ANTICIPER LES PROCÉDURES INFÉRENTIELLES PAR L'UTILISATION DE GRANDS MODÈLES DE LANGAGE

MÉMOIRE présenté pour l'obtention du MASTER

Métiers de l'enseignement, de l'éducation et de la formation

Mention second degré

Parcours mathématiques

Par

BALI Alexandre

Sous la direction de

MERCAT Christian

Membres du jury

NOM Prénom

NOM Prénom

Année universitaire 2025 – 2026

Date de soutenance : —/—/2026

Table des matières

1. Introduction

Au cours des expériences pédagogiques auxquelles j'ai pris part durant ma maîtrise en Métiers de l'Enseignement, de l'Éducation et de la Formation, mention mathématiques, que ce soit lors de mon stage d'observation et de pratique accompagnée au lycée La Martinière-Montplaisir, de mes permanences de tutorat en mathématiques, ou encore des séances que j'ai conduites en tutorat de mathématiques pour première année de licence Mathématiques-Informatique en parcours progressif, j'ai pu constater la croissance, sinon la proéminence, de l'utilisation des intelligences artificielles génératives, notamment les grands modèles de langage (*large language models*, LLM), principalement par le biais de la plateforme web *ChatGPT*, dans la résolution de problèmes. Les discussions que j'ai eues avec mon tuteur de stage et les autres professeurs de l'établissement, ainsi que les médias récents, corroborent cette tendance, aussi bien en mathématiques que dans les autres disciplines scolaires, suscitant des débats autour de leur place dans l'enseignement secondaire : ils mettent aussi bien en avant leur potentiel éducatif, qu'ils en dénoncent les usages à mauvais escient qu'en font les élèves pour tricher sur leurs devoirs et évaluations. Pour ne citer que quelques articles de presse :

- Diallo, K. (2024, juin 17). *« Je ne vois pas où est le problème » : les enseignants divisés face à ChatGPT*. BFM.
- Pineau, C. (2024, octobre 23). *« C'est très efficace » : ce prof a une technique infaillible pour démasquer les devoirs faits avec ChatGPT*. Le Figaro.
- Rahmil, D.-J. (2024, juillet 10). *Lutter ou accompagner ? Le choix difficile des profs face à ChatGPT*. L'ADN.

Face à cette réalité, la question de l'intégration de ces outils dans l'enseignement se pose alors. **Comment ces IA génératives peuvent être mobilisées de façon constructive en tant qu'enseignants des mathématiques ?** Dans ce mémoire, nous examinerons la pertinence des LLM dans l'analyse *a priori* de situations didactiques que nous, enseignants de mathématiques, pouvons proposer aux élèves pour les conduire à employer des raisonnements mathématiques : devoirs surveillés, question *flash*, exercices d'entraînement, etc.

En troisième année de licence, j'ai déjà pu étudier, pour mon rapport de stage en établissement au lycée de la Côtière, la mobilisation autonome du raisonnement

par l'absurde chez des élèves de seconde. Cette thématique m'avait permis de mettre en lumière quelques procédures logiques employées par les élèves. J'ai donc décidé d'étendre cette réflexion en orientant ma recherche sur l'analyse des inférences élèves, en comparant, d'une part, celles qu'on peut obtenir des réponses et productions d'élèves avec, d'autre part, celles que les utilisations que nous ferons des LLM mèneront à prévoir. Ce travail vise ainsi à mieux juger de la pertinence de l'utilisation des IA génératives dans la mise en œuvre réfléchie de situations didactiques dans l'enseignement des mathématiques.

2. Partie théorique

2.1. Rendre médiat l'immédiat : quels appuis théoriques pour identifier les inférences élèves ?

Dans ce mémoire, on se basera principalement sur la définition d'*inférences* donnée par Françoise Armengaud (*Encyclopædia Universalis*), une inférence est une procédure :

> « qui passe de propositions assertives, comme prémisses, à des propositions assertives, comme conclusions. Au sens strict, on distingue l'inférence du raisonnement en ce qu'elle peut être soit médiate soit immédiate [···] tandis que le raisonnement comporte nécessairement des médiations (il est discursif). On distingue aussi l'inférence démonstrative ou déduction [···] et l'inférence non démonstrative [···]. L'inférence déductive est qualifiée de valide ou de non valide ; cette qualification ne se retrouve pas exactement [autrement] [···]. L'inférence valide s'effectue conformément à des règles qui permettent de déduire d'un ensemble de prémisses toutes les conséquences logiques et elles seulement. » (Armengaud, 2006)

Ainsi, une inférence n'est pas nécessairement valide, elle peut même être non déductive, – par exemple, inductive ou abductive (Eduscol, 2019a), – et ne même pas être explicitée par l'élève.

L'intérêt principal que j'ai à étudier les inférences, plutôt que simplement les raisonnements, c'est pour pouvoir mieux rendre compte par la suite des limitations qu'imposent les grands modèles de langage.

Pour en parler, il faut cependant que je présente au préalable les diverses inférences utilisées dans le secondaire, qui seront par définition médiates, et donc, dans la nomen-

clature proposée par Armengaud, 2006, des raisonnements. Nous devrons alors d'abord en expliciter quelques uns.

2.1.1. Inférences utilisées dans le secondaire

Hormis les procédures personnelles des élèves, divers raisonnements sont utilisés dans le secondaire, que l'on retrouve notamment dans Eduscol, 2016; 2019a; 2019b :

- Raisonnement(s) par l'absurde,
- Disjonction de cas,
- *Modus ponens*,
- Double implication,
- Raisonnement par équivalence,

et autres. Dans ces brochures, ces raisonnements sont explicités dans un registre langagier : par exemple, *le* raisonnement par l'absurde dans Eduscol, 2019b, p. 28 est décrit ainsi : « Pour démontrer une proposition P, on suppose que sa négation (non P) est vraie, on en déduit des conséquences qui aboutissent à une contradiction. Cela montre que (non P) ne peut pas être vraie, et par conséquent que P est vraie. »

Cependant, dans le domaine de la logique mathématique, qui étudie précisément ces inférences, celles-ci sont souvent exprimés très différemment, principalement via la notion de règles d'inférence. Nous allons tenter d'intégrer quelques apports de la logique mathématique pour nous aider à mieux formaliser cette notion d'inférence.

2.1.2. Le filtrage par motifs comme base des inférences mathématiques

Les opérations et inférences utilisées en mathématiques scolaires proviennent de la logique formelle classique.

Or, selon Gao, 2017, p. 19, "proof is syntactic" [« une preuve est syntaxique »] en logique classique ; dans ce contexte, une preuve :

- commence avec une collection de prémisses,
- transforme ces prémisses selon une collection de règles d'inférence (via filtrage par motif),
- parviennent alors à une conclusion.

Cela rejoint effectivement la définition de ce qu'est une inférence donnée par Armengaud, 2006. Cette idée de "pattern matching" [« filtrage par motif »] rejoint quant à elle l'idée que certaines propositions suivent un schéma spécifique.

Cette approche syntaxique fonde le champ de la déduction naturelle, utilisée dans Gao, 2017, mais mise plus particulièrement en exergue dans Magnus et al., 2023. Ces approches tendent à utiliser des connecteurs logiques dans un registre symbolique plutôt que langagier ($\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$, $\neg$). Parmi ces symboles, $\Leftrightarrow$, tout particulièrement, peut être présent dans certains manuels, comme *Math'x Seconde*, 2010, p. 351, ainsi que dans *Déclic 2nde*, 2010, p. 329 (Grenier et al., 2019, p. 11).

2.1.3. Formalisation des inférences de la logique classique par la notation de Fitch

Magnus et al., 2023 nous servira de guide principal pour adopter une formalisation plus usuelle et systématique des schémas inférentiels. Celui-ci utilise ce qu'on appelle la notation de Fitch pour les preuves, qui est construite de la sorte :

- On pose initialement trois colonnes, la première et la deuxième délimitées l'une de l'autre par une ligne verticale :
  - Celle de gauche est composée d'entiers, en partant de 1, qui incrémentent d'une ligne à la suivante, permettant d'expliciter le numéro de chaque étape de la preuve.
  - Celle du milieu est composée des propositions, qu'elles soient justifiées ou admises.
  - Celle de droite, qui explicitent la justification utilisée, souvent avec des initiales ou une abréviation symbolique d'un côté, et de l'autre les numéros des étapes utilisées, conjointes par des virgules.
- Chaque hypothèse ajoutée, au-delà des conditions initiales, crée une nouvelle indentation délimitée par une ligne verticale à droite, ainsi qu'un trait horizontal en-dessous de l'hypothèse pour spécifier qu'il s'en agit d'une ; on parle alors de la sous-preuve associée à la proposition érigée en supposition.
- Chaque sous-preuve peut n'être citée, dans la colonne de droite, qu'à la couche précédente en terme d'imbrication de sous-preuves. En posant i la ligne sur laquelle se trouve l'hypothèse, et j une proposition sous cette hypothèse supplémentaire, et surtout, sans hypothèses supplémentaires, alors on citera ce bloc comme i–j.
- Chaque citation dans la colonne de droite doit faire appel uniquement aux lignes dans la même sous-preuve, ou dans les sous-preuves où elle se retrouve imbriquée.

Par exemple, voici une démonstration de la transitivité de l'implication matérielle ($\Rightarrow$) :

1	$A \Rightarrow B$	
2	$B \Rightarrow C$	
3	$\quad A$	
4	$\quad B$	$\Rightarrow$E 1, 3
5	$\quad C$	$\Rightarrow$E 2, 4
6	$A \Rightarrow B$	$\Rightarrow$I 3–5

Dans Magnus et al., 2023, on résume le séquent (théorème) démontré ici en juxtaposant les hypothèses initiales par des virgules, puis avec la conclusion avec $\therefore$ entre les deux, ce qui donne ici : « $A \Rightarrow B,\ B \Rightarrow C \therefore A \Rightarrow C$ ».

2.1.4. Comparaison entre les raisonnements du secondaire identifiés avec diverses inférences classiques

On retrouve de nombreuses inférences qui correspondent, pour certains, à différents usages des raisonnements établis dans les brochures d'Eduscol :

- Le raisonnement par l'absurde, comme défini précédemment dans Eduscol, 2019b, p. 28, est pratiquement identique à la règle de la preuve indirecte qui se trouve dans Magnus et al., 2023, p. 139. Cependant, les exemples donnés semblent davantage se ramener à l'introduction de la négation (p. 138) : en effet, « La fonction ‹ racine carrée › n'est pas dérivable en 0. » signifie que « non ($\sqrt{\cdot}$ dérivable en 0) », et pour le démontrer, on suppose « $\sqrt{\cdot}$ dérivable en 0 », ce qui n'est *a priori* pas identique à « non (non ($\sqrt{\cdot}$ dérivable en 0)) ».
- La disjonction de cas, telle que présentée dans Eduscol, 2019a, p. 22 ; Eduscol, 2019b, correspond à la loi du tiers exclu (Magnus et al., 2023, p. 167), puisque les cas considérés doivent y être disjoints. Or, tel que relate Battie, 2003, p. 44, certains ouvrage font fi de cette condition de cas disjoints, comme dans Arnaudiès & Fraysse, 1987 ; il s'agit alors d'une élimination de la disjonction, définie dans Magnus et al., 2023, p. 135, étant même référencée comme « proof by cases » [« disjonction de cas »] à la page 180.
- La « déduction », ou *modus ponens*, telle que définie dans Eduscol, 2016, p. 2, rejoint précisément l'élimination de la conditionnelle, définie dans Magnus et al., 2023, p. 123, et dont il est également mentionné qu'elle « is also sometimes called *modus ponens* » [« est aussi parfois appelée *modus ponens* »], pour davantage justifier leur équivalence.

- La double implication, définie dans Eduscol, 2019a, p. 16, cherche à démontrer une équivalence entre deux assertions, en montrant d'abord un sens puis un autre. Il s'agit dans Magnus et al., 2023, p. 132 de l'introduction de la biconditionnelle. Elle est à distinguer du raisonnement par équivalence, également définie dans Eduscol, 2019a, p. 16, et usant principalement de la transitivité de la biconditionnelle, qui est l'exercice C. 2. de Magnus et al., 2023, p. 143.

Récapitulons un peu ce qu'on a fait jusqu'ici. On a réussi à identifier diverses inférences utilisées dans le secondaire, et nous avons établi un lien entre ceux-ci et les inférences formelles de la logique classique, plus spécifiquement de la déduction naturelle, afin de nous donner un cadre plus rigoureux pour étudier et, espérons-le, identifier/classifier les inférences employées par les élèves.

Le problème, désormais, c'est que les inférences élèves, et même certains raisonnements mobilisés et parfois encouragés – dans une certaine mesure, tout du moins – en mathématiques au secondaire (je veux bien sûr parler des raisonnements inductifs), ne suivent pas toujours le cadre de la logique classique. La section suivante sert alors à montrer diverses inférences qui rentrent en conflit avec ce système classique.

2.1.5. Logique *naturelle*, polysémie et contre-classicismes : mise à l'épreuve de la notation de Fitch en-dehors de la logique classique

Un cas de polysémie, que nous avions abordé précédemment, est la question de la disjonction, du ‹ ou ›. Dans Eduscol, 2019b, p. 22, il semble que la disjonction soit interprétée comme exclusive :

> « On démontre la propriété sur des sous-ensembles E_1, E_2… disjoints deux à deux (‹ disjonction ›), dont la réunion est égale à E. Le professeur peut à cette occasion introduire le terme ‹ partition ›. [···] La validité du raisonnement repose sur le fait qu'on a bien une partition, [···] »

Pourtant, les disjonctions telles qu'utilisées par Gao, 2017, p. 7 ; Magnus et al., 2023, p. 70 sont inclusives : par exemple, on peut parfaitement traiter une disjonction de cas sur $a \vee a$. Il s'agit donc d'un désaccord, notable de par son inclusion dans les programmes du lycée (Ministère de l'Éducation nationale et de la Jeunesse, 2020, p. 21).

Quant à l'implication, Deloustal-Jorrand, 2001, p. 38, émet « l'hypothèse que [···] les élèves [utilisent] la propriété-en-acte suivante : ‹ $A \Rightarrow B$ *n'a pas d'intérêt lorsque* A *est fausse* ›. De là, peut aussi découler la propriété-en-acte : ‹ $A \Rightarrow B$ *est fausse*

lorsque A est fausse ›. » Cette règle, qu'elle abrège en P3b (p. 57), fait écho à une règle médiévale parfois appelée « *ex falso nihil sequitur* » [« rien ne suit d'une contradiction »], notamment présente chez les disciples du théologien Robert de Melun (1095–1167) (Lenzen, 2022). La notation de Fitch est suffisamment robuste pour ériger la formulation qu'en fait Deloustal-Jorrand en tant qu'inférence formelle :

$$\begin{array}{l|ll} m & \neg\mathscr{A} & \\ & \neg(\mathscr{A} \Rightarrow \mathscr{B}) & \text{P3b}\ \ m \end{array}$$

Pourtant, si on adjoint cette règle à la logique classique, on obtient que de toute formule φ, on peut démontrer $\therefore \varphi$, donc en particulier toute contradiction :

$$\begin{array}{ll|ll} 1 & \neg\varphi & & \\ \hline 2 & \neg(\varphi \Rightarrow \bot) & \text{P3b} & 1 \\ 3 & \quad \varphi & & \\ \hline 4 & \quad \bot & \neg\text{E} & 1,3 \\ 5 & \varphi \Rightarrow \bot & \Rightarrow\text{E} & 3\text{–}4 \\ 6 & \bot & \neg\text{E} & 2,5 \\ \blacksquare & \varphi & \text{IP} & 1\text{–}6 \end{array}$$

Uune règle, telle qu'P3b, qui lorsqu'adjointe à la logique classique engendre un système incohérent, est dite contre-classique. La notation de Fitch est alors suffisamment robuste pour identifier des inférences élèves qui ne sont pas valides classiquement.

Cela ne signifie pas que tout système logique admettant P3b pour théorème est incohérent. Il existe une famille de systèmes cohérents qui admettent P3b en théorème. En posant $\mathcal{L}$ le langage des inférences propositionnelles en notation de Fitch, on pose $\pi : \mathcal{L} \to \mathcal{L}$ l'application qui permute les symboles $\wedge$ et $\Rightarrow$, et laisse intacte le reste, $\pi(\text{P3b})$ devient un théorème de la logique classique :

$$\begin{array}{ll|ll} 1 & \pi(\neg\mathscr{A}) & & \\ \hline 2 & \neg\pi(\mathscr{A}) & \pi & 1 \\ 3 & \quad \pi(\mathscr{A}) \wedge \pi(\mathscr{B}) & & \\ \hline 4 & \quad \pi(\mathscr{A}) & \wedge\text{E} & 2 \\ 5 & \quad \bot & \neg\text{E} & 1,3 \\ 6 & \neg(\pi(\mathscr{A}) \wedge \pi(\mathscr{B})) & \neg\text{I} & 3\text{–}5 \\ 7 & \neg(\pi(\mathscr{A} \Rightarrow \mathscr{B})) & \pi & 6 \\ \blacksquare & \pi(\neg(\mathscr{A} \Rightarrow \mathscr{B})) & \pi & 7 \end{array}$$

Ainsi, pour tout système $\mathfrak{S}$ de théorèmes de la logique classique, et par cohérence de la logique classique, le système $\pi(\mathfrak{S}) \cup \{\text{P3b}\}$ est à son tour cohérent, puisqu'équivalent à $\mathfrak{S} \cup \{\pi(\text{P3b})\}$ un système de théorèmes de la logique classique.

2.1.6. Conclusion des apports théoriques sur les inférences

En somme, la conversion d'assertions de logique *naturelle* (Deloustal-Jorrand, 2001, p. 36) en propositions de logique formelle est non-triviale. Or, puisque le formalisme classique n'est pas dans les programmes de mathématiques dans le secondaire, nous devons nous attendre à ce que les élèves emploient de la logique *naturelle* en premier lieu, ce qui indique la pertinence de ce questionnement dans la portée de ce mémoire. De nombreuses tentatives ont été produites : pour ne citer qu'un exemple, « Durand-Guerrier s'est [···] attachée à *réconcilier* la logique formelle et la logique naturelle » (Deloustal-Jorrand, 2001, p. 36) dans le contexte de l'enseignement de la logique. Cependant, la motivation de ce mémoire est autre : nous cherchons à identifier les procédures inférentielles qu'emploient les élèves, qu'elles soient déductives et valides comme celles que nous enseignons, ou déductives mais non-valides, ou inductives, ou abductives, etc.

Cela nous apporte déjà beaucoup de matière pour déterminer, par la suite, comment nous établirons des critères de détermination d'utilisation d'une inférence ou une autre dans la partie méthodologie. Notamment, on sait que la notation de Fitch est suffisamment costaude pour formaliser des inférences non-classiques, voire contre-classiques. Nous baserons alors notre méthodologie d'identification des inférences élèves en passant par la formalisation en notation de Fitch, ce que nous devrons détailler dans notre partie méthodologie.

On passe maintenant à la partie modèles de langage.

2.2. Modèles de langage, grands modèles de langage : pistes pour leur exploitation dans l'identification des procédures élèves

Un modèle de langage, selon Hiemstra, 2009 :

> "assigns a probability to a piece of [···] text, based on some training data. For example, a language model based on a big English newspaper archive is expected to assign a higher probability to "a bit of text" than to "aw pit tov tags," because the words in the former phrase [···] occur more frequently in the data than the words in the latter phrase."
>
> [« assigne des probabilités à des bouts de texte [···], basés sur des données d'entraînement. Par exemple, un modèle de langage entraîné sur un répertoire d'archives d'un journal anglais conséquent devrait assigner une probabilité plus grande à ‹ a bit of text › qu'à ‹ aw pit tov tags ›, car les mots dans la première expression [···] apparaissent plus fréquemment dans les données que les mots de la dernière expression. »]

Ainsi, il est possible de construire des modèles de langage aussi simplistes que l'on souhaite. Cependant, comme nous le verrons dans notre étude d'un modèle markovien, nous dégagerons divers points qu'il a fallu améliorer afin de générer des expressions plus crédibles, ce qui nous mènerons à comprendre les bases du fonctionnement des grands modèles de langage.

2.2.1. Modèles markoviens et enjeux premiers

Les modèles markoviens du langage font partie des plus anciens, recensés notamment dans Chomsky, 1956. Selon Langville & Hilgers, 2006,

> "Chomsky argued that language models based on Markov chains do not capture some nested structures of sentences, which are quite common in many languages such as English. We now recognize that Chomsky's account of the limitations of Markovian language models was too harsh. The early 1980s saw a resurgence of the success of Markovian language models in speech recognition."
>
> [« Chomsky a défendu que les modèles de langage basés sur les chaînes de Markov ne captureraient pas certaines structures d'imbrication phrastiques, qui

sont assez courantes dans bien des langues telles que l'anglais. On sait désormais que le rapport de Chomsky sur les limitations des modèles de langage markoviens était trop sévère. Le début des années 1980 vit un regain du succès des modèles de langage markoviens dans la reconnaissance vocale. »]

De par la pertinence historique de l'utilisation des chaînes de Markov dans les modèles de langage, nous étudierons d'abord un modèle de langage markovien assez primitif comme suit :

```
def markovian_language_model (text, prompt=".", n=100) :
   from re import findall
   from numpy import unique, array, zeros, cumsum, searchsorted
   from random import random
   from warnings import filterwarnings
   pattern = r'\w+|[^\w\s]'
   states = findall(pattern, text)
   unique_states = list(unique(states))
   state_index = {state: i for i, state in enumerate(unique_states)}
   transitions = zeros((len(unique_states), len(unique_states)))
   for i in range(len(states) - 1) :
      transitions[state_index[states[i]], state_index[states[i+1]]] += 1
   filterwarnings("ignore")
   transitions /= transitions.sum(axis=1, keepdims=True)
   filterwarnings("default")
   output = [prompt]
   for _ in range(n) :
      current_index = state_index[output[-1]]
      if str(transitions[current_index][0]) == "nan" :
         return output
      if not any(transitions[current_index]) :
         break
      rand = random()
      cumm_prob = cumsum(transitions[current_index])
      next_state = unique_states[searchsorted(cumm_prob, rand)]
      output.append(next_state)
   return ' '.join(output)
```

Ce programme permet de générer procéduralement des mots et ponctuations à partir des diverses variables d'entrée du programme. La variable `text` sert de données textuelles d'entraînement, qui est ensuite segmenté en une séquence de mots et de ponctuations, qui deviennent les états de notre chaîne de Markov, les probabilités assignées à chaque transition suivant la distribution empirique obtenue par cette segmentation de `text`. La variable `prompt`, quant à elle, donne l'état de départ pour le texte qu'il génère : celui-ci doit être présent dans les données d'entraînement. Une fois cela fait, le modèle génère un texte de `n` tels états, séparés par des espaces dans le retour du programme, en suivant les états et probabilités des transitions générées par `text`.

En prenant `text` le texte de l'introduction de ce mémoire, converti en minuscules et sans sauts à la ligne, et en lui donnant également pour `prompt` « la », et pour `n` la valeur 50, on peut obtenir la sortie suivante :

```
la résolution de problèmes . comment ces outils dans les conduire
à mauvais escient qu ' utilisation des réponses et de tutorat de
seconde . en tant qu ' absurde chez des llm ) , pour première année de
l ' enseignants de façon constructive en dénoncent les élèves pour les
```

On peut d'emblée repérer divers points susceptibles d'être améliorés. D'une part, bien que le programme soit bien construit pour s'assurer que les paires de deux états consécutifs soient en succession réaliste, les séquences de trois états ou plus peuvent déjà donner lieu à des incohérences linguistique, comme les exemples de « l'enseignants » (où ‹ l ›, ‹ ' › et ‹ enseignants › sont trois états distincts), ou encore « dans les conduire ». De plus, de par la spécificité et la faible taille du texte d'entraînement, le vocabulaire disponible dans les états de la chaîne de Markov engendrée par sa segmentation est fortement limité, à peine plus de 200 mots uniques. Enfin, pour revenir à ce que nous disions précédemment, le fonctionnement du programme est très limité sur le plan des paramètres capables d'être manipulés. Tout cela mène à des limitations telles qu'avoir pour *prompt* un des états spécifiques de la chaîne de Markov engendrée par le texte d'entraînement, là où l'on pourrait s'attendre à ce que l'utilisateur emploie des néologismes, ou qu'il s'exprime des langues omises par le programme.

Ainsi, les modèles de langage plus sophistiqués essaient de prendre en compte autant que possible le contexte grammatical et syntaxique ambiant autour de chaque mot, ponctuation, et syntagme, segmenté à partir des données d'entraînement, lesquelles

devraient par ailleurs être bien plus fournies afin de donner au modèle un vocabulaire plus enrichi. Pour cela, l'introduction d'une approche de ‹ reconnaissance de motif › plus sophistiquée est primordiale pour produire des textes plus cohérents et pertinents, capacité que l'on peut construire avec la présence d'un grand nombre de paramètres. Nous avons alors établi trois axes d'amélioration principaux : la taille du corpus, la cohérence des textes générés, et le nombre de paramètres. Nous verrons comment ces points se résolvent dans le cas des grands modèles de langage.

2.2.2. Grands modèles de langage, température

Un grand modèle de langage est caractérisé ainsi dans Minaee et al., 2024, p. 1-2 :

> "Large language models (LLMs) mainly refer to transformer-based neural language models that contain tens to hundreds of billions of parameters, which are pre trained on massive text data, such as PaLM, LLaMA, and GPT-4, [···]. [···] LLMs are not only much larger in model size, but also exhibit stronger language understanding and generation abilities, and more importantly, emergent abilities that are not present in smaller-scale language models."
>
> [« Les grands modèles de langage (LLM) sont principalement les modèles de langage neuronaux à couches transformatives qui contiennent de dizaines à centaines de milliards de paramètres, lesquels sont entraîné au préalable sur des données textuelles massives, comme pour PaLM, LLaMA, et GPT-4, [···]. [···] Les LLM ont non seulement une taille de modèle plus grande, mais manifestent aussi de plus fortes compréhension et capacités de génération langagières ainsi que, plus notoirement encore, des capacités émergeantes qui ne sont pas présentes dans des modèles de langage plus réduits. »]

La cohérence de ces modèles peut être, par ailleurs, contrôlée par divers réglages qui influencent certaines des couches transformatives du modèle. C'est notamment le cas du réglage appelé « température ». Peeperkorn et al., 2024 le décrit ainsi :

> "The temperature parameter of an LLM regulates the amount of randomness, leading to more diverse outputs; therefore, it is often claimed to be the creativity parameter." (p. 1) "Temperature is a hyperparameter t that we find in stochastic models to regulate the [distribution of] a sampling process [···] its [distribution's] shape, redistributing the output probability mass, flattening the

distribution proportional to the chosen temperature." (p. 3) "We [···] observe a [···] moderately negative correlation between coherence and temperature" (p. 8) [« Le réglage température d'un LLM régule le seuil d'aléa, conduisant à des résultats plus variés ; ainsi, on affirme souvent qu'il s'agit du réglage de créativité. » (p. 1) « La température est un hyperparamètre t présent dans les modèles stochastiques pour réguler la loi de probabilité d'un échantillonage [···] la forme de sa loi, redistribuant sa fonction de masse résultante, de sorte à aplatir la loi proportionnellement à la température choisie. » (p. 3) « On [···] observe une [···] corrélation modérément négative entre cohérence et température. » (p. 8)]

De fait, si l'on souhaite obtenir des résultats plus variés, et des analyses *a priori* plus compréhensives, nous pourrions utiliser des températures plus hautes. Cependant, il faut également s'attendre, due à la « corrélation modérément négative entre cohérence et température », à des résultats moins cohérents en tout et pour tout.

Il faut cependant que nous gardions quelques réserves quant à l'impact de la température sur la créativité des résultats : "The influence of temperature is far more nuanced and weak than 'the creativity parameter' claim suggests." [« L'influence de la température est bien plus faible et nuancée que cette appellation de ‹ paramètre de créativité › suggère. »] (Peeperkorn et al., 2024, p. 8). Malgré tout, il s'agit d'un paramètre présent dans beaucoup de grands modèles de langage – tous ceux que nous étudierons dans ce mémoire, à vrai dire. Cela nous donne alors un réglage très simple à repérer, nous permettant de comparer le potentiel des divers grands modèles de langage que nous considérerons, vis à vis de l'aide qu'ils peuvent apporter pour l'analyse *a priori* de situations didactiques diverses et variées.

En tout cas, on a bien que les grands modèles de langage sont principalement statistiques plutôt que syntaxiques, comme vous le disiez ; il s'agit donc d'un axe de limitation majeure qu'il faudra explorer en détail.

Un enjeu majeur, également, de la partie méthodologique sera de déterminer comment faire en sorte que les grands modèles de langage que nous utiliserons puissent identifier des inférences élèves, d'une façon exploitable par les critères d'évaluation d'inférences usitées que nous aurons établis. Les LLM sauront-elles identifier ces inférences ? Il faudra donc jongler entre ce que les LLM arrivent à digérer d'une part, et le cadre méthodologique établi pour déterminer les inférences qui seront utilisées, ces deux disciplines devant alors communiquer l'une avec l'autre pour apporter un socle méthodologique aussi informé que possible.

2.3. Problématisation et conclusion de la partie théorique

Pour résumer, il semble que les grands modèles de langage, de par leur nature heuristique, risquent d'avoir des problèmes à détecter des inférences fines. Il s'agira donc de déterminer avec quelle précision les grands modèles de langage peuvent produire des analyses préalables de diverses situations didactiques, sur le plan purement inférentiel, dans un contexte mathématique.

La notion de « précision » de telles analyses préalables sera elle-même à préciser dans la partie méthodologie, mais dans la plupart des cas, nous emploierons la notion de « précision » comme la combinaison de deux critères :

- Justesse : Ordonner les analyses produites selon l'écart entre l'attendu et l'observé ; plus l'écart est important, plus la justesse de l'analyse préalable est médiocre, tandis qu'une analyse plus juste sera synonyme d'un moindre tel écart.
- Finesse : Plus il y a d'inférences proposées, plus l'analyse est fine. *A contrario*, une analyse préalable qui ne prend en compte que les raisonnements les plus basiques, et n'allant aucunement dans le détails des inférences immédiates possibles, sera une analyse plus grossière.

De par la nature syntaxique des inférences, il me semble important d'accorder du crédit à la finesse des analyses produites.

La problématique dégagée alors est la suivante : *« Peut-on rendre plus précise (juste et fine) les analyses préalables de procédures inférentielles d'élèves, identifiables et formulables en notation de Fitch, s'appuyant sur l'utilisation de grands modèles de langage, en faisant varier divers hyperparamètres comme celui de la température ? »*

Bibliographie et webographie

Armengaud, F. (2006, novembre 18). *Inférence*. Encyclopædia Universalis. https://www.universalis.fr/encyclopedie/inference/

Deloustal-Jorrand, V. (2001,). *L'implication. Quelques aspects dans les manuels et points de vue d'élèves-professeurs*. (n°55, p. 36, 65). Petit x. https://irem.univ-grenoble-alpes.fr/medias/fichier/55x3_1561102993655-pdf

Eduscol. (2016, mars 26). *Raisonner*. Ministère de l'Éducation nationale et de la Jeunesse. https://eduscol.education.fr/document/17224/download

Eduscol. (2019a, août 15). *Raisonnement et démonstration (Seconde générale et technologique)*. Ministère de l'Éducation nationale et de la Jeunesse. https://eduscol.education.fr/document/24580/download

Eduscol. (2019c, septembre 13). *Représentation des entiers naturels*. Ministère de l'Éducation nationale et de la Jeunesse. https://eduscol.education.fr/document/30025/download

Eduscol. (2019b, novembre 25). *Raisonnement et démonstration (Première générale)*. Ministère de l'Éducation nationale et de la Jeunesse. https://eduscol.education.fr/document/24583/download

Gao, A. (2017, septembre 26). *Semantic Entailment and Natural Deduction*. Université de Waterloo. https://cs.uwaterloo.ca/~a23gao/cs245_f17/slides/lecture6and7_to_post.pdf

Grenier, D., Bacher, R., Barbe, H., Bicaïs, H., Charlot, G., Decauwert, M., & Gezer, T. (2019, novembre 12). *Logique et Situations de Recherche pour la Classe*. (p. 11) Institut de recherche sur l'enseignement des mathématiques de Grenoble. https://irem.univ-grenoble-alpes.fr/recherche-action/raisonnement-logique-situations-de-recherche-pour-la-classe/logique-et-situations-de-recherche-pour-la-classe-498677.kjsp

Hiemstra, D. (2009). Language Models. In L. Liu & M. T. Özsu (éds.), *Encyclopedia of Database Systems* (p. 1591-1594). Springer US. https://doi.org/10.1007/978-0-387-39940-9_923

Lenzen, W. (2022). Rewriting the History of Connexive Logic. *Journal of Philosophical Logic*, *51*(3), 525-553. https://doi.org/10.1007/s10992-021-09640-6

Magnus, P., Button, T., Trueman, R., & Zach, R. (2023). *forall x: Calgary. An Introduction to Formal Logic*. (p. 117-167). Open Logic Project. https://forallx.openlogicproject.org/forallxyyc.pdf

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024, février 20). *Large Language Models: A Survey*. (p. 1-2). arXiv. https://arxiv.org/abs/2402.06196

Ministère de l'Éducation nationale et de la Jeunesse. (2020, juillet 9). *Programme de spécialité de mathématiques de terminale générale*. Le Bulletin officiel de l'Éducation nationale. https://eduscol.education.fr/document/24568/download

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024, mai 1). *Is Temperature the Creativity Parameter of Large Language Models?*. (p. 1, 3, 8). arXiv. https://arxiv.org/abs/2405.00492