Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Étude des facteurs de complexité des modèles de langage dans une tâche de compréhension de lecture à l'aide d'une expérience contrôlée sémantiquement

Elie Antoine[1] Frédéric Béchet[1,4] Géraldine Damnati[2] Philippe Langlais[3]
(1) Aix-Marseille Université, CNRS, LIS
(2) Orange Innovation, DATA&AI, Lannion
(3) RALI/DIRO, Université de Montréal, Canada
(4) International Laboratory on Learning Systems (ILLS - IRL CNRS), Montreal
{first.last}@lis-lab.fr , {first.last}@orange.com,
felipe@iro.umontreal.ca

RÉSUMÉ

Cet article propose une méthodologie pour identifier des facteurs de complexité inhérents aux tâches de traitement automatique du langage (TAL), indépendamment de la dimension des modèles. Il montre que la performance inférieure de certains exemples est attribuable à des facteurs de complexités spécifiques. Plutôt que de procéder à des évaluations générales, nous préconisons des évaluations restreintes portant sur des tâches, des ensembles de données et des langues spécifiques, décrites de manière linguistique. Appliquée à une tâche de compréhension de texte via un corpus de questions-réponses, notre méthode met en évidence des facteurs de complexité sémantique affectant divers modèles de tailles et d'architectures différentes. En outre, nous proposons plusieurs corpus de complexité sémantique croissante dérivés de ces facteurs, avançant que l'optimisation de leur traitement dépasse la simple augmentation de la taille des modèles.

ABSTRACT

Investigating the complexity factors of language models in a reading comprehension task using a semantically controlled experiment

The paper introduces a methodology focusing on identifying complexity factors specific to natural language processing (NLP) tasks, independent of model size. It suggests that certain examples consistently yield lower scores regardless of model size due to specific complexity factors. Rather than broad evaluations, the methodology advocates for constrained evaluations on specific tasks, datasets, and languages, described linguistically. The study demonstrates this approach on a reading comprehension task using a corpus of question-answer pairs. It proposes and validates semantic complexity factors affecting models of different sizes and architectures. Additionally, it defines multiple corpora of increasing semantic complexity derived from these factors. The study argues that improving the processing of these corpora requires more than just increasing model parameters.

MOTS-CLÉS : Compréhension de texte, Question/Réponse, Grand modèle de langage, Annotation sémantique, FrameNet.

KEYWORDS: Text Understanding, Question Answering, LLM, semantic annotation, FrameNet.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 384–396.

1 Introduction

Les modèles de langage génératifs constituent actuellement l'état de l'art pour presque toutes les tâches de traitement du langage naturel, en particulier grâce au développement de grands modèles de langage, dont le nombre et la taille ne cessent de croître. Toutefois, malgré l'intense activité de recherche les concernant, de nombreuses zones d'ombre demeurent quant à leurs capacités, limites, et risques. L'étude académique de ces modèles est entravée par des défis significatifs : les modèles "ouverts" nécessitent des ressources considérables non accessibles à tous, tandis que l'analyse des modèles "fermés", accessibles via des API, est limitée par des coûts potentiellement élevés et un manque de transparence sur leur fonctionnement et apprentissage.

Cette réalité encourage l'examen de modèles plus petits, dont la gestion, le développement, et l'analyse sont simplifiés. Cependant, cette approche est freinée par le risque que les améliorations apportées ne soient pas transférables à des modèles possédant plus de paramètres, posant ainsi la question de leur impact réel sur le développement futur.

Pour remédier à cette limitation, nous présentons dans cet article une approche axée sur l'identification de facteurs de complexité inhérents à des tâches spécifiques, indépendamment de la taille des modèles développés. Cette méthode postule que peu importe leur capacité individuelle, ces modèles sont uniformément influencés par ces facteurs de complexité.

Pour obtenir ces facteurs, au lieu d'effectuer des évaluations "boîte noire" couvrant de nombreuses tâches, jeux de données et langues (Liang *et al.*, 2023; Srivastava *et al.*, 2023) nous préconisons des analyses précises, confinées à des contextes spécifiquement définis et décrits sous un angle linguistique. Cette méthode permet de comparer des modèles de différentes tailles et cherche à révéler les facteurs de complexité linguistique inhérents aux capacités de résolution de tâches de chaque modèle.

Cet article présente un cas d'application de notre méthode à une tâche de compréhension de texte, en utilisant le corpus français Calor (Béchet *et al.*, 2019), qui comprend des paires Question-Réponse (QR) enrichies d'annotations sémantiques fines détaillant la relation entre questions et réponses. En analysant les résultats de différents modèles de tailles diverses sur ces paires et en tenant compte des annotations sémantiques, nous visons à identifier des facteurs de complexité pertinents pour tous les modèles.

Les principales contributions de cette étude consistent à proposer et à valider des facteurs de complexité sémantique qui ont un impact négatif sur des modèles d'architectures et de tailles diverses. En outre, l'étude définit des corpus de complexité sémantique croissante dérivés de ces facteurs, obtenus par des partitions du corpus Calor . Nous soutenons que l'amélioration du traitement de ces corpus nécessite plus qu'une simple augmentation du nombre de paramètres du modèle.

2 Un corpus de compréhension de texte sémantiquement contrôlé

L'étude de la tâche de question-réponse à partir de documents a été largement étudiée avec l'émergence des modèles de réseaux neuronaux profonds, stimulée par l'accès à d'importants corpus d'évaluation tels que SQuAD (Rajpurkar *et al.*, 2016) ou MultiRC (Khashabi *et al.*, 2018), qui fait partie du benchmark SuperGLUE (Wang *et al.*, 2019). Dans ces classements, les modèles de langage basés sur des Transformers sont aujourd'hui systématiquement plus performants que les humains. Cela

souligne la difficulté d'évaluer les modèles en raison de la nature subjective de la génération des réponses, tout en démontrant la capacité des modèles à capturer et à reproduire les biais inhérents aux données sur lesquelles ils sont entraînés.

Cette tâche présente plusieurs avantages malgré ses contraintes : (1) Elle facilite la comparaison directe entre modèles classificateurs et génératifs dans un contexte unifié, où les classificateurs, bien que moins complexes, parviennent à égaler les performances des génératifs ; (2) Elle peut être partiellement décrite par un modèle linguistique formel détaillant les liens syntaxiques et sémantiques entre questions et réponses, contrairement aux tâches exclusivement génératives comme le résumé ou la traduction, plus difficiles à formaliser linguistiquement.

Dans cette étude, nous utilisons le corpus Calor contenant des paires question/réponse enrichies d'annotations sémantiques selon le modèle Berkeley Framenet (Baker *et al.*, 1998). Ce corpus, destiné à l'origine à l'évaluation de la génération de questions, rassemble des textes en français issus de Wikipedia et de ClioTexte [1], un ensemble de documents historiques utilisés dans l'enseignement en France, couvrant principalement trois sujets : la Première Guerre mondiale, l'archéologie, et l'antiquité, avec des sources variées, depuis les documents historiques de ClioTexte jusqu'aux articles encyclopédiques de Wikipedia.

Les annotations s'articulent autour de *cadres sémantiques* (***Frames***), définissant des scénarios prototypiques (tels que *décider*, *perdre*, *attaquer*, *vaincre*, etc.). Le processus d'annotation consiste d'abord à identifier l'*Unité Lexicale ou Lexical Unit (**LU**)* déclenchant le Cadre, suivie par celle des éléments constitutifs du cadre, ou *Frame Elements (**FE**)*.

Un exemple est donné dans la figure 1 pour une phrase annotée avec les deux *Frames*, `Losing` déclenché par le mot `perdu` et `Attack` déclenché par `assauts`.

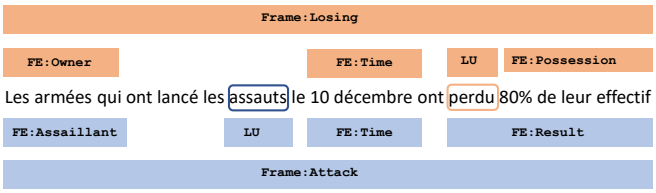

FIGURE 1 – Exemple de phrase annotée avec les cadres sémantiques FrameNet

La création d'un corpus de *questions/réponses* structuré sémantiquement a été réalisée selon la méthode suivante : initialement, une sélection aléatoire d'un *cadre* sémantique F et d'un *élément de cadre* E est effectuée dans une phrase spécifique. Des annotateurs humains sont ensuite chargés de formuler une question dont E constitue la réponse, exploitant pour cela la relation sémantique existant entre F et E. Les autres *FE* présents dans la phrase servent de contexte C à la question, offrant ainsi une contextualisation supplémentaire. Par la variation des éléments F, E, et C au sein d'une même phrase, il est possible de générer un large éventail de questions, accompagnées de leurs réponses et catégorisées selon leur classe sémantique correspondant au type du *FE*.

Prenons pour exemple la phrase mentionnée dans la Figure 1. Avec F = `Perdre`, E = `Propriétaire`, et C = `Temps, Possession`, il est possible pour les annotateurs de formuler des questions comme :

— *Qui a perdu la majorité de ses troupes le 10 décembre ?*

1. https://clio-texte.clionautes.org/

— *Quelles armées ont été décimées lors des attaques du 10 décembre ?*

Pour encourager une diversité lexicale dans la formulation des questions, la phrase source n'était pas révélée aux annotateurs. Cette approche leur donnait la latitude de choisir librement les mots lorsqu'ils construisaient les questions et de décider des éléments contextuels C à intégrer. Lorsque les réponses exigeaient une analyse de coréférences, les chaînes de coréférence menant à la réponse appropriée étaient annotées. Dans l'exemple mentionné précédemment, la réponses (`les armées`) nécessitaient une résolution de coréférence pour déterminer la réponse exacte, bien que l'annotation des cadres se fasse au niveau de la phrase. Il en résulte que répondre correctement pourrait demander d'extraire des informations au-delà de la phrase donnée, dans d'autres sections du texte. Cette méthodologie a permis d'enrichir le corpus Calor de **1821** questions issues de 54 cadres sémantiques différents.

3 Modèles de langage pour la compréhension de la lecture

Dans cette étude, nous comparons six modèles de langage pré-entraînés sur notre tâche de question-réponse en utilisant le corpus Calor . Parmi ces modèles, l'un est un modèle de classification basé sur une architecture BERT développée pour la langue française, CamemBERT (Devlin *et al.*, 2019; Martin *et al.*, 2020). Trois d'entre eux sont des modèles génératifs multilingues basés sur **T5** (T5-LARGE, FLAN-T5-LARGE (Wei *et al.*, 2021), MT5-LARGE (Xue *et al.*, 2021)), et les deux autres sont des grand modèle de langage (LLM) actuels : LLAMA2 (Touvron *et al.*, 2023), Mixtral 8x7B (Jiang *et al.*, 2024) et chatGPT-3.5 [2].

Tous ces modèles pré-entraînés, à l'exception de GPT3.5, ont été affinés sur notre tâche de question-réponse (*QR*) en utilisant le corpus français FQuAD (d'Hoffschmidt *et al.*, 2020). Ce corpus, construit de manière similaire à SQuAD (Rajpurkar *et al.*, 2016), contient des questions basées sur des documents Wikipédia en français. Il convient de noter que si la plupart des textes du corpus Calor proviennent également de Wikipédia, Calor est nettement plus difficile que FQuAD. Cette différence s'explique par le fait que les textes de Calor sont spécialisés et que les réponses annotées sont significativement plus longues, correspondant aux éléments de cadre de nos annotations sémantiques.

Pour l'évaluation sur le corpus Calor , nous avons harmonisé les formats des corpus FQuAD et Calor , facilitant ainsi l'évaluation directe des systèmes affinés avec FQuAD sur Calor . Cette unification a permis d'appliquer un affinage spécifique à chaque modèle en vue de leur évaluation. Pour CamemBERT et les variantes de T5, nous avons procédé à un affinage sur le corpus FQuAD pendant deux époques. Concernant LLAMA2, nous avons employé la méthode Low-Rank Adaptation (LoRA), tandis que pour GPT-3.5 et Mixtral 8x7B, une approche d'amorçage à respectivement un et deux exemples a été utilisé. Cette dernière consiste à fournir au modèle un exemple d'entrée et de sortie dans le format désiré, lui indiquant explicitement de se concentrer sur l'extraction de contenu à partir du document source.

Les performances de ces sept modèles ont été évaluées sur le corpus Calor . Pour comparer l'efficacité des modèles extractifs (comme CamemBERT) et abstractifs (les autres modèles considérés), la métrique ROUGE-L, via l'outil ROUGE [3] a été utilisé. Une évaluation humaine a également été menée, nous n'indiquons ici que le pourcentage de réponses annotées comme correctes. Les résultats

2. API de `https://chat.openai.com`
3. Nous utilisons l'implémentation de google research disponible ici

obtenus, illustrés dans le tableau 1, montrent les scores moyens en ROUGE-L et en pourcentage de réponses correctes via l'annotation manuelle pour l'ensemble des 1821 questions composant le corpus.

Model	type	adapt	#param	Rouge-L	% réponses correctes
CamemBERT	classif.	FT	335M	0.82	78
T5-LARGE	gene.	FT	738M	0.81	77
FLAN-T5-LARGE	gene.	FT	783M	0.80	79
MT5-LARGE	gene.	FT	1.2B	0.80	77
LLAMA-2	gene.	LoRA	7B	0.69	72
Mixtral-8x7b	gene.	prompt	47B	0.80	82
GPT 3.5	gene.	prompt	175B	0.72	82

TABLE 1 – Description des 7 modèles de langage utilisés dans nos expériences avec le score ROUGE-L moyen et le pourcentage de réponses annotées comme correctes sur le corpus d'évaluation

Dans l'ensemble, les performances des différents modèles sur notre corpus sont considérablement plus basses par rapport à celles rapportées dans des tâches analogues telles que SQuAD [4] ou MultiRC dans SuperGLUE [5]. Cet écart peut être attribué en partie aux caractéristiques du corpus Calor , aux différences entre le corpus d'affinage FQuAD et le corpus d'évaluation Calor comme discuté précédemment, et aussi à l'absence d'optimisation du modèle via hyperparamétrage.

Les scores ROUGE-L des modèles de génération basés sur T5 et du modèle de classification basé sur CamemBERT sont relativement proches, alors que ceux de deux des grands modèles de langage, LLAMA-2 et GPT3.5, sont significativement en retard par rapport aux autres modèles. Ces résultats ne sont pas inattendus, étant donné que la tâche de question-réponse employée dans cette étude favorise les modèles extractifs par rapport aux modèles génératifs. Ce biais provient du fait que les références dans le corpus d'évaluation Calor sont extractives, comprenant des segments du texte original. Par conséquent, les métriques ROUGE favorisent intrinsèquement les modèles qui ne font que reproduire les segments sans introduire de mots supplémentaires.

Il est donc essentiel d'interpréter avec prudence les comparaisons directes de scores entre modèles basées sur cette métrique. L'intention de cette étude est moins de juger de la supériorité d'un modèle sur un autre en termes de performances brutes que d'explorer les facteurs linguistiques qui affectent les résultats de chaque modèle, indépendamment de leur efficacité absolue.

L'étude initiale que nous avons menée visait à examiner la corrélation potentielle entre le type de relation sémantique liant une question et sa réponse, et la performance des modèles. Pour discerner ces relations sémantiques, nous avons utilisé les cadres sémantiques (**Frames**) qui ont formé la base de l'élaboration des questions, comme détaillé dans le paragraphe 2. En segmentant le corpus Calor en 54 sous-corpus correspondant au nombre de cadres présents au total dans notre corpus, nous avons pu évaluer la performance de chaque modèle en fonction du cadre considéré.

S'il n'y avait pas de corrélation entre la performance et la relation sémantique, la variance des scores entre les sous-corpus devrait être minimale. Cependant, nos résultats indiquent le contraire.

Ceci est illustré dans la figure 2 où la distribution des scores de GPT-3.5 pour chaque sous-corpus de Frame est représentée. Comme on peut le voir, cette distribution n'est pas uniforme, validant l'hypothèse que la performance du modèle est liée aux relations sémantiques sous-jacentes. Ce même

4. https://rajpurkar.github.io/SQuAD-explorer
5. https://super.gluebenchmark.com/leaderboard

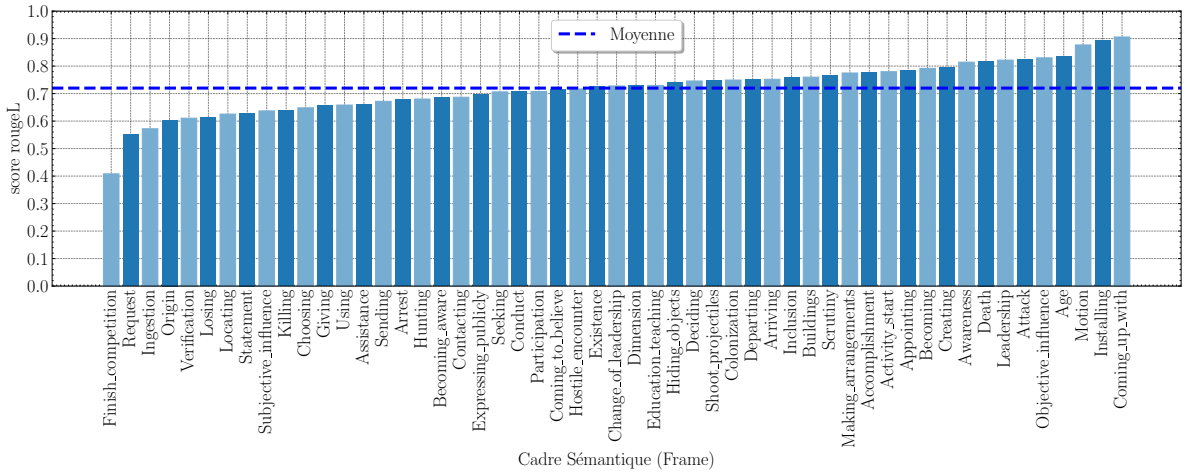

FIGURE 2 – Performance du modèle GPT-3.5 pour chaque Frame triée par la mesure rouge-L

type de distribution se retrouve pour tous les modèles, même si les mêmes Frames ne se retrouvent pas toujours dans les mêmes positions. Cette distribution se retrouve également dans l'évaluation humaine des réponses produites par les modèles. Dans tous les cas, le type de relations sémantiques a une influence significative sur la performance de tous les modèles. Dans la section suivante, nous proposerons différents facteurs de complexité pouvant expliquer ce phénomène.

4 Facteurs de complexité sémantique

Nous avons vu dans la section précédente que les performances des différents modèles de question-réponse n'étaient pas uniformes sur l'ensemble des cadres sémantiques. Cette section introduit une méthodologie destinée à identifier des facteurs de complexité sémantique susceptibles de justifier ces variations de performances. Nous étudierons aussi dans quelle mesure ces facteurs sont applicables de manière générale aux différents modèles analysés.

Méthode

1. Nous formulons plusieurs hypothèses concernant les facteurs de complexité sémantique ($F = f_1, f_2, \ldots$) sous forme de questions binaires applicables aux exemples du corpus d'évaluation. Par exemple : *La recherche de la réponse nécessite-t-elle de résoudre une chaîne de coréférence ?*
2. Pour chaque facteur f, nous divisons le corpus d'évaluation en deux sous-ensembles E_f et $\bar{E}f$ contenant respectivement les exemples répondant "oui" (supposés "difficiles") et "non" (supposés "plus faciles") à la question posée par le facteur f. Lorsqu'un seuil numérique est nécessaire, nous sélectionnons une valeur pour répartir le plus équitablement le corpus.
3. Pour chaque facteur f et modèle m, nous calculons la performance ROUGE-L du modèle m sur les partitions Ef et $\bar{E}f$: $RL(m, Ef)$ et $RL(m, \bar{E}f)$. Ensuite, nous calculons : $\delta = \lfloor (RL(m, Ef) - RL(m, \bar{E}f)) * 100 \rfloor$ représentant la dégradation en termes de points de ROUGE-L due au facteur de complexité f.

4. Enfin, nous calculons une mesure de significativité statistique pour δ avec le test U de Mann-Whitney avec un niveau de risque de 5% entre les deux partitions Ef et $\bar{E}_f$. Ce test prend en compte la valeur de δ et les caractéristiques de chaque ensemble dans la partition.

Facteurs de complexité sémantique. Pour cette étude, nous avons classé les facteurs en deux catégories : les facteurs génériques indépendants des relations sémantiques ($f0$ et $f1$) et ceux qui dépendent des relations sémantiques ($f2$ à $f6$). Pour ces derniers, nous avons été guidés par certains facteurs de complexité proposés pour l'analyse syntaxique automatique dans Frames dans (Marzinotto *et al.*, 2018).

$f0$: **biais dans le corpus d'adaptation.** La première explication pourrait être l'association entre la distribution des cadres sémantiques dans le corpus d'adaptation et les scores du modèle dans le corpus d'évaluation. Malgré l'absence de corrélation directe entre ces corpus, une grande quantité de certaines relations sémantiques dans le corpus FQuAD pourrait être corrélée avec la performance du modèle sur des questions/réponses présentant des relations similaires dans le corpus d'évaluation. Pour ce faire, nous avons automatiquement annoté le corpus FQuAD avec des cadres et classé les cadres en fonction de leur fréquence. Ensuite, nous divisons les exemples FQuAD en deux sous-ensembles, afin d'obtenir une représentation équilibrée en termes de fréquence des cadres : E_f comprend les exemples correspondant aux cadres les moins fréquents, tandis que $\bar{E}_f$ comprend le reste.

$f1$: **coréférence.** La nécessité de résoudre une coréférence constitue un facteur de complexité potentiel. Comme mentionné dans la section 2, les chaînes de coréférence sont annotées pour les arguments des relations sémantiques liant les questions et les réponses, ce qui nous permet de diviser le corpus en deux : les exemples nécessitant la résolution d'une chaîne de coréférence pour trouver la réponse E_f, et les autres $\bar{E}_f$.

$f2$: **nature de la relation sémantique du déclencheur.** Les déclencheurs d'un cadre sémantique dans le modèle FrameNet, appelés *Unité lexicale - LU*, peuvent être verbaux ou nominaux. Il a été démontré dans (Marzinotto *et al.*, 2018) que les relations déclenchées par une LU nominale sont plus difficiles à traiter. Nous avons donc divisé les exemples du corpus d'évaluation en fonction de la nature de la LU : soit nominale E_f, soit verbale $\bar{E}_f$.

$f3$: **présence du déclencheur du cadre sémantique dans la question.** Lorsque le même terme déclenche la relation sémantique dans le texte et apparaît dans la question, l'établissement d'un lien entre la question et la réponse est évidemment plus simple. Pour tenir compte de ce facteur, nous classons les exemples dans le sous-ensemble E_f lorsque le déclencheur est absent de la question et de la réponse, et dans $\bar{E}_f$ lorsque l'unité lexicale (LU) (ou l'une de ses inflexions) est présente dans les deux.

$f4$: **Distance entre le déclencheur et la réponse en termes d'arcs de dépendance.** La distance entre le déclencheur du cadre et la réponse dans le texte peut constituer un facteur de complexité, étant donné qu'une plus grande distance tend à être corrélée à un plus grand nombre d'attracteurs. Pour quantifier ce facteur, nous calculons la distance en termes d'arcs de dépendance à l'aide d'une

	Facteur de complexité						
modèles/facteurs	f0	f1	f2	f3	f4	f5	f6
proportion de E_f (%)	42%	6%	37%	45%	12%	59%	46%
CamemBERT	-1	-4	-1	-2	-7	-3	-1
T5	-1	-9	-2	-1	-7	-5	-2
FLAN	-2	-4	-3	-2	-4	-5	-3
MT5	0	-13	-1	-1	-10	-4	-2
llama-2	0	-3	-1	3	-3	-7	-2
mixtral-8x7b	0	1	-2	-1	-5	-6	0
GPT-3.5	0	4	-1	0	-4	-4	-3

TABLE 2 – Validation des facteurs de complexité : chaque cellule montre δ pour chaque modèle et facteur, avec des marquages jaunes pour différences significatives. La ligne "proportion" indique le pourcentage de chaque partition E_f dans le corpus total.

analyse syntaxique du corpus. Nous regroupons les exemples pour lesquels il existe au moins deux arcs de dépendance entre l'unité lexicale (LU) et la réponse dans le sous-ensemble E_f, et ceux pour lesquels il n'existe qu'un seul arc dans $\bar{E}_f$.

$f5$: **Nombre d'arguments dans le cadre.** Certaines relations sémantiques ont un plus grand nombre d'arguments (Frame Elements - FEs) que d'autres. La quantité d'éléments de cadre dans la relation sémantique sous-jacente à la paire question-réponse est également un facteur influençant la difficulté : un plus grand nombre d'éléments de cadre implique que le processus de liaison dispose de plus de contexte pour identifier avec précision la réponse. À l'inverse, un nombre réduit d'arguments rend la tâche plus ambiguë. Nous regroupons les exemples ne comportant pas plus de deux FE dans le sous-ensemble E_f, et ceux comportant plus de deux arguments dans $\bar{E}_f$.

$f6$: **Mesure de l'entropie de la distribution des LUs pour un cadre donné.** Certains cadres sont systématiquement déclenchés par les mêmes termes, tandis que d'autres présentent une diversité beaucoup plus grande, ce qui entraîne une ambiguïté dans leurs déclenchements. Cette mesure de la "surprise" peut être quantifiée par l'entropie de la distribution des LU dans le corpus d'entraînement. Une entropie plus élevée indique une plus grande ambiguïté dans le déclenchement des cadres. Nous incluons des exemples dans le sous-ensemble E_f pour les cas dont la valeur d'entropie est supérieure à un seuil α, et dans $\bar{E}_f$ pour les cas inférieurs au même seuil. Le seuil α est calculé comme la valeur moyenne de l'entropie sur l'ensemble du corpus.

5 Validation expérimentale

Le tableau 2 présente les résultats pour ces 7 facteurs de complexité. Dans chaque case, pour un modèle m et un facteur f, la valeur correspond à l'impact de f sur m exprimé par la différence en termes de points ROUGE-L δ présentée précédemment. Les cases colorées en jaune correspondent aux facteurs qui ont validé le test U de Mann-Whitney pour la significativité statistique avec un risque de 5

Comme on peut le voir, le facteur générique $f0$ correspondant au lien entre la fréquence d'une Frame

dans le corpus d'adaptation et dans le corpus d'évaluation a très peu d'influence sur les résultats.

En ce qui concerne $f1$, on constate que le fait de devoir résoudre une chaîne de coréférence est un facteur de complexité pour tous les modèles sans être significatif, mais qu'il n'a principalement un impact que pour les "petits" modèles. Dans les faits, même s'il existe une certaine perte de performance avec les LLMs, celle-ci est moins importante qu'avec les autres modèles, voir même un gain significatif dans le cas de GPT-3.5. Ceci suggère que les LLMs, par leur taille, ont une bien meilleure capacité à gérer les coréférences, au moins celles considérées dans cette tâche de question-réponse.

Parmi tous les autres facteurs, nous pouvons observer que la nature du déclencheur du cadre ($f2$) est effectivement un facteur de complexité pour tous les modèles, bien qu'il ne soit statistiquement significatif pour aucun d'entre eux. Le facteur $f3$ est également validé pour tous les modèles sauf LLAMA et GPT-3.5, mais il n'est significatif que pour CamemBERT. Concernant la distance entre le déclencheur et la réponse ($f4$), elle affecte négativement de façon plus importante les "petits" modèles, ce qui peut orienter vers l'hypothèse que les LLMs encodent mieux la structure syntaxique des phrases et modélisent ainsi la structure profonde des énoncés.

Les deux facteurs de complexité qui sont validés pour tous les modèles et qui sont pour la plupart statistiquement significatifs sont le nombre d'éléments du cadre dans la relation sémantique ($f5$) et la mesure de l'entropie dans la distribution des déclencheurs de ces mêmes relations ($f6$).

Il est intéressant de noter que les facteurs les plus fiables sont ceux les plus étroitement liés à la définition des cadres (nombre d'arguments pour $f5$ et entropie dans la distribution des déclencheurs pour $f6$) plutôt qu'à leur utilisation dans un contexte particulier (choix de la forme syntaxique du déclencheur dans $f2$, répétition du déclencheur dans la question dans $f3$, et complexité de l'arbre syntaxique dans $f4$). Par conséquent, les facteurs $f5$ et $f6$ peuvent être assimilés à une mesure de l'ambiguïté sémantique intrinsèque dans les relations question/réponse.

Cela peut être illustré par quelques exemples de notre corpus. Par exemple, le Frame *Request* peut avoir plus de 20 déclencheurs dans le lexique de Berkeley Framenet [6]. Dans notre corpus d'entraînement, il compte 118 occurrences avec 18 déclencheurs différents, résultant en une des mesures d'entropie les plus élevées, et un score ROUGE-L variant de 0,55 à 0,84 selon le modèle.

En contraste, la Frame *Installing*, défini comme "Un *Agent* place un *Composant* dans un *Emplacement Fixe* de sorte que le *Composant* est attaché et interconnecté et par là fonctionnel", a seulement deux déclencheurs dans le dictionnaire Framenet, *installer* et *installation*. Il compte 58 occurrences dans notre corpus d'entraînement avec 2 déclencheurs principaux et est l'un des cadres *faciles* avec une entropie faible et un score ROUGE-L variant de 0,79 à 0,90 selon le modèle.

De même, ($f5$) démontre que certains cadres présentent un nombre moyen inhabituellement bas d'éléments de cadre dans leurs exemples (≤ 2). Par exemple, le Frame *Origin* contient seulement deux FEs essentiel s(*Origin* et *Entity*), sans FEs non-essentiels présents dans Berkeley Framenet. Par contraste, les Frames *Giving* et *Contacting* comportent respectivement trois et cinq FEs essentiel, ainsi que de nombreux FEs non-essentiels dans FrameNet. Ce schéma reflète le phénomène observé avec $f6$, où le Frame *Origin* obtient un score en dessous de la moyenne tandis que *Contacting* et *Giving* sont classifiés comme Frames 'faciles'.

6. https://framenet.icsi.berkeley.edu/frameIndex

6 Travaux connexes

Notre travail se situe dans le domaine de l'évaluation des modèles. Notre approche contraste avec les évaluations à grande échelle qui couvrent plusieurs tâches, corpus et langues (Laskar *et al.*, 2023; Liang *et al.*, 2023; Srivastava *et al.*, 2023; Brown *et al.*, 2020; Wang *et al.*, 2019). Il se rapporte à des études ciblées traitant de phénomènes linguistiques spécifiques tels que les négations (Truong *et al.*, 2022, 2023; Zhang *et al.*, 2023; Ravichander *et al.*, 2022), l'ambiguïté dans les tâches d'inférence (Liu *et al.*, 2023), et l'extraction d'informations ouverte (Lechelle *et al.*, 2019), qui utilisent des ensembles de données petits et méticuleusement organisés pour évaluer avec précision les capacités des modèles pour la tâche. Notre étude fait écho à cette dernière, en explorant des évaluations linguistiques ciblées.

Cette étude est également liée à d'autres efforts de recherche qui ont été dirigés vers l'évaluation de la fiabilité des LLMs "fermés" accessibles uniquement via une API comme ChatGPT sur des benchmarks dans le domaine de la question-réponse basée sur la connaissance (KBQA) (Tan *et al.*, 2023), ainsi que son applicabilité générale à travers un large éventail de tâches NLP (Kocoń *et al.*, 2023; Laskar *et al.*, 2023). Ces études suggèrent que bien que ChatGPT présente une performance robuste sur un ensemble de tâches très large et diversifié, il peut également être devancé par des modèles spécialisés spécifiques à la tâche. Nous avons trouvé des résultats similaires dans notre étude mais montrons également que malgré ses forces, un modèle tel que ChatGPT peut être sensible aux mêmes facteurs de complexité qu'un modèle beaucoup plus petit comme T5 ou même CamemBert.

Dans l'ensemble, cette étude promeut l'idée que nous avons besoin d'un cadre d'évaluation plus précis et peut être reliée à d'autres études telles que (Ribeiro *et al.*, 2020) qui identifient des *échecs critiques* dans les modèles commerciaux et à la pointe de la technologie en proposant une méthodologie de test agnostique au modèle et à la tâche ou (Gehrmann *et al.*, 2023) insistant sur le fait que pour comparer les modèles, nous avons besoin d'un processus d'annotation plus "*soigné [...] pour caractériser leur qualité de sortie et les distinguer*".

7 Conclusion

Dans cette étude, nous avons mené une expérience utilisant un corpus de questions-réponses annoté sémantiquement pour identifier des facteurs de complexité sémantique inhérents à cette tâche, indépendamment de l'architecture et de la taille des modèles. Cette investigation est cruciale car elle souligne les avantages potentiels de se concentrer sur des modèles avec moins de paramètres pour gérer des ensembles de données difficiles. Il convient de noter que les gains de performance réalisés avec des modèles plus petits pourraient être éclipsés par des modèles plus grands. En partitionnant les ensembles de données basés sur des phénomènes linguistiques d'une complexité équivalente, indépendamment du modèle, nous pouvons nous attendre à ce que les améliorations de performance se généralisent à travers les modèles. Nos résultats démontrent que les facteurs de complexité sémantique identifiés catégorisent efficacement le corpus en sous-corpus de niveaux de difficulté variables, indépendamment du modèle employé.

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1 : The 17th International Conference on Computational Linguistics*.

BÉCHET F., ALOUI C., CHARLET D., DAMNATI G., HEINECKE J., NASR A. & HERLEDAN F. (2019). Calor-quest : generating a training corpus for machine reading comprehension models from shallow semantic annotations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, p. 19–26.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

GEHRMANN S., CLARK E. & SELLAM T. (2023). Repairing the cracked foundation : A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, **77**. DOI : 10.1613/jair.1.13715.

JIANG A. Q., SABLAYROLLES A., ROUX A., MENSCH A., SAVARY B., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., HANNA E. B., BRESSAND F. *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv :2401.04088*.

KHASHABI D., CHATURVEDI S., ROTH M., UPADHYAY S. & ROTH D. (2018). Looking beyond the surface : A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 252–262.

KOCOŃ J., CICHECKI I., KASZYCA O., KOCHANEK M., SZYDŁO D., BARAN J., BIELANIEWICZ J., GRUZA M., JANZ A., KANCLERZ K. *et al.* (2023). Chatgpt : Jack of all trades, master of none. *Information Fusion*, p. 101861.

LASKAR M. T. R., BARI M. S., RAHMAN M., BHUIYAN M. A. H., JOTY S. & HUANG J. (2023). A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 431–469, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.29.

LECHELLE W., GOTTI F. & LANGLAIS P. (2019). WiRe57 : A fine-grained benchmark for open information extraction. In A. FRIEDRICH, D. ZEYREK & J. HOEK, Éds., *Proceedings of the 13th Linguistic Annotation Workshop*, p. 6–15, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4002.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A. & BENJAMIN NEWMAN E. A. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

LIU A., WU Z., MICHAEL J., SUHR A., WEST P., KOLLER A., SWAYAMDIPTA S., SMITH N. & CHOI Y. (2023). We're afraid language models aren't modeling ambiguity. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 790–807, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.51.

MARTIN L., MULLER B., SUÁREZ P. J. O., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D., SAGOT B. *et al.* (2020). Camembert : a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

MARZINOTTO G., BÉCHET F., DAMNATI G. & NASR A. (2018). Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan. HAL : hal-01731385.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

RAVICHANDER A., GARDNER M. & MARASOVIC A. (2022). CONDAQA : A contrastive reading comprehension dataset for reasoning about negation. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 8729–8755, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.598.

RIBEIRO M. T., WU T., GUESTRIN C. & SINGH S. (2020). Beyond accuracy : Behavioral testing of NLP models with CheckList. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4902–4912, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.442.

SRIVASTAVA A., RASTOGI A., RAO A., SHOEB A. A. M., ABID A., FISCH A., BROWN A. R., SANTORO A., GUPTA A. & ADRIÀ GARRIGA-ALONSO E. A. (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

TAN Y., MIN D., LI Y., LI W., HU N., CHEN Y. & QI G. (2023). Can chatgpt replace traditional kbqa models ? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, p. 348–367 : Springer.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

TRUONG T. H., BALDWIN T., VERSPOOR K. & COHN T. (2023). Language models are not naysayers : an analysis of language models on negation benchmarks. In A. PALMER & J. CAMACHO-COLLADOS, Éds., *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, p. 101–114, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.starsem-1.10.

TRUONG T. H., OTMAKHOVA Y., BALDWIN T., COHN T., LAU J. H. & VERSPOOR K. (2022). Not another negation benchmark : The NaN-NLI test suite for sub-clausal negation. In Y. HE, H. JI, S. LI, Y. LIU & C.-H. CHANG, Éds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 883–894, Online only : Association for Computational Linguistics.

WANG A., PRUKSACHATKUN Y., NANGIA N., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, **32**.

WEI J., BOSMA M., ZHAO V. Y., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv :2109.01652*.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mt5 : A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 483–498.

ZHANG Y., YASUNAGA M., ZHOU Z., HAOCHEN J. Z., ZOU J., LIANG P. & YEUNG S. (2023). Beyond positive scaling : How negation impacts scaling trends of language models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 7479–7498, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.472.