, le type d'énoncé et l'intervalle d'α utilisés.

C.2 Matrices de généralisation pour pour chaque ML

Dans la section 4.2, nous avons exploré la capacité des modèles de langue à généraliser leurs connaissances temporelles d'une granularité à une autre. Nous avons fourni deux matrices (une pour les questions au format instruction et une pour les questions au format brut) qui contiennent le taux de généralisation entre chaque paire de granularités moyennées sur les 5 ML les plus robustes. En complément de ces performances moyennes, la matrice des taux de généralisation des modèles individuels est présentée dans la figure 8.

11. Nous rappelons que « connu » dans le contexte de cet article signifie que ML en question a une robustesse égale à 1 sur le fait en question, c.-à-d, tous les contextes corrects sont préférés aux contextes incorrect par le ML.

12. L'exactitude mesure la proportion d'éléments identiques entre deux vecteurs, c'est-à-dire le nombre de positions où les valeurs sont égales, rapporté au nombre total d'éléments comparés.

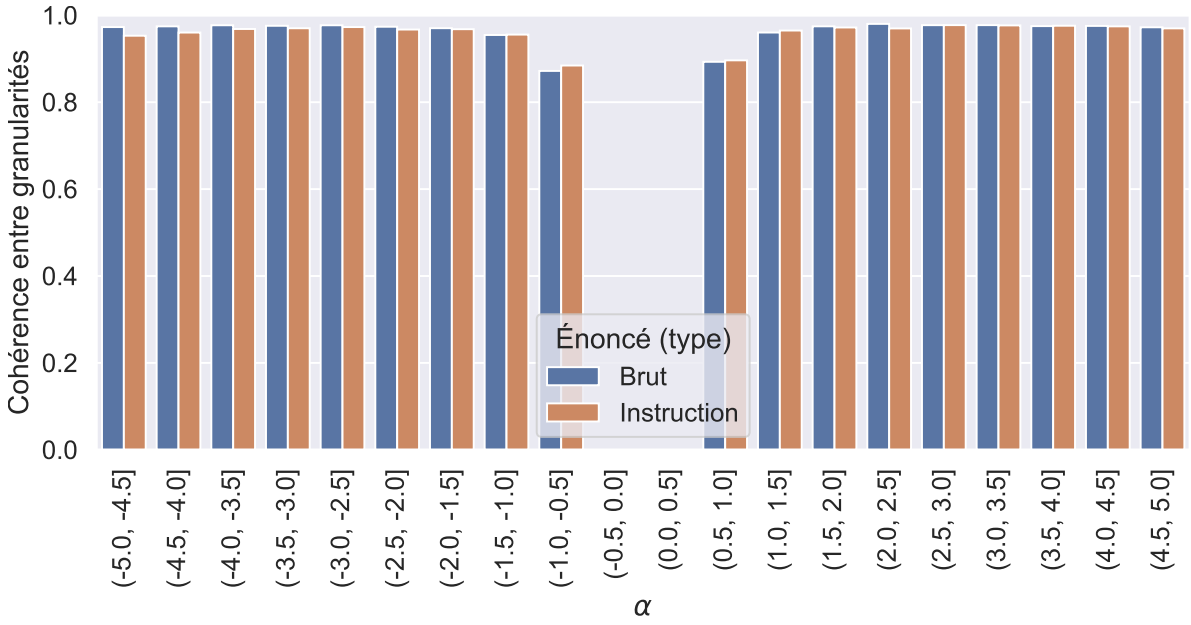


FIGURE 7 – Pour chaque segment α, la corrélation de robustesse locale moyenne entre toutes les paires de granularités est calculée sur tous les faits et les 5 ML les plus robustes.

C.3 Amorces explicatives

Nous avons étudié dans la section 4.2 si l'inclusion d'explications de concepts temporels dans l'amorce pouvait aider les ML à mieux généraliser leurs connaissances entre les granularités. Deux amorces préfixées à chaque instruction dans TimeStress ont été utilisées :

Amorce 1 : nature hiérarchique des dates

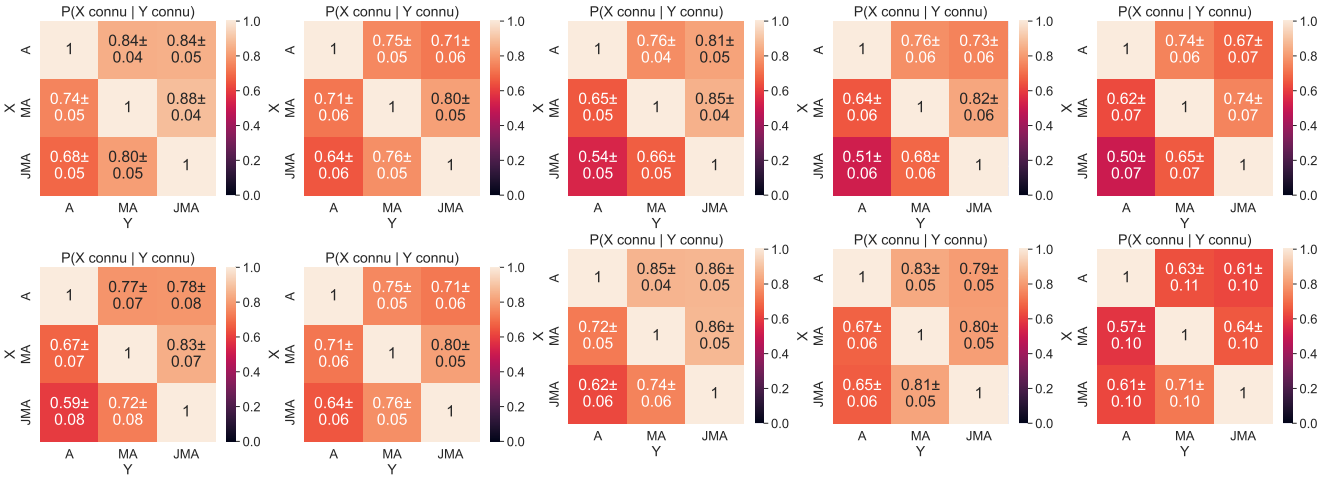
```
A date is a specific point in time, expressed through a year, a month, and a day. A
year is divided into months, and a month is divided into days. Answer the following
question.
```

Amorce 2 : transfert de connaissances entre granularités

```
A date is a specific point in time. If a fact is valid for a specific year, it holds
true for all dates within that year. If a fact is valid for a specific month of a
specific year, it holds true for all dates within that month. Answer the following
question.
```

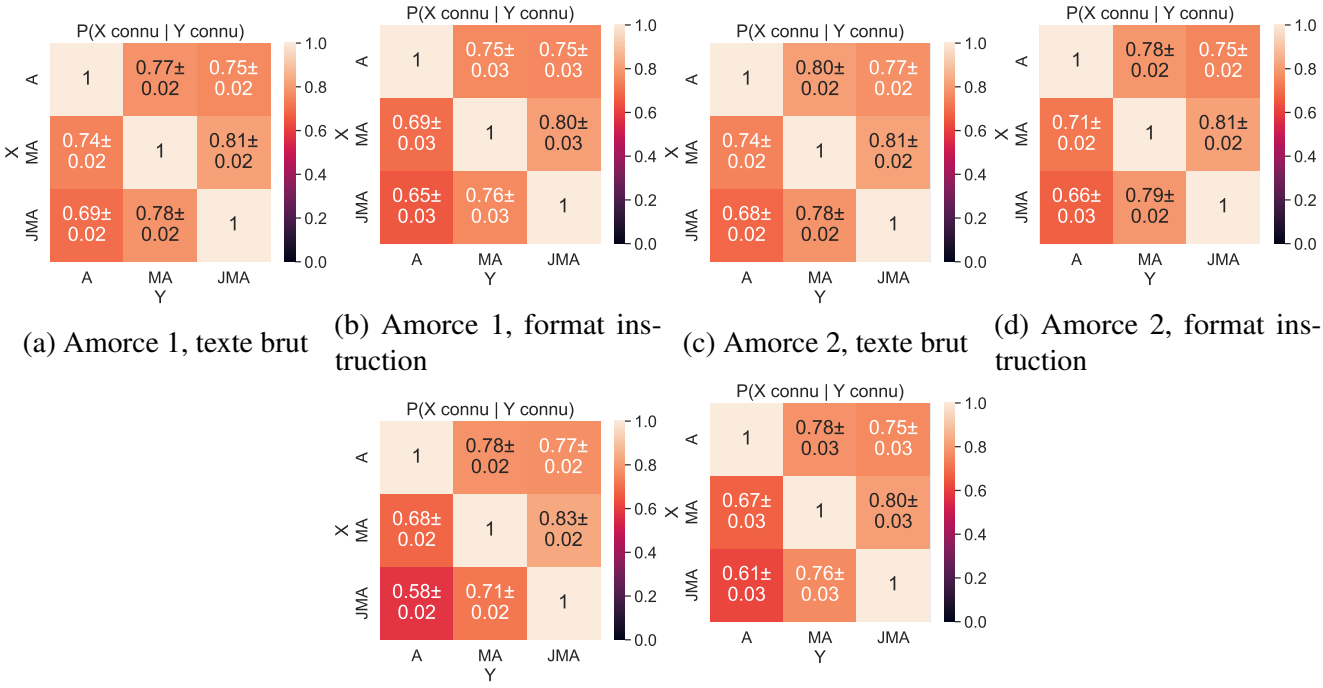
La première explique la nature hiérarchique des dates tandis que la seconde est plus directe et explique comment la connaissance d'un fait temporel peut être généralisée entre les granularités.

En complément des performances moyennes de la section 4.2, la figure 9 montre les matrices de généralisation moyennes sur les 5 mêmes modèles que dans la figure 5, en utilisant du texte brut et un format d'instruction.



(a) gemma-2-27b-it (b) gemma-2-9b-it (c) Llama-3.1-70B-Instruct (d) Mistral-Nemo-Instruct-2407 (e) Mistral-7B-Instruct-v0.3

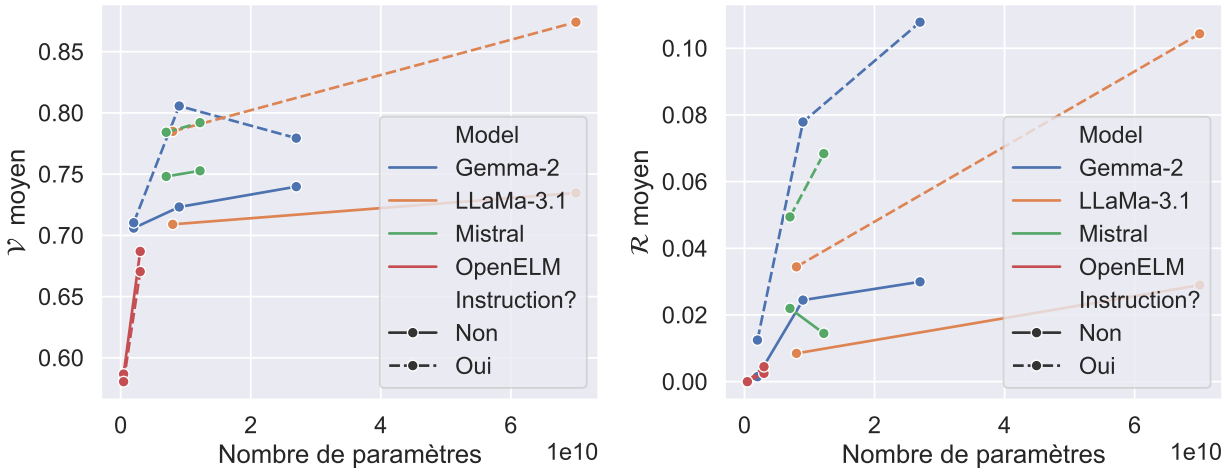
FIGURE 8 – Matrices de généralisation entre les paires de granularité des 5 ML les plus robustes. Dans la **première ligne**, les énoncés sont présentés aux ML dans un format **brut**, et dans la **deuxième ligne**, ils sont présentés dans un format d'**instruction**.



(a) Amorce 1, texte brut (b) Amorce 1, format instruction (c) Amorce 2, texte brut (d) Amorce 2, format instruction

(e) Pas d'amorce, texte brut (f) Pas d'amorce, format instruction

FIGURE 9 – Effet de l'ajout d'explications sur les concepts temporels par le biais d'une amorce explicative (cf. Annexe C.3)



(a) Taux de victoire

(b) Robustesse

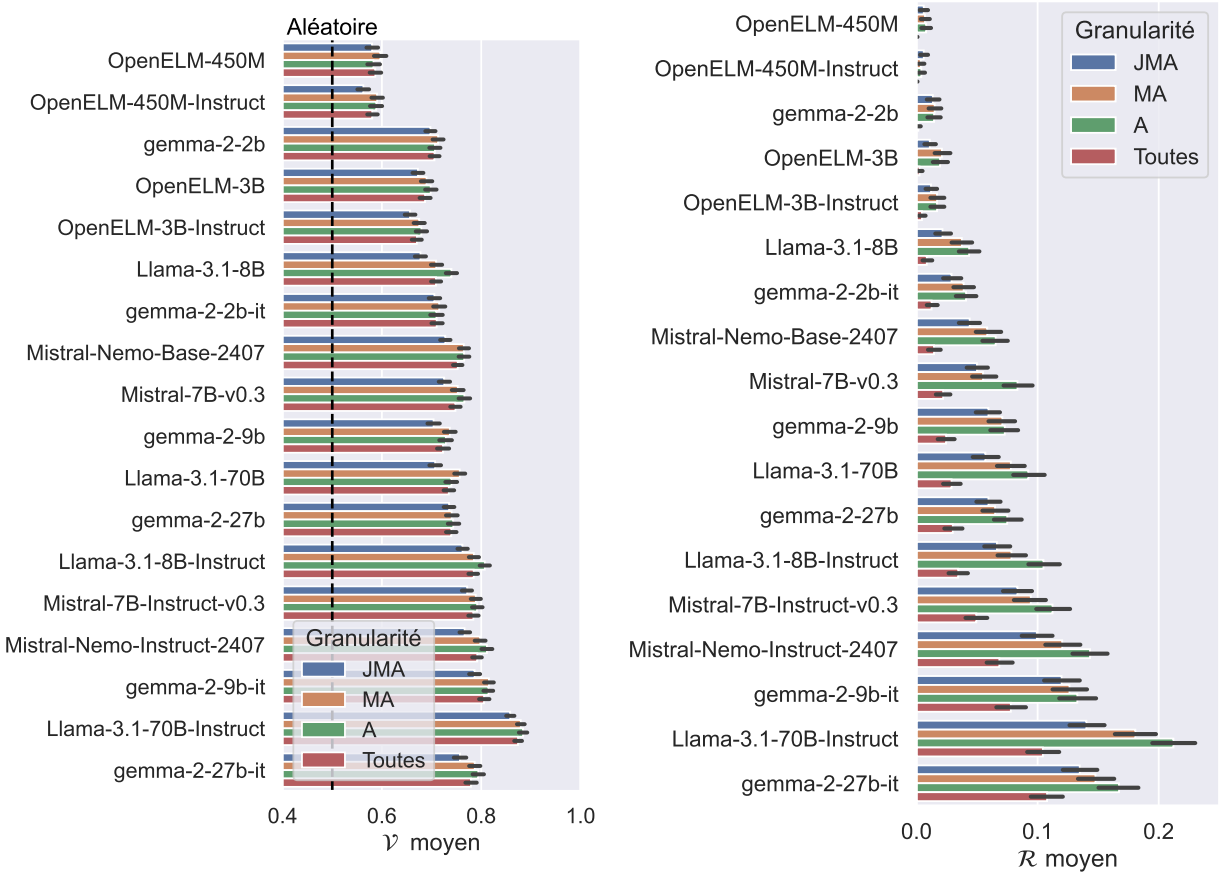
FIGURE 10 – Relation entre le nombre de paramètres d'un ML et la métrique utilisée (sur l'ensemble des granularités A, MA et JMA). Les modèles pré-entraînés sont représentés par des lignes droites, tandis que les modèles affinés sur des instructions sont représentés par des lignes pointillées.

D Résultats supplémentaires

— Le score de robustesse moyen et le taux de victoire sur les 18 ML étudiés sont présentés dans la figure 11.
— La relation entre le nombre de paramètres des ML et leur performances est présenté dans la figure 10.
— La figure 14 montre l'évolution de $logP(o|f,d)$ par rapport à la distance relative de la date à la période de validité α, ce qui est équivalent à la figure 3 mais avec plus de détails.
— La figure 15 montre les relations sur lesquelles les MLs sont les plus robustes en moyenne (format d'énoncés bruts).
— La figure 16 montre des exemples où les MLs ont été vulnérables a des contextes incorrects *faciles*.
— La figure 13 montre la distribution des années des contextes temporels dans l'ensemble du jeu de données TimeStress.
— La figure 17 montre l'influence de la distance des faits par rapport au présent (ici, l'année 2021), ainsi que leurs durées sur la robustesse et le taux de victoire des 5 ML les plus robustes. L'unité de temps utilisée pour ces deux mesures est l'**année**.

E Limites de l'étude

L'étude évalue les ML en utilisant une approche basée sur les probabilités pour évaluer leur compréhension des faits temporels. Bien que cette méthode ne capture pas entièrement les performances des modèles dans des scénarios de génération de texte, elles sont fortement liées car le texte généré est échantillonné à partir de la distribution de probabilité du ML. En outre, notre approche permet une exploration précise de relations non fonctionnelles spécifiques où plusieurs réponses correctes existent. Cela est plus difficile avec des métriques basées sur la génération, car les ML peuvent



(a) Taux de victoire

(b) Robustesse

FIGURE 11 – Métriques moyennes sur l'ensemble de fait dans TimeStress pour les 18 ML étudiés avec intervalles de confiance à 95%.

produire une autre réponse correcte, des réponses inattendues ou hors sujet. De plus, des recherches antérieures ont montré que les métriques basées sur les probabilités corrèlent raisonnablement bien avec les performances génératives des modèles dans des contextes d'évaluation des connaissances factuelles, où le modèle est censé générer des entités spécifiques (Dong *et al.*, 2023; Lyu *et al.*, 2024), ce qui est étroitement lié à notre approche.

Ensuite, les résultats de notre étude se limitent au format des énoncés que nous avons choisi, c.-à-d, un contexte temporel suivi d'une question et d'une réponse. Il est possible que les ML aient de meilleures performances sur un format différent. Cependant, leurs limites actuelles sur nos données sont déjà problématiques.

Enfin, le jeu de données TimeStress est constitué d'énoncés en anglais, ce qui pourrait limiter l'applicabilité de nos résultats à d'autres langues en raison d'éventuelles différences linguistiques pouvant affecter la compréhension temporelle. Cependant, des recherches futures peuvent facilement élargir le champ d'application en adaptant l'amorce de GPT-4o utilisée pour verbaliser les faits afin de cibler des langues supplémentaires. Pour ce qui est des étiquettes des entités, elles sont disponibles dans d'autres langues dans Wikidata.

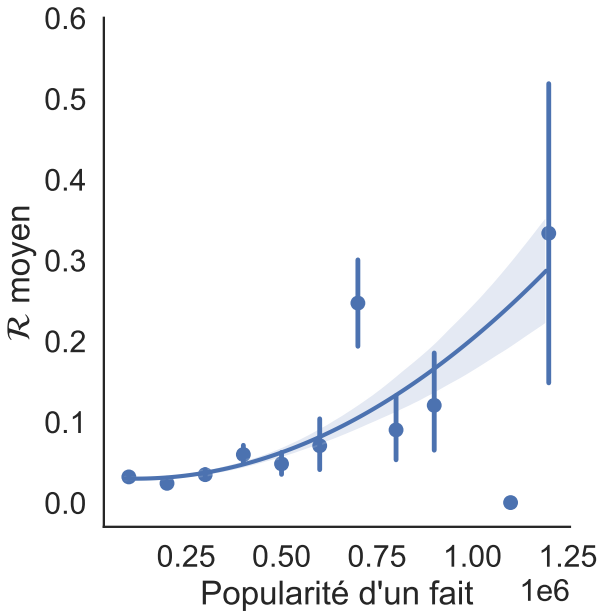


FIGURE 12 – Relation entre la popularité des faits et la métrique de robustesse calculée sur l'ensemble des ML et des granularités A, MA et JMA.

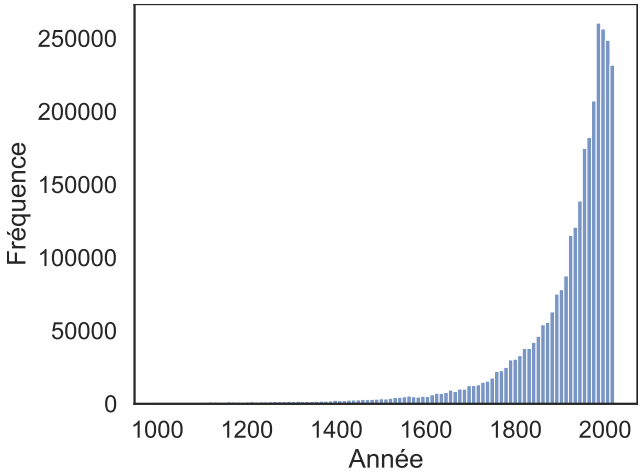


FIGURE 13 – Distribution des années des contextes temporels dans TimeStress.

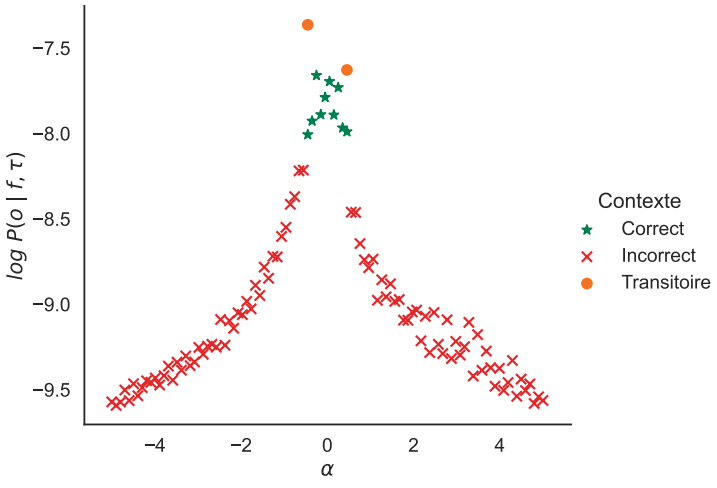


FIGURE 14 – L'évolution de $logP(o|f,d)$ par rapport à la distance relative de la date à la période de validité α. Chaque point est une moyenne sur de nombreux points de données.

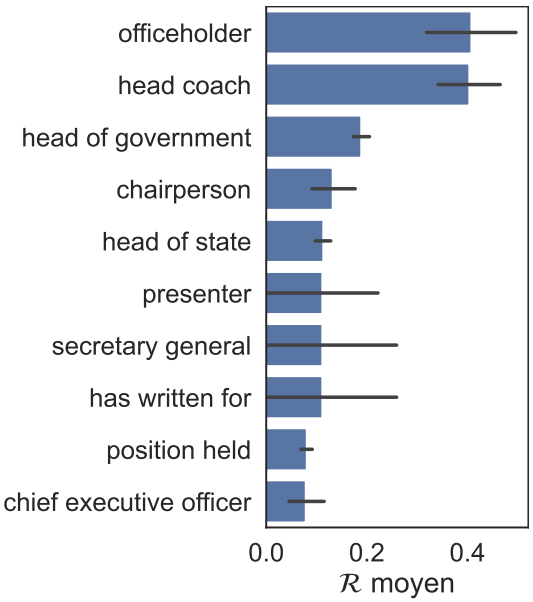
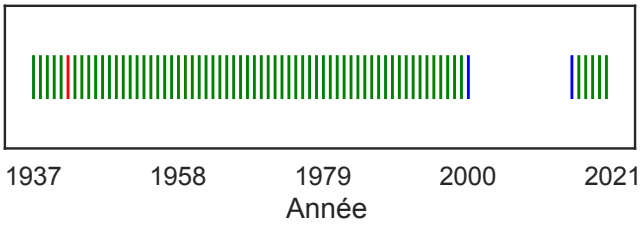
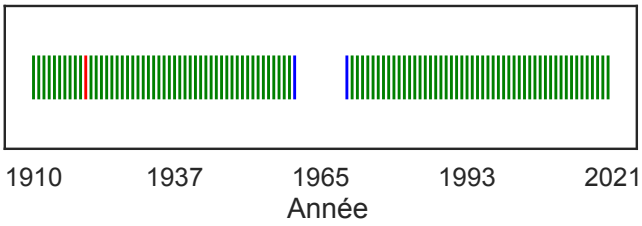


FIGURE 15 – Les 10 relations sur lesquelles les MLs sont les plus robustes en moyenne (format d'énoncés bruts).



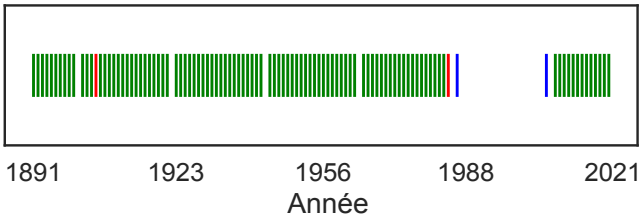
Question : In [YEAR], who led the government of Texas? Rick Perry

Type : Instruction
Modèle : Mistral-Nemo-Instruct-2407



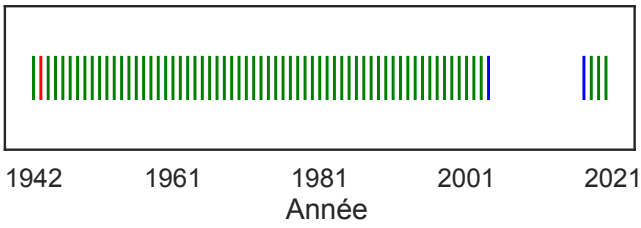
Question : In [YEAR], of which band was Paul McCartney a member? The Beatles

Type : Instruction
Modèle : Llama-3.1-8B-Instruct



Question : In [YEAR], who was the owner of Pixar? Steve Jobs

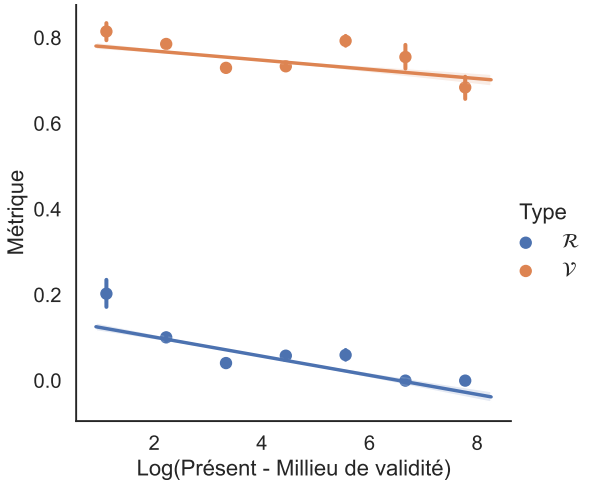
Type : Brut
Modèle : gemma-2-9b-it



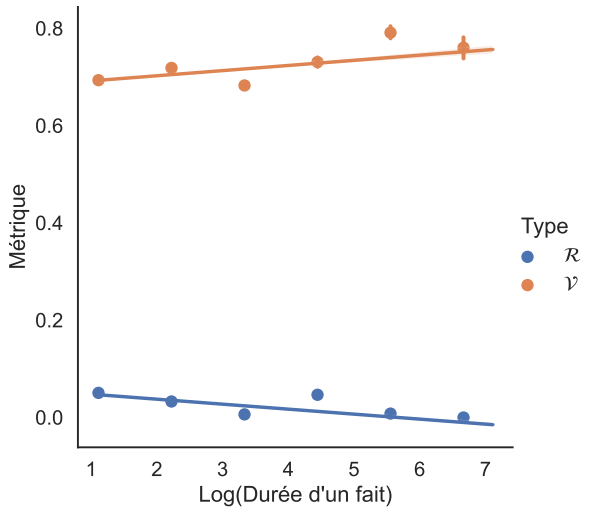
Question : In [YEAR], which football club was Wayne Rooney associated with? Manchester United F.C.

Type : Instruction
Modèle : gemma-2-27b-it

FIGURE 16 – Exemples de vulnérabilité face à des contextes incorrects *faciles* pour différents ML. La couleur bleue représente les frontières de la période de validité, la couleur verte représente les contextes incorrects qui ne sont jamais préférés aux contextes corrects, et la couleur rouge, au contraire, représente les contextes incorrects qui ont été préférés à un contexte correct ou plus. Les erreurs critiques sont en rouge.



(a) Logarithme de la distance du fait par rapport au présent.



(b) Logarithme de la durée du fait.

FIGURE 17 – L'influence de deux facteurs sur la robustesse et le taux de victoire des 5 ML les plus robustes. Toutes les corrélations sont significative où l'hypothèse nulle est l'absence de corrélation linéaire.

Fait temporel	Énoncé	Statut
(Alexander Graham Bell, country of citizenship, United States of America, 1882, 1922)	In July 1734, what was Alexander Graham Bell's country of citizenship? United States of America	Incorrect
(Lauren Bacall, spouse, Jason Robards, 1961-07-04, 1969-09-10)	In July 1984, who was the spouse of Lauren Bacall? Jason Robards	Incorrect
(Vatican City, head of state, John Paul II, 1978-10-16, 2005-04-02)	In July 2006, who held the highest authority in Vatican City? John Paul II	Incorrect
(Gareth Barry, member of sports team, Manchester City F.C., 2009, 2014)	In July 2020, which football team included Gareth Barry as a player? Manchester City F.C.	Incorrect
(Pierce Brosnan, spouse, Cassandra Harris, 1980, 1991)	In 1954, who did Pierce Brosnan have as his wife? Cassandra Harris	Incorrect
(Metallica, has part, Jason Newsted, 1987, 2001-01-17)	In 1971, who was included in Metallica's lineup? Jason Newsted	Incorrect
(Eliza Dushku, unmarried partner, Rick Fox, 2009, 2014)	In 2003, who was Eliza Dushku in a relationship with? Rick Fox	Incorrect
(United Kingdom, head of state, George VI, 1936-12-11, 1952-02-06)	On July 1, 1892, who served as the king of the United Kingdom? George VI	Incorrect
(Linda Lee Cadwell, spouse, Bruce Lee, 1964, 1973-07-20)	In 1929, who was the spouse of Linda Lee Cadwell? Bruce Lee	Incorrect
(George Harrison, part of, The Beatles, 1960, 1970)	On July 2, 1971, what was the name of the band that George Harrison was associated with? The Beatles	Incorrect
(Philippines, head of state, Corazon Aquino, 1986-02-25, 1992-06-30)	On July 2, 1969, who served as the leader of the Philippines? Corazon Aquino	Incorrect
(Jawaharlal Nehru, position held, Prime minister of India, 1947-08-15, 1964-05-27)	In 1985, what position did Jawaharlal Nehru hold? Prime Minister of India	Incorrect
(Vienna, country, Austria-Hungary, 1867-03-30, 1918-11-11)	In July 1769, which country did Vienna belong to? Austria-Hungary	Incorrect
(Mileva Marić, spouse, Albert Einstein, 1903, 1919)	In July 1907, who was Mileva Marić married to? Albert Einstein	Correct
(Mayte Garcia, spouse, Prince, 1996, 2000)	In July 1979, who was the spouse of Mayte Garcia? Prince	Incorrect
(Abkhazia, country, Soviet Union, 1921, 1991)	In July 1956, which country did Abkhazia belong to? Soviet Union	Correct
(Georgia, member of, Commonwealth of Independent States, 1993-12-03, 2009-08-18)	In 1930, what group included Georgia as a member? Commonwealth of Independent States	Incorrect
(Abraham Lincoln, member of political party, Whig Party, 1834, 1854)	In 1808, which political party was Abraham Lincoln a member of? Whig Party	Incorrect
(Wales, located in the administrative territorial entity, Kingdom of England, 1284, 1707-04-30)	On July 1, 1072, which territorial entity included Wales? Kingdom of England	Incorrect
(Frédéric Chopin, residence, Paris, 1831, 1849)	On July 2, 1847, which city was home to Frédéric Chopin? Paris	Correct

TABLE 3 – Échantillon aléatoire de TimeStress.