ChatGPT vs CTRL+F: Impact sur l'apprentissage des élèves de l'enseignement secondaire

Léane Jourdan[1] * Quentin Lemesle [2] *
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
(2) Univ Rennes, CNRS, IRISA, EXPRESSION, 22300 Lannion, France
leane.jourdan@univ-nantes.fr, quentin.lemesle@irisa.fr

RÉSUMÉ

L'usage des Grands Modèles de Langue par les élèves du secondaire suscite de vives interrogations, sans que leur impact réel sur l'apprentissage ait été rigoureusement évalué. Nous présentons une expérience comparative menée auprès de 125 collégiens français répartis en trois groupes disposant chacun d'un outil distinct : un assistant conversationnel (ChatGPT), une recherche par mot-clé dans un PDF (CTRL+F) et un manuel papier. Les élèves réalisent deux fois la même évaluation d'histoire à une semaine d'intervalle, afin de mesurer la performance immédiate et la mémorisation. Nos résultats montrent qu'aucun outil ne confère d'avantage significatif sur la performance immédiate. En seconde session, le groupe utilisant le manuel affiche le gain le plus marqué et les temps de complétion les plus homogènes, tandis que CTRL+F présente la progression la plus faible. Ces résultats interrogent la valeur ajoutée de ChatGPT en contexte scolaire et soulignent la viabilité d'alternatives frugales.

ABSTRACT

ChatGPT vs. CTRL+F : Impact on Secondary-School Students' Learning

The use of Large Language Models by secondary-school students raises strong concerns, yet their actual impact on learning remains poorly evaluated. We present a comparative experiment with 125 French middle-school students assigned to three conditions : a conversational assistant (ChatGPT), keyword search within a PDF (CTRL+F), and a paper textbook. Students complete the same history task twice, one week apart, to measure both immediate performance and retention. Our results show that none of the tools provides a significant advantage on immediate performance. In the second session, the Paper group shows the largest gain and the most homogeneous completion times, while CTRL+F exhibits the weakest progression. These results question the added value of ChatGPT in classroom settings and highlight the viability of frugal alternatives.

MOTS-CLÉS : grands modèles de langue, éducation, enseignement secondaire.

KEYWORDS: large language models, education, secondary school.

1 Introduction

L'émergence des Grands Modèles de Langue (GML) (Brown *et al.*, 2020) et leur accessibilité croissante au grand public ont conduit à une adoption rapide de ces outils dans de nombreux domaines, dont l'éducation. Parmi les usages observés, celui des élèves du secondaire, pour la réalisation de

*. Contribution égale

devoirs ou la révision, soulève des enjeux particuliers : il est fréquemment perçu avec méfiance, voire interdit, sans que son impact réel sur les processus d'apprentissage n'ait été rigoureusement évalué.

Dans cet article, nous évaluons l'impact de différents outils sur l'apprentissage d'élèves de cinquième et quatrième. Nous menons une expérience comparative dans laquelle les élèves réalisent une tâche de fouille documentaire en contexte d'examen d'histoire, en s'appuyant soit sur ChatGPT, soit sur une recherche par mots-clés dans un document numérique (CTRL+F), soit sur des ressources Papier. L'examen est répété à une semaine d'intervalle afin de mesurer la rétention des informations.

Nous considérons les Questions de Recherche suivantes : ***QR1*** Les assistants GML améliorent-ils les performances des apprenants ? ***QR2*** Sont-ils plus efficaces qu'une alternative simple et frugale ? ***QR3*** Quel outil favorise la rétention des connaissances sur le temps long ? ***QR4*** Comment les outils sont-ils perçus par les apprenants ?

Nos résultats montrent qu'aucun des trois outils ne confère d'avantage significatif sur la performance immédiate (note de l'élève au devoir). En revanche, on observe des différences marquées une semaine après : les groupes Papier et ChatGPT présentent un gain important entre les deux sessions, tandis que le groupe CTRL+F affiche une progression faible. Le groupe Papier se distingue également par des temps de complétion plus homogènes, une meilleure maîtrise du support en session 2 et montre une tendance à mieux retenir la correction.

Nos principales contributions sont les suivantes :

— Un protocole expérimental combinant mesure de performance immédiate et mesure de rétention, impliquant 125 élèves de collège répartis en trois groupes.
— Une évaluation empirique de ChatGPT face à des alternatives plus simples en contexte de fouille documentaire scolaire.
— Une mise en évidence de l'intérêt limité de ChatGPT en contexte scolaire, dont les performances ne justifient pas le coût d'utilisation et la perception contrastée qu'il suscite chez les élèves.

2 Travaux connexes

L'intégration des GML aux environnements scolaires suscite un intérêt croissant de la communauté scientifique (Denny *et al.*, 2024), soulevant des questions fondamentales quant à leur impact sur l'apprentissage : favorisent-ils l'engagement intellectuel des étudiants, ou le réduisent-ils ? (Abdelghani *et al.*, 2023). Ces interrogations ont mené à de nombreuses études empiriques récentes.

Un premier axe de recherche porte sur l'utilisation des GML comme médiateurs de l'interaction pédagogique. Savelka *et al.* (2023) montrent qu'un GML affiné peut classifier automatiquement le type de demande d'aide formulée par un étudiant lors d'un cours d'introduction à la programmation, avec des performances proches d'une décision humaine. Cette tâche est non triviale (Turpin *et al.*, 2023; Kunz & Kuhlmann, 2024) ; de plus, les étudiants débutants risquent d'éprouver des difficultés à exprimer précisément leur problème (Lodge *et al.*, 2018). Un tel système ouvre la voie à un guidage automatique qui préserve la réflexion de l'étudiant en lui évitant de recevoir directement la réponse.

Un second axe s'intéresse à l'effet des GML sur la compréhension et la rétention des connaissances. Kreijkes *et al.* (2026) ont mené une expérience auprès de 344 élèves âgés de 14 à 15 ans en Angleterre, répartis en deux groupes ; chacun devait étudier deux passages de texte puis répondre à des questions. Dans un groupe, les élèves utilisaient pour un passage un GML et pour l'autre la prise de notes sur

ordinateur. Dans l'autre groupe, ils utilisaient pour un passage un GML seul et pour l'autre un GML combiné à la prise de notes. L'ordre des deux modalités était aléatoire au sein de chaque groupe. Aucun effet statistiquement significatif sur la compréhension ou la rétention n'est montré, mais les élèves ayant utilisé uniquement la prise de notes tendent à préférer cette modalité.

Dans la continuité de la théorie de la charge cognitive (Gerlich, 2025), qui postule que déléguer des tâches à des outils externes réduit l'effort de traitement nécessaire à la mémorisation à long terme, Barcaui (2025) a étudié la rétention d'information selon la modalité d'apprentissage. Leur expérience implique 120 participants âgés de 18 à 24 ans au Brésil, répartis en deux groupes : l'un autorisé à utiliser des GML ou des moteurs de recherche à base d'intelligence artificielle, l'autre soumis à une interdiction stricte de ces outils. Leur tâche consiste à effectuer des recherches sur un sujet lié à l'intelligence artificielle et à en présenter les résultats oralement à leurs pairs. Une interrogation surprise menée 45 jours plus tard révèle que, bien que la différence ne soit pas statistiquement significative, le groupe ayant utilisé les GML présente une tendance à une rétention de l'information moindre.

Kosmyna *et al.* (2025) ont quant à eux directement mesuré l'activité cognitive de 54 participants universitaires répartis en trois groupes selon l'outil utilisé : ChatGPT, un moteur de recherche classique, ou aucun outil. La tâche consistait à rédiger une dissertation sur quatre sessions. Les résultats montrent que les participants ayant utilisé ChatGPT étaient significativement moins capables de citer le contenu de leur essai ou les consignes de l'examen, et présentaient une activité cérébrale réduite durant la tâche, suggérant une moindre implication cognitive.

Ces travaux partagent plusieurs limites communes. D'une part, le support numérique est omniprésent, tant pour la consultation des ressources que pour l'évaluation, ce qui ne permet pas d'isoler l'effet de l'outil de celui du médium. D'autre part, les faiblesses propres aux GML, hallucinations et imprécisions factuelles (Ji *et al.*, 2023; Zhou *et al.*, 2023), sont rarement prises en compte comme variable expérimentale. Enfin, ces études se concentrent quasi exclusivement sur des tâches de lecture ou de rédaction, laissant peu de place à d'autres formes d'évaluation comme l'analyse d'images ou la fouille documentaire. Bien que peu d'impact statistiquement significatif n'ait été relevé, on peut observer des tendances communes à travers ces différents travaux : les résultats moyens des groupes utilisant un GML sont systématiquement plus faibles que ceux des groupes contrôles.

Notre travail se distingue sur plusieurs points. Nous ciblons des élèves de l'enseignement secondaire français, population jusqu'ici peu représentée dans la littérature. L'évaluation se déroule sur support papier, ce qui empêche tout copier-coller et oblige les participants à lire et reformuler activement les réponses du GML. Surtout, notre protocole inclut la comparaison à un outil simple et frugal, la recherche par mot-clé (CTRL+F), absente des études existantes à notre connaissance.

3 Méthode

3.1 Design expérimental

L'expérience prend la forme d'un devoir d'histoire de 10 questions à réaliser en 40 minutes, pour lequel les élèves s'appuient sur un manuel scolaire qu'ils n'ont jamais utilisé. Le devoir est réalisé lors de deux sessions séparées d'une semaine, afin de mesurer à la fois la performance immédiate et la rétention des connaissances. Le protocole est représenté de façon schématique en Figure 1. Chaque séance suit l'organisation suivante :

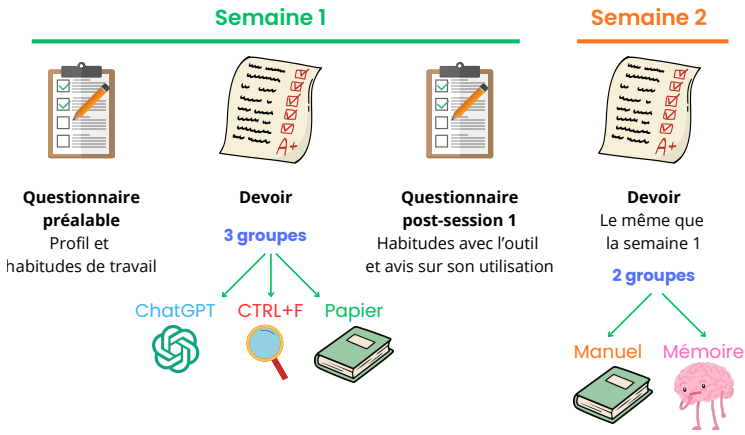


FIGURE 1 – Déroulé du protocole expérimental

— **Semaine 1**
  — **Questionnaire préalable** : Profil et habitudes de travail
  — **Devoir** : 40 min
  — **Questionnaire post-session 1** : Habitudes d'utilisation de l'outil et avis sur son usage
  — **Correction** : Faite au tableau devant toute la classe avec indication d'où se trouvait l'information à chercher dans le manuel.
— **Semaine 2**
  — **Devoir** : 40 min, identique à celui de la semaine 1, sans annonce préalable.

Les participants à l'expérience sont répartis en trois groupes aléatoirement. Chacun des groupes dispose d'un outil différent pour réaliser son devoir :

— **groupe ChatGPT** : accès au PDF du manuel et à la version gratuite d'un assistant conversationnel basé sur un GML (ChatGPT), permettant environ cinq échanges en mode RAG (*retrieval-augmented generation*), le PDF était ouvert sur leur ordinateurs ;
— **groupe CTRL+F** : accès au PDF du manuel avec consigne d'utiliser la recherche par mots-clés avec la commande CTRL+F ;
— **groupe Papier** : groupe contrôle disposant d'une version imprimée du manuel. On émet l'hypothèse que ce groupe aura plus de difficultés à trouver les informations dans un document long et aura donc des résultats inférieurs, voire manquera de temps pour finir le devoir.

Pour tous les groupes, le devoir est réalisé sur support papier afin d'empêcher tout copier-coller et de contraindre les participants à reformuler activement les informations obtenues.

La seconde session vise à mesurer la rétention des connaissances. Indépendamment de l'outil utilisé en première session, chaque groupe est subdivisé en deux sous-groupes selon les modalités suivantes :

— **Manuel** : accès à la version imprimée du manuel, afin d'observer l'évolution de la capacité de recherche documentaire ;
— **Mémoire** : aucun support à disposition, afin d'évaluer ce qui a été retenu.

Un questionnaire préalable recueille le profil de chaque participant (niveau scolaire, moyennes, habitudes de révision). Des questionnaires post-session mesurent le ressenti des élèves sur l'outil assigné. Les différents questionnaires sont fournis en annexe B.

Les élèves ne sont pas informés de cette organisation à l'avance. Durant les deux sessions, ils ignorent l'objectif de l'étude : il est seulement indiqué qu'ils participent à une expérience scientifique. Avant le devoir, les élèves des groupes ChatGPT et CTRL+F reçoivent une brève formation à leur outil respectif, sous la forme d'une démonstration au tableau, assortie d'une vérification individuelle de leur capacité à ouvrir la barre de recherche (CTRL+F) ou à importer le document dans ChatGPT.

À l'issue de l'expérience, les réponses sur papier sont reportées dans un tableur afin de constituer le jeu de données analysé en Section 4. Les données nominatives permettant d'identifier les élèves (nom, prénom, classe) ne sont pas saisies, et les documents papier sont détruits à l'issue de la saisie.

3.2 Participants et cadre expérimental

L'expérience se déroule en janvier 2026 dans un collège public de France métropolitaine. Cinq classes y participent : quatre classes de cinquième et une classe de quatrième, soit 125 élèves au total. Les participants sont répartis en trois groupes de taille équivalente : 41 élèves dans le groupe ChatGPT, 42 dans le groupe CTRL+F et 42 dans le groupe Papier.
Durant toutes les séances, une enseignante est présente pour encadrer la classe aux côtés des auteurs de l'article. L'expérience a été validée et autorisée par la directrice de l'établissement ; celle-ci a prévenu les collégiens que deux chercheurs allaient intervenir pendant leurs cours.

3.3 Matériel

L'établissement met à notre disposition des salles informatiques équipées d'un ordinateur par élève, avec accès à Internet pour ChatGPT et au logiciel Adobe Acrobat Reader pour la lecture du PDF.

Le document support utilisé est le manuel d'Histoire Géographie et Enseignement Moral et Civique (EMC) de quatrième des éditions Hatier. Nous nous concentrons sur un seul chapitre, « La Révolution française et l'Empire », pour deux raisons : d'une part, l'histoire offre une plus grande diversité de types de documents (textes, images, cartes) ; d'autre part, aucune des classes participant à l'expérience n'avait étudié ce chapitre au moment de l'expérience. Compte tenu de la durée de l'évaluation (40 minutes), nous limitons le corpus à 18 pages. Le chapitre nous a été fourni au format PDF par l'éditeur et est également consultable en ligne. [1] Ce format permet d'effectuer des recherches par mots-clés (CTRL+F) y compris dans les contenus des cartes et des schémas. Pour le groupe Papier, le chapitre est imprimé en couleur et relié.

Pour le groupe ChatGPT, un compte OpenAI distinct est préalablement créé par élève sous l'abonnement gratuit et est pré-connecté. Dans cette configuration, les élèves disposent d'environ cinq échanges en mode RAG et peuvent également s'appuyer sur les connaissances internes du modèle.

3.4 Construction de l'évaluation

Le devoir est élaboré à partir de cinq exemples d'évaluations portant sur le chapitre étudié, fournis par les enseignants d'histoire-géographie du collège. Il comprend 10 questions réparties en trois types : **questions à choix multiples (QCM)**, **questions ouvertes** et **analyse de documents**. Les deux

1. https://www.editions-hatier.fr/livre/histoire-geographie-emc-4e-ed-2022-livre-eleve-9782401085695 Nous utilisons les pages 68 à 85.

premières catégories comportent chacune quatre questions. Chacune relève d'un des quatre niveaux de difficulté suivants, définis en fonction des propriétés de chaque outil :

— **Facile pour les deux** : la réponse est présente de façon explicite dans le texte, le contexte est formulé de manière identique à celui de la question ;
— **Difficile pour CTRL+F** : la réponse est présente dans le texte, mais la question en paraphrase le contexte (par exemple, « au nom de dieu et du souverain » plutôt que « au nom du roi et de la religion »), ce qui ne pose pas de difficulté à ChatGPT ;
— **Difficile pour ChatGPT** : le contexte n'est pas paraphrasé, mais la réponse n'est pas formulée explicitement dans le texte et nécessite une étape de raisonnement ;
— **Difficile pour les deux** : le contexte est paraphrasé et la réponse requiert une étape de raisonnement ou de lecture d'image.

Le devoir a été soumis à la relecture d'une enseignante du collège, puis ajusté selon ses retours pour adapter le vocabulaire et les attentes au niveau des élèves. Le devoir corrigé et annoté avec l'indication du type de chaque question est disponible en Annexe A.

Pour obtenir la note finale nous considérons chaque type de questions comme ayant un poids équivalent. Ainsi, chacune des questions du QCM et des questions ouvertes vaut 1 point et chacune des questions d'analyse vaut 2 points, car il y a deux réponses attendues pour ces questions. Ce qui donne un score sur 12 points, que nous ramenons sur 30 pour faciliter l'interprétation.

4 Résultats

Dans cette section, nous décrivons et analysons les résultats obtenus par les élèves au devoir et leurs réponses aux différents questionnaires. Nous étudions les résultats selon le type d'outils utilisé ainsi que selon le type de question puis nous menons une analyse du ressenti des élèves.

Afin d'écarter les *copies blanches*, nous commençons par appliquer un filtre basé sur le nombre de réponses laissées vides. La distribution de ce nombre par groupe est représentée en Figure 2. Nous excluons ainsi les copies comportant plus de 10 questions sans réponse au total sur les deux devoirs. Les élèves qui ont été absents à une des deux sessions sont également exclus. Après ce filtrage, l'échantillon retenu comprend 39 élèves dans le groupe ChatGPT, 38 dans le groupe CTRL+F et 39 dans le groupe Papier, soit 116 élèves au total. La répartition des groupes reste équilibrée.

4.1 Profil de l'échantillon

Nous commençons par caractériser nos participants grâce à leurs réponses aux questionnaires préalables et post-première session (questionnaires en Annexe B). Sur les 116 élèves retenus, 113 sont locuteurs natifs du français et les 3 autres sont bilingues.

La Figure 3 présente la distribution des méthodes de révision habituellement utilisées par les élèves. La relecture des notes de cours constitue la méthode la plus répandue (72%), suivie par les discussions avec les parents et l'utilisation de fiches de révision, tandis que ChatGPT est peu utilisé dans ce contexte (18%). Les temps de révision moyens par groupe sont reportés en Annexe C. La Figure 4 présente la distribution des moyennes générales et en histoire-géographie par groupe. Les groupes suivent une tendance similaire et sont centrés autour de moyennes de 15. Le groupe CTRL+F est celui qui présente le plus d'élèves en difficulté et le groupe Papier est le plus uniforme.

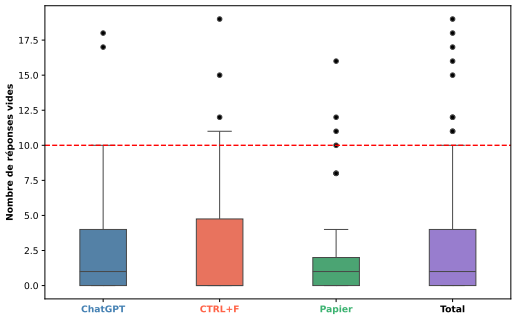


FIGURE 2 – Distribution du nombre de réponses laissées vides par groupe et pour l'ensemble de l'échantillon. Les points au-delà de la ligne rouge correspondent aux copies exclues de l'analyse.

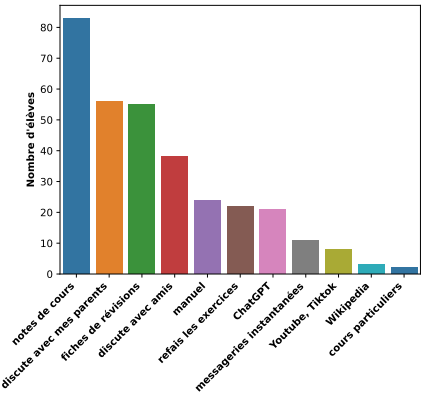


FIGURE 3 – Distribution des méthodes de révision déclarées par les élèves.

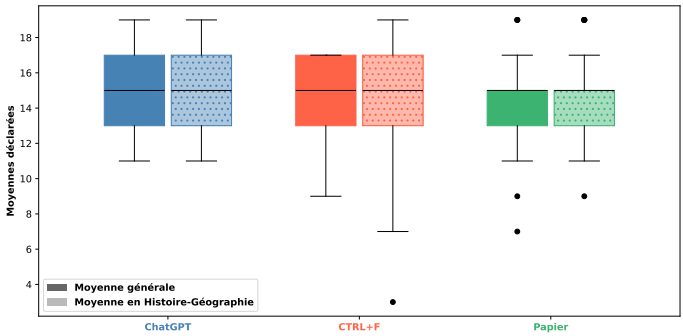


FIGURE 4 – Distribution des moyennes scolaires par groupe.

Enfin, après le premier devoir, seuls 15,8% des élèves du groupe CTRL+F déclaraient connaître le raccourci avant l'expérience, alors que 61,5% du groupe ChatGPT l'avaient déjà utilisé au moins une fois, avec 25,6% du groupe l'utilisant au moins une fois par semaine.

4.2 Résultats aux devoirs

Nous analysons les scores moyens par groupe et leur évolution entre les sessions, puis nous détaillons la réussite par question et par niveau de difficulté. Nous examinons ensuite les temps de complétion, puis l'évolution fine des réponses correctes entre les deux sessions.

Résultats globaux. Les scores moyens obtenus par groupe aux deux sessions sont reportés dans la Table 1. Lors de la première session, tous les groupes obtiennent des scores proches, avec un fort chevauchement des intervalles de confiance, malgré les différences de niveau entre élèves. On ne constate pas de différence significative entre les méthodes sur la performance immédiate.

Entre deux les sessions, on observe une évolution contrastée selon le groupe. Le groupe ChatGPT pré-

Groupe	Devoir	Moyenne	IC 95%	Δ	p	
ChatGPT	D1	16,86	[15,64 ; 18,07]			
	D2	19,46	[17,57 ; 21,34]	+2, 60	0,0081	**
	D1 Mémoire	*17,72*	*[16,13 ; 19,31]*			
	D2 Mémoire	*17,13*	*[14,22 ; 20,04]*	−0, 59	*0,7157*	*ns*
	D1 Manuel	*16,19*	*[14,36 ; 18,02]*			
	D2 Manuel	*21,25*	*[18,86 ; 23,64]*	+5, 06	*≪0,0001*	***
CTRL+F	D1	16,78	[15,04 ; 18,51]			
	D2	16,97	[14,02 ; 19,93]	+0, 20	0,8894	ns
	D1 Mémoire	*17,92*	*[14,95 ; 20,88]*			
	D2 Mémoire	*16,83*	*[13,12 ; 20,55]*	−1, 08	*0,3541*	*ns*
	D1 Manuel	*15,90*	*[13,09 ; 18,71]*			
	D2 Manuel	*21,81*	*[19,11 ; 24,50]*	+5, 90	*0,0002*	***
Papier	D1	16,67	[15,26 ; 18,07]			
	D2	19,68	[17,28 ; 22,08]	+3, 01	0,0151	*
	D1 Mémoire	*15,59*	*[12,95 ; 18,23]*			
	D2 Mémoire	*17,28*	*[14,20 ; 20,36]*	+1, 69	*0,1588*	*ns*
	D1 Manuel	*17,25*	*[15,61 ; 18,89]*			
	D2 Manuel	*23,69*	*[21,78 ; 25,59]*	+6, 44	*≪0,0001*	***

TABLE 1 – Comparaison des scores moyens (sur 30) entre Devoir 1 (D1) et Devoir 2 (D2) par groupe, et décomposition selon le sous-groupe Mémoire / Manuel au Devoir 2. T-test apparié, Intervalle de Confiance (IC) à 95 %. *** $p < 0{,}001$, ** $p < 0{,}01$, * $p < 0{,}05$, ns : non significatif.

sente une augmentation très significative (*p<0,001*), le groupe Papier une augmentation significative (*p<0,05*), tandis que le groupe CTRL+F n'atteint pas le seuil de progression significative (*p>0,05*).

Lors de la seconde session, les résultats diffèrent selon la condition d'accès au support. Pour le groupe Mémoire, toutes les configurations présentent une moyenne légèrement supérieure à celle du devoir 1, avec un fort chevauchement des intervalles de confiance. Bien que l'augmentation de la note ne soit pas significative, il faut noter que ce groupe ne disposait d'aucun support et témoigne donc d'une attention portée lors de la correction et d'une bonne capacité de rétention de l'information pour toutes les modalités. Le groupe Manuel affiche en revanche une amélioration très significative, particulièrement chez les élèves ayant déjà utilisé le manuel papier au Devoir 1. On observe d'ailleurs une hiérarchie nette entre les trois modalités d'origine : Papier > CTRL+F > ChatGPT. Cela reflète vraisemblablement le degré de familiarité antérieur avec le support : les élèves du groupe Papier l'ont manipulé au Devoir 1, ceux du groupe CTRL+F également via le PDF, tandis que ceux du groupe ChatGPT ont eu moins tendance à l'utiliser bien qu'il était ouvert sur leur ordinateur. On observe toutefois une tendance d'amélioration plus forte pour le sous-groupe Papier-Mémoire (*p≈0,16*). À l'inverse, pour CTRL+F, on observe une tendance de dégradation des résultats pour les élèves utilisant leur mémoire. Enfin, pour le sous-groupe ChatGPT-Mémoire, nous n'observons aucune tendance.

Nous nous intéressons également aux différences de notes entre les groupes : qu'il s'agisse de la note en session 1, en session 2 ou du gain entre les deux, elles ne révèlent pas de différences statistiquement significatives (Tableau 2). On note toutefois que le groupe CTRL+F présente systématiquement la moyenne et le gain les plus faibles en session 2. En particulier, la valeur $p \approx 0{,}2$ observée entre les

Mesure	ANOVA	Groupe 1	Groupe 2	Δ	IC 95%	p-adj	
D1	$F = 0,018$ $p = 0,982$ ns	CTRL+F	ChatGPT	$+0,083$	$[-2,349\,;\,2,515]$	0,996	ns
		CTRL+F	Papier	$-0,110$	$[-2,542\,;\,2,323]$	0,994	ns
		ChatGPT	Papier	$-0,192$	$[-2,609\,;\,2,224]$	0,981	ns
D2	$F = 1,536$ $p = 0,220$ ns	CTRL+F	ChatGPT	$+2,481$	$[-1,580\,;\,6,543]$	0,319	ns
		CTRL+F	Papier	$+2,706$	$[-1,356\,;\,6,767]$	0,258	ns
		ChatGPT	Papier	$+0,224$	$[-3,811\,;\,4,259]$	0,990	ns
Gain	$F = 1,627$ $p = 0,201$ ns	CTRL+F	ChatGPT	$+2,399$	$[-1,595\,;\,6,392]$	0,331	ns
		CTRL+F	Papier	$+2,816$	$[-1,178\,;\,6,809]$	0,220	ns
		ChatGPT	Papier	$+0,417$	$[-3,551\,;\,4,384]$	0,966	ns

TABLE 2 – Comparaison inter-groupes (**ChatGPT**, **CTRL+F**, **Papier**) pour les scores au Devoir 1 (D1) et au Devoir 2 (D2) et le gain (D2−D1) avec son Intervalle de Confiance (IC) à 95%. ANOVA à un facteur suivie d'un test post-hoc de Tukey HSD (*Family-Wise Error Rate*=0,05). Δ : différence de moyennes. ns : non significatif ($p \geq 0,05$).

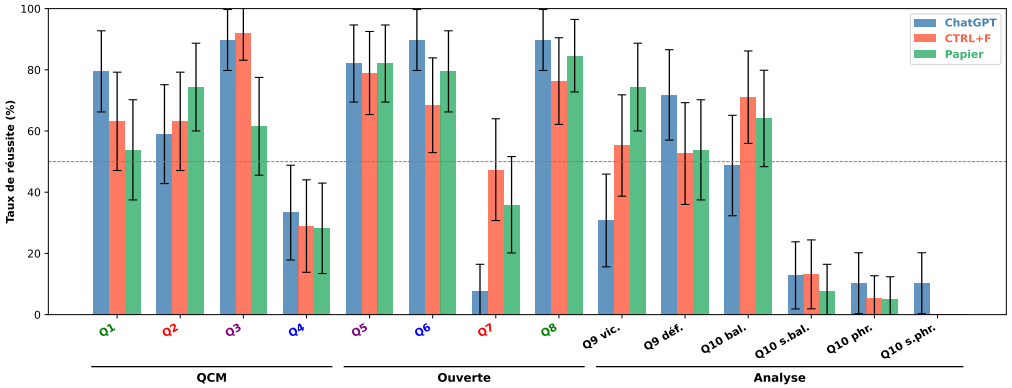
groupes CTRL+F et Papier indique une tendance qui ne peut être entièrement écartée compte tenu de la taille de l'échantillon. Nous n'observons ni écart significatif, ni tendance entre les groupes Papier et ChatGPT. Cela semble indiquer que le système n'a que très peu d'impact sur la réussite et la rétention d'information des élèves.

Pour CTRL+F, nous avons observé que les élèves ont des difficultés à utiliser l'outil. Certains ont essayé de l'utiliser comme un GML en posant une question dans la barre de recherche. De plus, les collégiens font des fautes d'orthographe et ont tendance à écrire phonétiquement. Cet outil fonctionnant par recherche exacte, aucun résultat n'est retourné si la recherche est mal écrite.

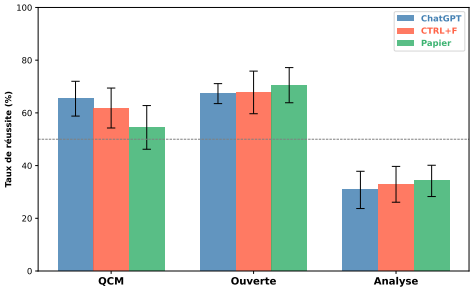
Réussite par question. La Figure 5 détaille les taux de réussite par question, par type et par niveau de difficulté au premier devoir. Pour les trois types de questions, les différences entre groupes sont faibles, à l'exception des QCM pour lesquels les groupes ayant travaillé sur ordinateur obtiennent de meilleurs résultats. Concernant les niveaux de difficulté, le groupe ChatGPT réussit significativement moins bien les questions conçues pour être difficiles pour les deux outils numériques (contexte paraphrasé et étape de raisonnement), tandis que le groupe CTRL+F affiche sur ces questions des résultats comparables à ceux du groupe Papier. Par ailleurs, le groupe ChatGPT tend à obtenir de meilleurs résultats sur les questions faciles pour les deux outils numériques et sur celles difficiles pour CTRL+F (le contexte est paraphrasé mais la réponse est explicitement présente dans le document).

Temps de complétion du devoir. Sur le temps de complétion (Figure 6), le groupe Papier se distingue par une variance plus faible dès la session 1, et par la vitesse d'exécution la plus élevée en session 2, avec CTRL+F. Ce résultat contraste avec notre hypothèse initiale, selon laquelle le groupe contrôle n'aurait pas le temps de parcourir l'intégralité du document dans le délai imparti. Il semble au contraire que la familiarité avec cette modalité réduise le temps d'adaptation et de mise au travail.

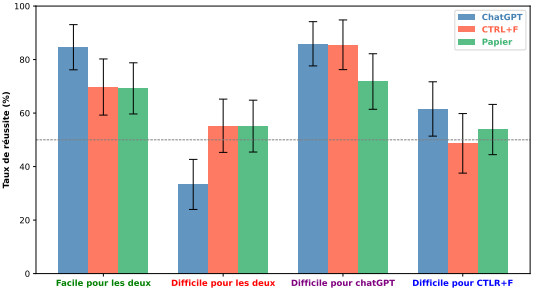
Évolution détaillée entre les devoirs. Afin de visualiser l'évolution des réponses correctes entre les deux sessions, et notamment la rétention des informations apportées lors de la correction au tableau,



(a) Taux de réussite par question



(b) Taux de réussite par type de question



(c) Taux de réussite par difficulté

FIGURE 5 – Taux de réussite aux questions par groupe au devoir 1.

nous utilisons un diagramme de Sankey (Figure 7), dont une version détaillée question par question est disponible en Annexe F. On observe que le groupe CTRL+F présente le plus de variations et également le plus grand taux d'oubli, à un niveau comparable à celui du groupe ChatGPT. Le groupe Papier, en revanche, montre une tendance plus marquée à la correction entre les deux sessions, ce que nous interprétons comme un indicateur d'une attention plus soutenue portée par ces élèves lors de la restitution corrigée.

4.3 Ressenti des élèves

Le ressenti des élèves face à la tâche est présenté en Table 3. Dans l'ensemble, CTRL+F et le manuel papier sont jugés utiles pour la réalisation du devoir. ChatGPT et CTRL+F sont perçus comme faciles à utiliser, mais l'intention de réutilisation reste modérée. Le manuel papier, bien qu'également jugé utile, est considéré comme moins facile d'utilisation et suscite une faible intention de réutilisation.

Concernant l'appréciation de l'expérience (*« Avez-vous aimé travailler avec [outil] ? »*), le groupe Papier affiche une distribution proche d'une loi normale, centrée en position neutre. ChatGPT suscite des réactions plus polarisées, avec une proportion notable de jugements très positifs et très négatifs. CTRL+F se distingue par une appréciation globalement favorable, avec peu de réactions extrêmes.

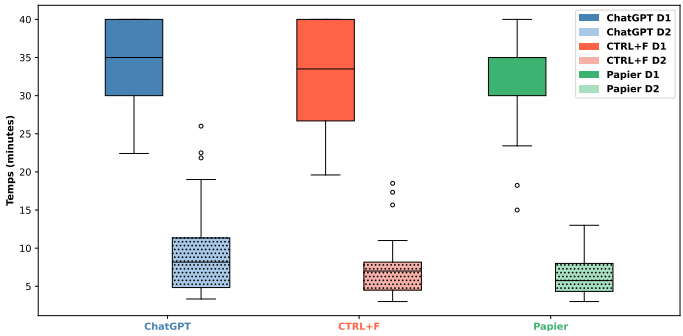


FIGURE 6 – Distribution des temps de réponse au devoir par groupe et par session.

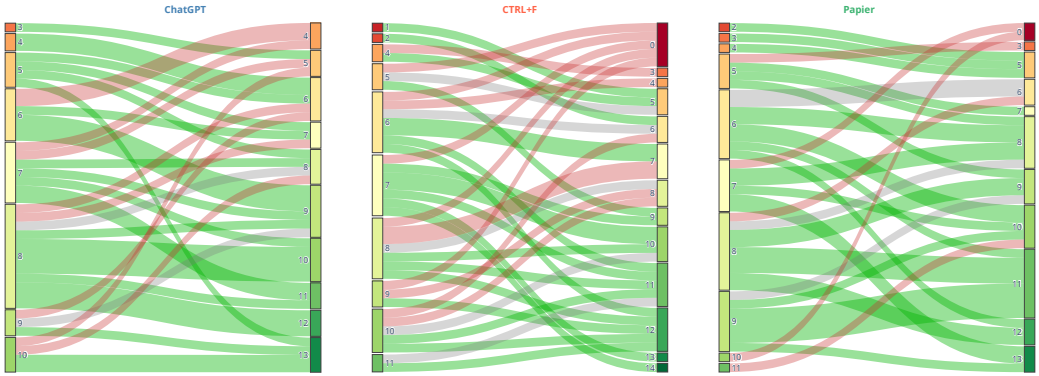


FIGURE 7 – Diagramme de Sankey représentant l'évolution du nombre de bonnes réponses par élève entre le devoir 1 et le devoir 2, par groupe. En rouge les élèves dont le nombre de bonnes réponses a baissé, en gris ceux dont le nombre est resté stable et en vert ceux dont le nombre a augmenté.

5 Discussion et Conclusion

Bien que les résultats ne soient pas significatifs, les p-values de 0,16 (Papier) et 0,71 (ChatGPT) sur l'amélioration des notes au Devoir 2 pour le groupe Mémoire laissent entrevoir un signal en faveur du manuel. Au regard de l'état de l'art, qui rapporte des effets similaires dans des contextes, tâches et populations différentes, cette convergence renforce l'hypothèse. S'il est prématuré de conclure que ChatGPT a un impact négatif sur l'apprentissage, le faisceau d'indices ne peut être ignoré.

Nous avons observé plusieurs phénomènes qualitatifs lors des corrections au tableau en fin de session 1. Les élèves du groupe ChatGPT paraissaient moins concernés par la correction et nous avons été surpris par leur réaction face à une erreur : souvent frustrés, ils s'exclamaient « il s'est trompé ! ». Ils semblent rejeter la faute sur l'outil et se dédouaner de leurs erreurs ; les détachant de leur travail et les déresponsabilisant. À l'inverse, les élèves du groupe CTRL+F ont plutôt apprécié l'outil malgré une frustration manifeste à l'usage : ils ne comprenaient pas pourquoi la commande ne se comportait pas comme dans les barres de recherche qu'ils utilisent au quotidien. Nous pensons que cette inadéquation explique en partie les performances plus faibles du système.

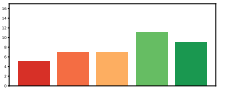
Groupe	Utile ?	Facile ?	Réutiliser ?	Appréciation Distribution	Appréciation Moy. / 5
ChatGPT	64,1	**89,7**	**41,0**		$3{,}31 \pm 1{,}36$
CTRL+F	81,6	84,2	36,8		$3{,}21 \pm 1{,}09$
Papier	**87,2**	66,7	17,9		$2{,}85 \pm 1{,}18$

TABLE 3 – Réponses au questionnaire de fin de première session. Pour les colonnes 2 à 4, les valeurs indiquées correspondent au pourcentage d'élèves ayant répondu *Oui*. Pour l'appréciation, les catégories sont, de gauche à droite : *Pas du tout*, *Pas vraiment*, *Neutre*, *Plutôt*, *Beaucoup*.

Ces observations restent empiriques ; il serait intéressant que de futurs travaux étudient ces phénomènes de ressenti direct des élèves pour mieux cerner leurs implications sur l'apprentissage. Dans l'attente d'autres études pour confirmer ces résultats, nous appelons à privilégier les systèmes les plus frugaux, dans la mesure où aucun outil ne produit d'amélioration significative de l'apprentissage.

Enfin, un compromis intéressant semble résider dans un CTRL+F amélioré, autorisant non plus seulement une recherche exacte mais une recherche approximative ou sémantique. Un tel outil resterait frugal tout en offrant une expérience plus naturelle aux collégiens.

6 Limitations

Taille et homogénéité de l'échantillon. Bien que notre échantillon de 116 élèves soit conséquent par rapport aux études similaires, la subdivision en groupes et sous-groupes ramène l'effectif par condition à une vingtaine d'élèves. Cela limite la représentativité statistique et explique, en partie, que plusieurs tendances observées n'atteignent pas le seuil de significativité. L'expérience s'est déroulée dans un unique établissement public de France métropolitaine, sur des élèves de quatrième et cinquième : la généralisation à d'autres niveaux, types d'établissements ou contextes culturels doit donc être faite avec prudence.

Familiarité préalable inégale avec les outils. Les trois modalités ne placent pas les élèves sur un pied d'égalité en termes de familiarité. Si le support papier est familier, ChatGPT est connu de 61,5% des élèves du groupe concerné alors que seuls 15,8% connaissent CTRL+F. La courte démonstration au tableau ne compense pas une pratique antérieure. L'effet observé pour CTRL+F mêle donc l'effet propre de l'outil et un effet d'apprentissage en cours de session.

Version de ChatGPT. Nous avons utilisé la version gratuite de ChatGPT, qui limite à environ cinq échanges en mode *RAG*. Les versions payantes ou les modèles plus récents pourraient produire des résultats différents, notamment sur les questions d'analyse de documents en fin de devoir.

Domaine et tâche. L'évaluation porte sur un unique chapitre d'histoire et sur une tâche de fouille documentaire. La transposition à d'autres disciplines (sciences, langues), à d'autres types de questions (production écrite, résolution de problèmes) ou à des temps de travail plus longs reste à étudier.

Mesure de la rétention. L'intervalle d'une semaine entre les deux sessions, contraint par l'organisation scolaire, permet de mesurer uniquement la rétention à moyen terme. La correction collective effectuée au tableau entre les deux sessions introduit une variable que nous ne contrôlons pas finement : la variation de l'attention d'un élève à l'autre. De plus, l'ordre des questions est resté identique entre les devoirs, pouvant induire un biais de mémorisation.

Remerciements

Nous tenons à remercier chaleureusement Madame Frédérique Jourdan pour son aide lors de l'élaboration du devoir et pour sa participation active à la mise en place de l'expérience : communication avec le personnel et les collégiens, logistique et présence lors de chaque session. Sans elle, nous n'aurions pas pu mener à bien cette expérience.

Nous tenons également à remercier Madame Nathalie Bourdin, directrice du collège, pour avoir accueilli cette expérience dans son établissement et encouragé notre intervention. Nous adressons nos remerciements à tous les collégiens qui ont accepté de participer à l'expérience et qui ont manifesté un vif intérêt à son sujet.

Nous remercions également les éditions Hatier de nous avoir partagé leur support PDF pour la réalisation de notre expérience.

Enfin, nous remercions Madame Nelly Barbot pour sa relecture approfondie et ses remarques pertinentes, ainsi que Monsieur Amine Boulahmel pour son aide à l'élaboration du protocole de recherche.

Références

Abdelghani R., Sauzéon H. & Oudeyer P.-Y. (2023). Generative ai in the classroom : Can students remain active learners ?

Barcaui A. (2025). Chatgpt as a cognitive crutch : Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, **12**, 102287. doi : https ://doi.org/10.1016/j.ssaho.2025.102287.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I. & Amodei D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

Denny P., Gulwani S., Heffernan N. T., Käser T., Moore S., Rafferty A. N. & Singla A. (2024). Generative ai for education (gaied) : Advances, opportunities, and challenges.

Gerlich M. (2025). Ai tools in society : Impacts on cognitive offloading and the future of critical thinking. *Societies*, **15**(1). doi : 10.3390/soc15010006.

Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y. J., Madotto A. & Fung P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38. doi : 10.1145/3571730.

KOSMYNA N., HAUPTMANN E., YUAN Y. T., SITU J., LIAO X.-H., BERESNITZKY A. V., BRAUNSTEIN I. & MAES P. (2025). Your brain on chatgpt : Accumulation of cognitive debt when using an ai assistant for essay writing task.

KREIJKES P., KEWENIG V., KUVALJA M., LEE M., HOFMAN J. M., VITELLO S., SELLEN A., RINTEL S., GOLDSTEIN D. G., ROTHSCHILD D., TANKELEVITCH L. & OATES T. (2026). Effects of llm use and note-taking on reading comprehension and memory : A randomised experiment in secondary schools. *Computers & Education*, **243**, 105514. DOI : https ://doi.org/10.1016/j.compedu.2025.105514.

KUNZ J. & KUHLMANN M. (2024). Properties and challenges of LLM-generated explanations. In S. L. BLODGETT, A. CERCAS CURRY, S. DEV, M. MADAIO, A. NENKOVA, D. YANG & Z. XIAO, Éds., *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, p. 13–27, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.hcinlp-1.2.

LODGE J. M., KENNEDY G., LOCKYER L., ARGUEL A. & PACHMAN M. (2018). Understanding difficulties and resulting confusion in learning : An integrative review. *Frontiers in Education*, **Volume 3 - 2018**. DOI : 10.3389/feduc.2018.00049.

SAVELKA J., DENNY P., LIFFITON M. & SHEESE B. (2023). Efficient classification of student help requests in programming courses using large language models.

TURPIN M., MICHAEL J., PEREZ E. & BOWMAN S. R. (2023). Language models don't always say what they think : Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

ZHOU W., ZHANG S., POON H. & CHEN M. (2023). Context-faithful prompting for large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 14544–14556, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.968.

A Devoir corrigé et annoté

Nom:
Prénom:
Classe :

Évaluation d'histoire
La Révolution française et l'Empire

Durée: 40 min

Exercice 1 : QCM

Facile pour les deux
Difficile pour les deux
Difficile ChatGPT
Difficile CTLR+F

1 - Avant la prise des Tuileries, que s'est-il passé en Juin 1791?
- □ Le peuple va chercher Louis XVI à Versailles
- □ Les fédérés s'emparent du palais
- ☑ **Le roi s'enfuit avec sa famille**
- □ La France déclare la guerre à l'Autriche

2 - Quelle est l'inscription sur la face de la monnaie à l'effigie de Bonaparte ?
- □ Napoléon Empereur
- ☑ **Premier Consul**
- □ Liberté égalité fraternité
- □ Monnaie de Paris

3 - D'après le texte de Lewis Goldsmith, que fait la police à ses citoyens ?
- □ La police les menaces
- □ La police les fouillent régulièrement
- □ La police protège les mendiants
- ☑ **La police les espionne dans leur vie quotidienne**

4 - Durant la Terreur, un citoyen ne va pas à la guerre si : (2 réponses)
- ☑ **il a 30 ans**
- □ il est noble
- □ il s'est battu pendant la révolution
- ☑ **il est marié**

FIGURE 8 – Devoir corrigé, page n°1

Exercice 2 : Questions ouvertes

5 - Pendant la conquête de l'Europe, combien de soldats l'armée française a-t-elle de plus que l'armée Prusse ?

4 000 de plus

6 - Lorsque la Première République s'installe, quel département s'y oppose au nom de Dieu et du souverain ?

La Vendée

7 - Est-ce que Olympe de Gouges s'est opposée au rétablissement du suffrage censitaire ?

Non car elle était décédée au moment du rétablissement du suffrage censitaire (Doc 3 : Discussions menées en 1795 alors qu'elle décède en 1793)

8 - Comment les parisiens surnomment-ils Robespierre ?

L'incorruptible

Exercice 3 : Analyse de document

Document 4 page 83 : Carte L'Empire napoléonien à son apogée (1811)

9 - Dans quelles villes Napoléon a-t-il eu sa première victoire et sa dernière défaite ?

Première victoire : Austerlitz - 1805
Dernière défaite : Waterloo - 1815

Document 1 page 74 : Les journées d'octobre 1789

10 - Quels sont les deux symboles que tient le personnage dans cette image ? → Selon vous, que peuvent-ils représenter ?

Le bonnet phrygien : La révolution
La balance : La justice

FIGURE 9 – Devoir corrigé, page n°2

B Questionnaires

Questionnaire préalable

1. Le français est-il votre langue maternelle ?

   ☐ Oui ☐ Non

2. Quel est votre niveau d'étude ?

   ☐ 5ème ☐ 4ème

3. Quelle est votre dernière moyenne générale ?

   ☐ 0-2 ☐ 2-4 ☐ 4-6 ☐ 6-8 ☐ 8-10
   ☐ 10-12 ☐ 12-14 ☐ 14-16 ☐ 16-18 ☐ 18-20

4. Quelle est votre dernière moyenne en histoire-géographie-EMC ?

   ☐ 0-2 ☐ 2-4 ☐ 4-6 ☐ 6-8 ☐ 8-10
   ☐ 10-12 ☐ 12-14 ☐ 14-16 ☐ 16-18 ☐ 18-20

5. Combien de temps passez-vous en général à réviser avant un devoir ?

   Entre et

6. Comment révisez-vous ? *(Plusieurs choix possibles)*

   ☐ Je discute avec mes amis
   ☐ Je relis mes notes prises en cours
   ☐ J'utilise des groupes de discussion sur les messageries instantanées et les réseaux sociaux (Whatsapp, Instagram,...)
   ☐ J'utilise le manuel
   ☐ J'utilise ChatGPT (ou une autre IA)
   ☐ Je discute avec mes parents
   ☐ Je refais les exercices faits en cours
   ☐ Je regarde du contenu vidéo (YouTube, TikTok, ...)
   ☐ Je fais des fiches de révision
   ☐ J'utilise Wikipedia
   ☐ Je prends des cours particuliers

Questionnaire post première session (Groupe ChatGPT)

1. Aviez-vous déjà utilisé ChatGPT avant aujourd'hui ?
   ☐ Oui ☐ Non
2. Si oui, à quelle fréquence l'utilisez-vous ?
   ☐ Tous les jours
   ☐ Une à plusieurs fois par semaine
   ☐ Une à plusieurs fois par mois
   ☐ À de rares occasions
   ☐ Je m'en suis servi(e) une fois et je ne l'ai pas réutilisé depuis.
3. De quelle façon l'utilisez-vous ? *(Plusieurs choix possibles)*
   ☐ Par saisie écrite ☐ Par saisie orale
4. Dans quel but l'utilisez-vous ? (Par exemple : Vérifier un fait, discuter, réviser, demander des recettes de cuisine, . . .)

5. Avez-vous trouvé l'utilisation de ChatGPT pertinente dans le cadre de cet examen ?
   ☐ Oui ☐ Non
6. Avez-vous trouvé ChatGPT facile d'utilisation ?
   ☐ Oui ☐ Non
7. Réutiliseriez-vous ChatGPT dans le cadre de vos révisions ?
   ☐ Oui ☐ Non
8. Avez-vous aimé travailler avec ChatGPT ?

Beaucoup	Plutôt	Neutre	Pas vraiment	Pas du tout
☐	☐	☐	☐	☐

Questionnaire post première session (Groupe CTRL+F)

1. Connaissiez-vous l'existence du raccourci CTRL+F avant aujourd'hui ?
   ☐ Oui ☐ Non
2. Si oui, pensiez-vous souvent à l'utiliser quand vous cherchez quelque chose dans un document ?
   ☐ Oui ☐ Non
3. Avez-vous trouvé l'utilisation de CTRL+F pertinente dans le cadre de cet examen ?
   ☐ Oui ☐ Non
4. Avez-vous trouvé CTRL+F facile d'utilisation ?
   ☐ Oui ☐ Non
5. Réutiliseriez-vous CTRL+F dans le cadre de vos révisions ?
   ☐ Oui ☐ Non
6. Avez-vous aimé travailler avec CTRL+F ?

Beaucoup	Plutôt	Neutre	Pas vraiment	Pas du tout
☐	☐	☐	☐	☐

Questionnaire post première session (Groupe Papier)

1. Avez-vous trouvé l'utilisation d'un document papier pertinente dans le cadre de cet examen ?
   ☐ Oui ☐ Non
2. Avez-vous trouvé le document facile d'utilisation ?
   ☐ Oui ☐ Non
3. Réutiliseriez-vous un manuel dans le cadre de vos révisions ?
   ☐ Oui ☐ Non
4. Avez-vous aimé travailler avec un document papier ?

Beaucoup	Plutôt	Neutre	Pas vraiment	Pas du tout
☐	☐	☐	☐	☐

C Temps de révisions

La Table 7 rapporte les temps de révision approximatifs habituels des élèves d'après les informations recueillies dans le questionnaire préalable.

Groupe	Temps moyen (min)	IC 95%	Nombre d'élèves
Tous	73,6	[34,7 ; 112,5]	115
ChatGPT	43,7	[33,3 ; 54,1]	38
CTRL+F	85,3	[11,0 ; 159,6]	38
Papier	91,3	[-0,9 ; 183,5]	39

TABLE 7 – Temps de révision moyen par groupe

D Réponses laissées vides

La figure 10 montre le nombre moyen de réponses vides par groupe lors du premier et du second devoir. On peut observer que le groupe ChatGPT a beaucoup moins répondu en moyenne lors du second devoir que lors du premier.

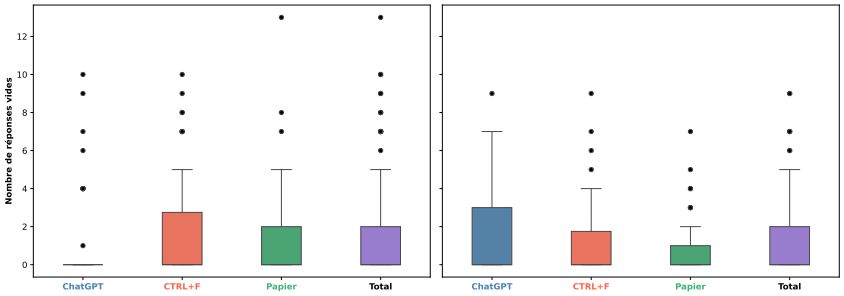


FIGURE 10 – Nombre de réponses vides par groupe sur le devoir n°1 et n°2.

E Résultat moyen au devoir par niveau

La Table 8 rapporte les notes moyennes obtenues aux devoirs par classes. On ne note pas de différences significatives entre les 5ème et les 4ème.

Classe	n	D1 (IC 95 %)	D2 (IC 95 %)
4ème	24	17,50 [15,82 ; 19,18]	19,74 [16,20 ; 23,28]
5ème	92	16,58 [15,63 ; 17,52]	18,45 [16,93 ; 19,97]

TABLE 8 – Moyennes et intervalles de confiance à 95 % des notes aux devoirs 1 et 2, par classe.

F Évolution détaillée entre les devoirs par question

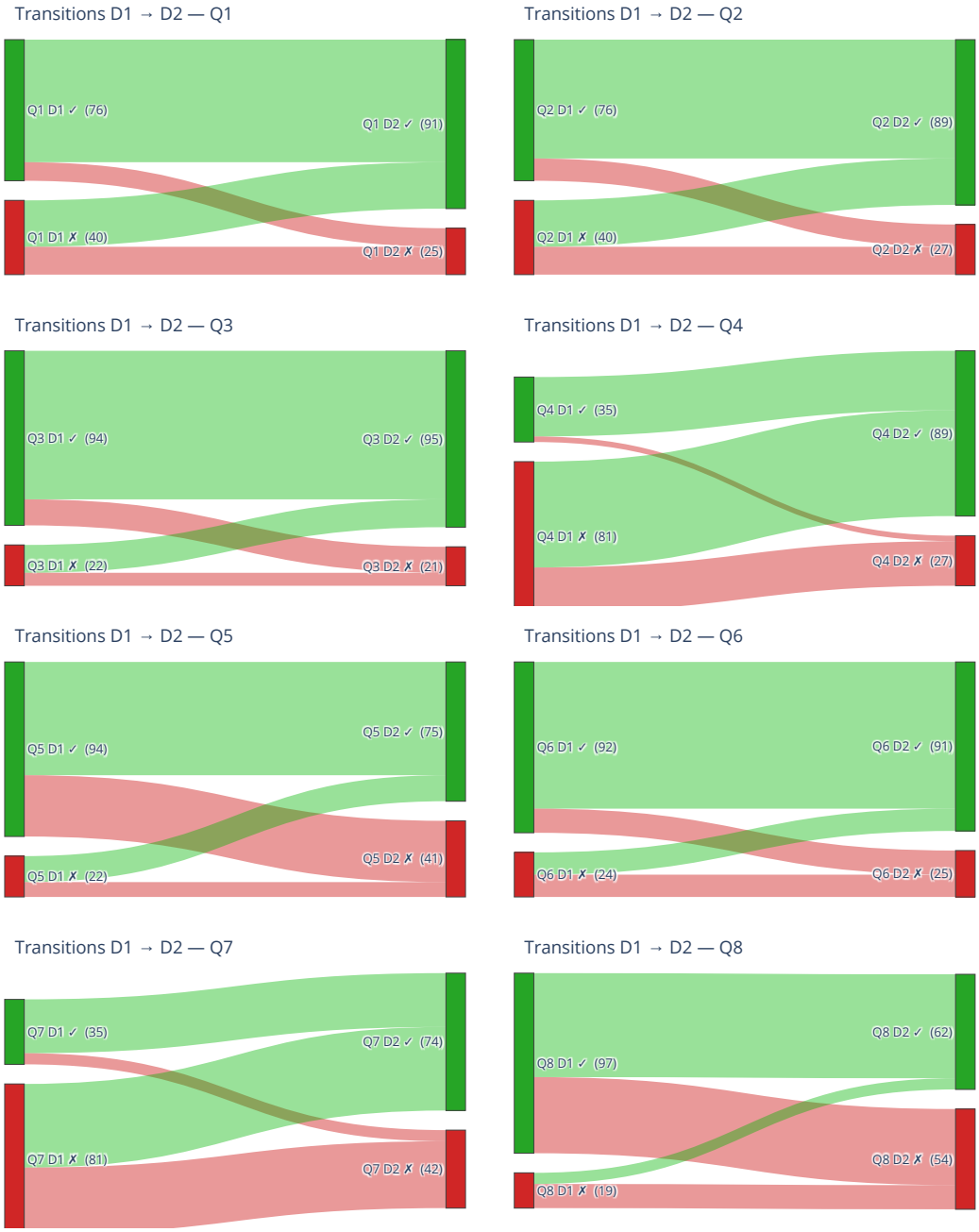


FIGURE 11 – Évolution des flux de réponses correctes et incorrectes entre les deux devoirs pour les questions 1 à 8.

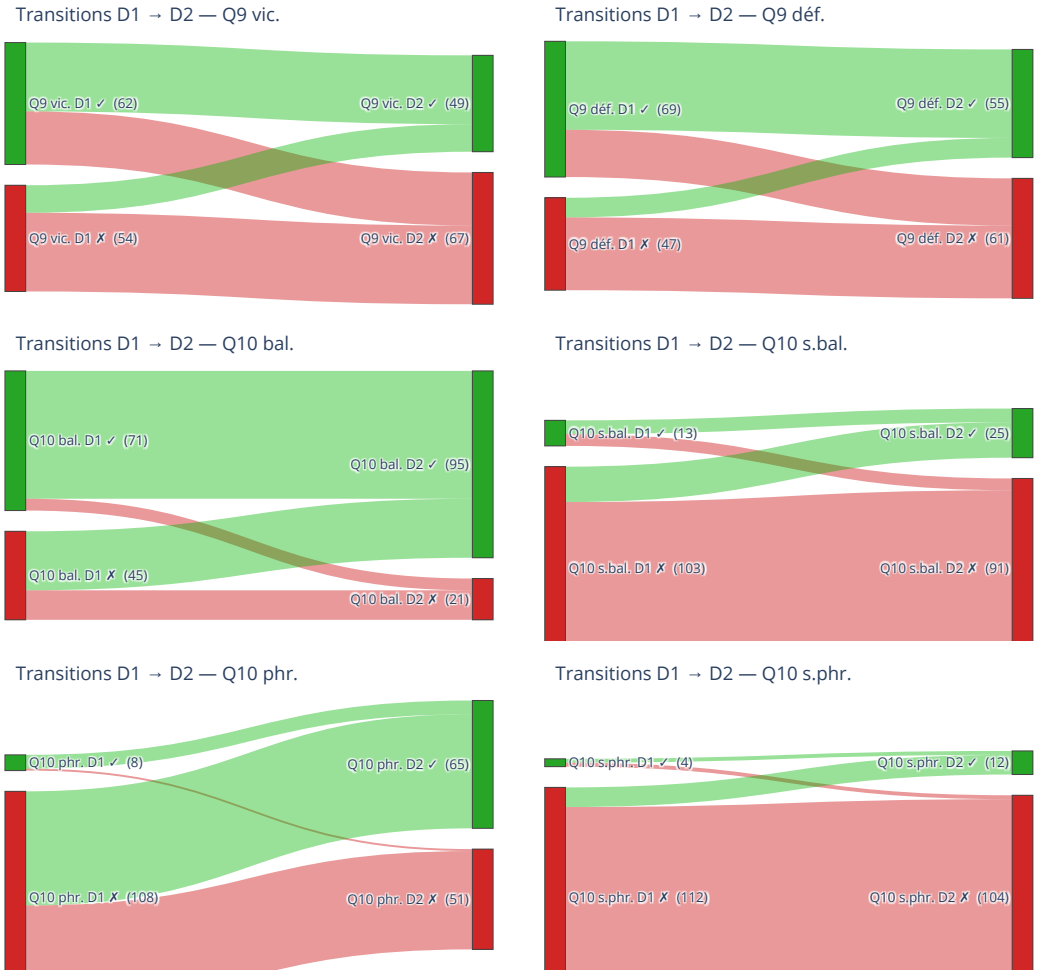


FIGURE 12 – Évolution des flux de réponses correctes et incorrectes entre les deux devoirs pour les questions 9 et 10.