

Louvain School of Management

Assessing AI's Ability to Recognize Human Emotions in Text

Author : **Thomas VANBEVER**
Supervisor : **Vincenzo VERADI**
Academic year : 2025 – 2026
Master [60] : Science de Gestion

Résumé

Titre : Assessing AI's Ability to Recognize Human Emotions in Text

Ce mémoire examine dans quelle mesure plusieurs grands modèles de langage sont capables de reconnaître les émotions de manière comparable à des répondants humains lorsqu'ils sont confrontés aux mêmes stimuli textuels. L'étude compare les réponses de Claude, Gemini, Piccolo et ChatGPT à une base humaine de référence construite à partir d'un questionnaire soumis à une quarantaine de participants. Le corpus analysé se compose de 28 titres d'articles, incluant des formulations explicites, implicites, métaphoriques ou ambiguës. La roue des émotions de Plutchik est utilisée comme cadre théorique commun, autour de huit émotions : la joie, la tristesse, la colère, la peur, l'anticipation, la surprise, le dégoût et la confiance.

La méthodologie repose sur deux niveaux d'analyse. Le premier consiste à comparer l'émotion dominante attribuée par les modèles à celle issue de la moyenne humaine. Le second prend en compte l'ensemble du profil émotionnel à l'aide d'indicateurs d'écart, notamment le MAE, le nMAE, le taux de couverture par intervalle de confiance et le taux de déviations statistiquement significatives. Des tests de Wilcoxon et de Friedman complètent l'analyse afin d'évaluer les différences entre modèles et selon les émotions.

Les résultats montrent que Piccolo est le modèle le plus proche de la référence humaine pour l'identification de l'émotion dominante. En revanche, lorsque l'ensemble du profil émotionnel est pris en compte, Claude apparaît légèrement plus proche de la base humaine. Gemini est le modèle qui s'en éloigne le plus. L'analyse par émotion révèle également que certaines émotions, comme l'anticipation, le dégoût et la confiance, différencient davantage les modèles. Ces résultats suggèrent que les grands modèles de langage peuvent approcher certaines logiques humaines de reconnaissance émotionnelle, mais que cette proximité reste partielle, variable et dépendante du type d'émotion étudié.

Mots-clés : grands modèles de langage ; reconnaissance des émotions ; informatique affective ; roue de Plutchik ; ChatGPT ; Claude ; Gemini ; Piccolo ; analyse émotionnelle ; comparaison humain-machine.

Remerciements

Avant tout, Je tiens à exprimer ma sincère gratitude à mon promoteur, Vincenzo veradi, pour ses conseils, ses encouragements et son soutien tout au long de ce projet. Son expertise m'a véritablement aidé à mener à bien ce mémoire de master.

Je souhaite également adresser mes remerciements à mes amis et à ma famille pour leurs encouragements et leur soutien dans le cadre de ce travail.

Enfin, je tiens à remercier toutes les personnes qui ont contribué à ce travail, que ce soit directement ou indirectement. Chaque conseil, chaque discussion et toute l'aide apportée ont été grandement appréciés et ont joué un rôle dans l'aboutissement de ce projet.

Merci à toutes et à tous pour vos contributions et votre soutien.

Note préliminaire

Deux outils basés sur l'intelligence artificielle ont été utilisés pour aider à la rédaction de ce travail : ChatGPT d'OpenAI¹ et Grammarly de Grammarly Inc². Ces deux outils ont été utilisés uniquement pour corriger les fautes d'orthographe et de grammaire et améliorer la structure ainsi que la lisibilité de ce mémoire. Ils n'ont pas été utilisés pour générer des réflexions, résultats ou explications. Ils ne sont donc pas responsables du contenu de ce travail.

1. <https://chatgpt.com/>

2. <https://app.grammarly.com/>

Table des matières

1	Introduction	2
2	Revue de la littérature	4
2.1	Grands modèles de langage	4
2.1.1	Diversité des modèles conversationnels contemporains	6
2.2	Informatique affective et reconnaissance des émotions	7
2.3	La roue des émotions de Plutchik	8
2.4	Capacités émotionnelles des grands modèles de langage	11
2.4.1	Capacités affectives fines et émergence de compétences émo- tionnelles	11
2.4.2	Des performances variables selon les tâches affectives	11
2.4.3	Comparaisons entre familles de modèles de langage	12
2.4.4	Comparaison avec la référence humaine et emotional awareness	13
2.4.5	L'influence du protocole : prompting, langue et adaptation du modèle	13
2.4.6	De la reconnaissance émotionnelle à l'empathie : une fron- tière à nuancer	14
2.4.7	Apports et limites des modèles généralistes	15
2.4.8	Synthèse	15
3	Méthodologie	16
3.1	Approche méthodologique	16
3.2	Dispositif d'étude	17
3.2.1	Constitution du corpus	17
3.2.2	Collecte des réponses humaines	17
3.2.3	Collecte des réponses des modèles de langage	18
3.3	Méthode d'analyse	19
3.3.1	Première analyse : identification de l'émotion dominante . . .	19
3.3.2	Deuxième analyse : comparaison du profil émotionnel complet	20
3.4	Limites méthodologiques	24

4	Résultats	28
4.1	Résultats de la première analyse	28
4.1.1	Résultats par modèle	29
4.1.2	Analyse du rang de l'émotion dominante humaine	30
4.2	Résultats de la deuxième analyse	30
4.2.1	Résultats globaux	31
4.3	Synthèse des principaux résultats	38
5	Conclusion	42
A	Titre du questionnaire	49
B	Tableau démographique	51
C	Résultats des grands modèle de langage	52
D	Graphiques radar par titre	57

Chapitre 1

Introduction

L’essor des grands modèles de langage (LLM) a profondément renouvelé les ambitions du traitement automatique du langage. Ils ont longtemps été mobilisés pour des tâches de génération, de traduction ou de classification mais ils ont rapidement démontré des capacités inattendues sur une grande variété de tâches. Dans un contexte où les systèmes conversationnels sont de plus en plus mobilisés, que ce soit dans des environnements éducatifs, informationnels, commerciaux ou d’assistance, leur capacité à interpréter, reconnaître et formuler des contenus émotionnels devient un enjeu décisif.

Cette problématique s’inscrit dans le champ de l’informatique affective qui a pour objectif de permettre aux systèmes artificiels de détecter, comprendre et modéliser les émotions humaines. Les travaux récents consacrés aux LLM montrent des résultats prometteurs en analyse de sentiment, en catégorisation émotionnelle ou encore en production de réponses émotionnellement adaptées mais des limites subsistent lorsque les signaux affectifs sont implicites, ambigus ou contextualisés.

Dans ce mémoire, nous proposons de comparer les capacités de reconnaissance émotionnelle de plusieurs grands modèles de langage à celles d’un groupe de répondants humains. Plus précisément, ce travail vise à examiner dans quelle mesure Claude, Gemini, Piccolo et ChatGPT sont capables d’identifier des émotions à partir d’un même corpus de titres d’articles, et dans quelle mesure leurs réponses se rapprochent des évaluations humaines. Afin de structurer cette comparaison, nous nous appuyons sur la roue des émotions de Plutchik comme cadre théorique commun.

La question de recherche centrale de ce mémoire peut dès lors être formulée de la manière suivante : dans quelle mesure les grands modèles de langage étudiés sont-ils capables de reconnaître les émotions de manière comparable à des répondants

humains lorsqu'ils sont confrontés aux mêmes stimuli textuels ?

Pour répondre à cette question, nous avons constitué un questionnaire de titres d'articles présentant à la fois des formes explicites et implicites d'expression émotionnelle. Ces titres ont été soumis à un groupe de participants humains ainsi qu'à plusieurs grands modèles de langage selon un protocole commun. L'objectif est d'analyser les convergences et les écarts entre les réponses produites par les modèles et celles issues de la base humaine de référence.

La suite de ce mémoire est structurée comme suit. Le chapitre 2 présente la revue de littérature et expose les principaux travaux relatifs aux grands modèles de langage, à l'informatique affective et à la reconnaissance des émotions. Le chapitre 3 présente la méthodologie adoptée, le dispositif d'étude, la constitution du corpus, les modalités de collecte des données et la méthode d'analyse retenue. Le chapitre 4 est consacré à la présentation et à l'analyse des résultats obtenus. Enfin, le chapitre 5 revient sur les principaux enseignements du mémoire, ses limites et les perspectives qu'il ouvre pour des travaux futurs.

Chapitre 2

Revue de la littérature

Ce chapitre présente la revue de littérature consacrée aux capacités émotionnelles des grands modèles de langage. L’objectif est d’examiner les principaux travaux portant sur la reconnaissance des émotions et sur l’informatique affective afin de situer ce mémoire dans le champ scientifique existant. Les recherches récentes montrent que des modèles généralistes sont de plus en plus évalués sur des tâches affectives variées, avec des résultats parfois encourageants mais souvent inégaux selon les tâches et les méthodes mobilisées.

Dans un premier temps, nous reviendrons sur les notions théoriques et conceptuelles de base relatives à la reconnaissance des émotions par l’intelligence artificielle. Nous présenterons ensuite le cadre théorique retenu dans ce mémoire, à savoir la roue des émotions de Plutchik. Nous examinerons ensuite les principaux travaux portant sur les capacités émotionnelles des grands modèles de langage. Enfin, les principales limites des travaux existants seront mises en évidence afin de faire ressortir les lacunes auxquelles ce mémoire entend répondre, c’est-à-dire la comparaison de plusieurs modèles conversationnels à partir d’un même cadre théorique et d’une même base humaine de référence.

2.1 Grands modèles de langage

Un grand modèle de langage est un modèle de langage possédant un grand nombre de paramètres. Il est composé de réseaux de neurones profonds entraînés sur de grandes quantités de texte non étiqueté à l’aide de l’apprentissage auto-supervisé. Contrairement aux approches plus anciennes, souvent développées pour une tâche unique et précise, les grands modèles de langage peuvent accomplir une grande variété de tâches simplement à partir de consignes formulées en langage naturel,

appelées prompts [8].

Cette évolution s’inscrit dans le prolongement de l’architecture Transformer, qui a profondément renouvelé le traitement automatique du langage grâce au mécanisme d’attention [28]. Les modèles de la famille GPT ont ensuite montré qu’un modèle pré-entraîné sur de très grands volumes de textes pouvait accomplir des tâches variées à partir de quelques exemples, voire sans aucun exemple explicite [6]. Pour ce mémoire, cette propriété est centrale : elle explique pourquoi des modèles généralistes peuvent être utilisés sur des tâches qui ne figuraient pas dans leur entraînement initial, comme la reconnaissance des émotions ou l’analyse affective de titres. Les modèles conversationnels récents reposent aussi sur des procédures d’alignement comme l’apprentissage par retour humain, dont l’objectif est de produire des réponses plus utiles, plus sûres et plus conformes aux attentes des utilisateurs [20].

Dans ce contexte, un prompt est une consigne textuelle en langage naturel destinée à guider le modèle vers une tâche spécifique [9]. Il en découle différentes configurations, telles que :

- le *zero-shot*, qui est une technique d’ingénierie de prompt où le modèle accomplit une tâche sans aucun exemple préalable, en s’appuyant uniquement sur ses connaissances issues du pré-entraînement ;
- le *few-shot*, qui est une technique d’ingénierie de prompt consistant à fournir au modèle quelques exemples d’entrées et de sorties souhaitées dans le prompt.

Ces configurations jouent un rôle important dans les travaux récents sur les grands modèles de langage. Elles permettent d’utiliser un même modèle sur des tâches très différentes sans avoir à entraîner un nouveau système spécialisé. Mais elles mettent aussi en lumière une dépendance forte à la formulation de la consigne : un même problème peut recevoir des réponses différentes selon le vocabulaire utilisé, l’ordre des informations, le nombre d’exemples fournis ou le format de sortie attendu. Cette sensibilité au prompt est particulièrement importante dans ce mémoire, puisque la manière dont la tâche est présentée au modèle peut influencer directement la reconnaissance d’une émotion.

On peut donc se demander dans quelle mesure les grands modèles de langage sont capables d’identifier des émotions significatives, alors même qu’ils n’ont pas été initialement conçus pour cela.

2.1.1 Diversité des modèles conversationnels contemporains

Les grands modèles de langage ne forment pas un ensemble homogène. Même s'ils reposent souvent sur des principes communs liés au pré-entraînement massif et à l'architecture Transformer, ils se distinguent par leur taille, leurs données d'entraînement, leurs modalités d'entrée, leurs mécanismes d'alignement et leurs objectifs de conception. Pour ce mémoire, cette hétérogénéité a une implication concrète : deux modèles peuvent recevoir le même prompt et produire des réponses différentes, non seulement en termes de performance brute, mais aussi dans leur style, leur manière de gérer l'ambiguïté ou leur niveau de prudence.

ChatGPT appartient à la famille des modèles GPT développés par OpenAI. Les travaux consacrés à ce modèle montrent qu'il peut accomplir plusieurs tâches affectives à partir de consignes en langage naturel, notamment l'analyse de sentiments, l'identification d'émotions et la génération de réponses émotionnellement adaptées [5, 1]. Cela dit, la littérature étant très largement centrée sur ChatGPT, les résultats qui en sont issus ne peuvent pas être étendus automatiquement à d'autres systèmes conversationnels.

Gemini, développé par Google DeepMind, s'inscrit dans une logique différente. Le rapport technique consacré à cette famille de modèles le présente comme un système multimodal capable de traiter simultanément du texte, des images, de l'audio et de la vidéo [12]. Cette dimension est intéressante pour l'informatique affective, puisque les émotions humaines ne s'expriment pas uniquement par le langage. Même si ce mémoire porte sur des titres textuels, Gemini rappelle que l'analyse des émotions peut s'inscrire dans un cadre plus large que le seul texte.

Claude, développé par Anthropic, constitue une autre famille de modèles conversationnels. La famille Claude 3 comprend notamment Opus, Sonnet et Haiku, qui représentent des compromis différents entre capacité, rapidité et coût [2]. Claude est également associé à une approche d'alignement appelée *Constitutional AI*, qui cherche à encadrer le comportement du modèle à partir de principes explicites et d'un retour généré par l'IA elle-même [3]. Dans une tâche de reconnaissance émotionnelle, cet aspect peut avoir une influence non négligeable : un modèle plus fortement aligné peut produire des classifications plus nuancées ou éviter certaines interprétations trop directes.

LLaMA, développé par Meta, occupe une position différente dans la littérature car il s'agit d'une famille de modèles plus ouverte, largement utilisée dans la recherche. Le rapport sur LLaMA 3 présente des modèles capables de traiter des tâches multilingues, de raisonnement, de code et d'usage d'outils [14]. L'ouverture de ce type de modèle facilite son adaptation et son évaluation dans des contextes expérimentaux. Même si LLaMA n'est pas directement évalué dans ce mémoire,

sa place dans la littérature montre que l'étude des capacités émotionnelles ne doit pas se restreindre aux modèles propriétaires les plus médiatisés.

Les modèles de la famille Mistral et Mixtral montrent par ailleurs que les performances des grands modèles de langage ne dépendent pas uniquement de leur taille. Mixtral repose sur une architecture dite *Mixture-of-Experts*, dans laquelle seuls certains experts sont activés pour traiter un token donné [15]. Cette diversité technique illustre les différentes stratégies possibles pour produire des modèles performants et renforce l'intérêt d'une comparaison entre plusieurs systèmes plutôt qu'une évaluation limitée à un seul modèle.

Enfin, Piccolo occupe une place particulière dans ce mémoire. Contrairement à ChatGPT, Claude ou Gemini, Piccolo ne constitue pas une famille de modèles de langage indépendante. Il s'agit d'une interface institutionnelle développée par l'UCLouvain afin de donner accès à des modèles d'intelligence artificielle générative dans un cadre universitaire [27]. L'outil est accessible aux membres de la communauté UCLouvain et repose notamment sur des modèles de la société Mistral, tels que Mistral Small, Mistral Medium ou Codestral. Dans ce mémoire, le terme Piccolo désigne donc l'environnement d'accès utilisé, tandis que le modèle sous-jacent appartient à la famille Mistral.

L'intérêt d'inclure Piccolo dans la comparaison est double. D'une part, il permet d'intégrer un outil effectivement proposé dans un contexte académique local. D'autre part, il permet d'observer dans quelle mesure un modèle accessible via une interface institutionnelle universitaire se rapproche ou s'éloigne des réponses produites par des systèmes propriétaires plus largement diffusés, tels que ChatGPT, Claude ou Gemini.

2.2 Informatique affective et reconnaissance des émotions

L'informatique affective est un domaine de recherche qui vise à mesurer, comprendre et parfois modéliser les émotions humaines [5]. Dans le champ du traitement du langage naturel, elle s'intéresse aux différentes manières dont les émotions peuvent être exprimées, interprétées et représentées à travers le langage. Elle ne se réduit donc pas à la détection d'une émotion ponctuelle mais cherche à analyser plus largement les phénomènes affectifs présents dans les interactions humaines et les productions de langage.

Ce domaine est généralement associé aux travaux de Picard, qui a contribué à le définir comme l'étude des systèmes capables de reconnaître, interpréter et parfois

simuler les affects humains [21]. Dans les travaux plus récents, le champ s’est élargi avec l’arrivée des grands modèles de langage. Zhang et al. distinguent deux grandes orientations : la compréhension affective, qui vise à reconnaître ou interpréter des émotions et la génération affective, qui vise à produire des réponses émotionnellement adaptées [30]. Cette distinction est utile pour situer notre travail : nous ne nous intéressons pas à la production de réponses empathiques mais bien à la capacité de plusieurs modèles à attribuer une émotion à des titres à partir d’un même cadre théorique.

La reconnaissance des émotions occupe une place centrale dans ce champ car elle consiste à identifier les émotions exprimées dans le langage. Toutefois, l’informatique affective appliquée au traitement du langage naturel couvre un ensemble bien plus large que la seule classification émotionnelle. Comme le montrent Amin et al., elle rassemble aussi l’analyse des sentiments, la détection du sarcasme, l’évaluation de la personnalité et la détection de la subjectivité, entre autres [1]. L’étude des émotions dans le langage porte donc à la fois sur des phénomènes directement exprimés et sur des signaux plus implicites parfois plus difficiles à saisir.

Les travaux dans ce domaine s’appuient souvent sur des jeux de données annotés par des humains. SemEval-2018 Task 1, par exemple, propose plusieurs tâches liées aux affects dans les tweets, dont la classification émotionnelle, l’intensité émotionnelle et la valence [18]. GoEmotions fournit quant à lui un jeu de données annoté avec des catégories émotionnelles fines, construites à partir de commentaires Reddit [10]. Ces ressources illustrent à quel point l’évaluation de la reconnaissance émotionnelle dépend de la qualité des annotations humaines. Elles justifient aussi le choix fait dans ce mémoire de comparer les réponses des modèles à une base humaine de référence.

Tout cela montre que la reconnaissance des émotions ne peut pas s’étudier sans un cadre théorique clair permettant de définir ce qu’est une émotion et comment elle peut être représentée. C’est pourquoi il est nécessaire, avant d’examiner les travaux sur les grands modèles de langage, de revenir sur le modèle théorique retenu dans ce mémoire.

2.3 La roue des émotions de Plutchik

Pour étudier la reconnaissance des émotions, il est nécessaire de s’appuyer sur un cadre théorique permettant d’identifier et de structurer les différentes émotions analysées. Bien qu’il existe plusieurs modèles dans la littérature, ce mémoire retient uniquement la roue des émotions de Plutchik comme cadre d’analyse.

Ce choix s’inscrit dans une conception catégorielle des émotions, où celles-ci sont

organisées en catégories fondamentales pouvant être rapprochées les unes des autres selon leurs relations d'opposition, de proximité ou d'intensité [22, 23]. Cette approche se distingue des modèles dimensionnels comme celui de Russell, qui représente les affects selon des axes continus tels que la valence et l'activation [24]. Plutchik a été retenu ici parce que son modèle permet de travailler directement avec des catégories mobilisables dans une tâche de classification ce qui correspond mieux aux contraintes pratiques de ce mémoire.

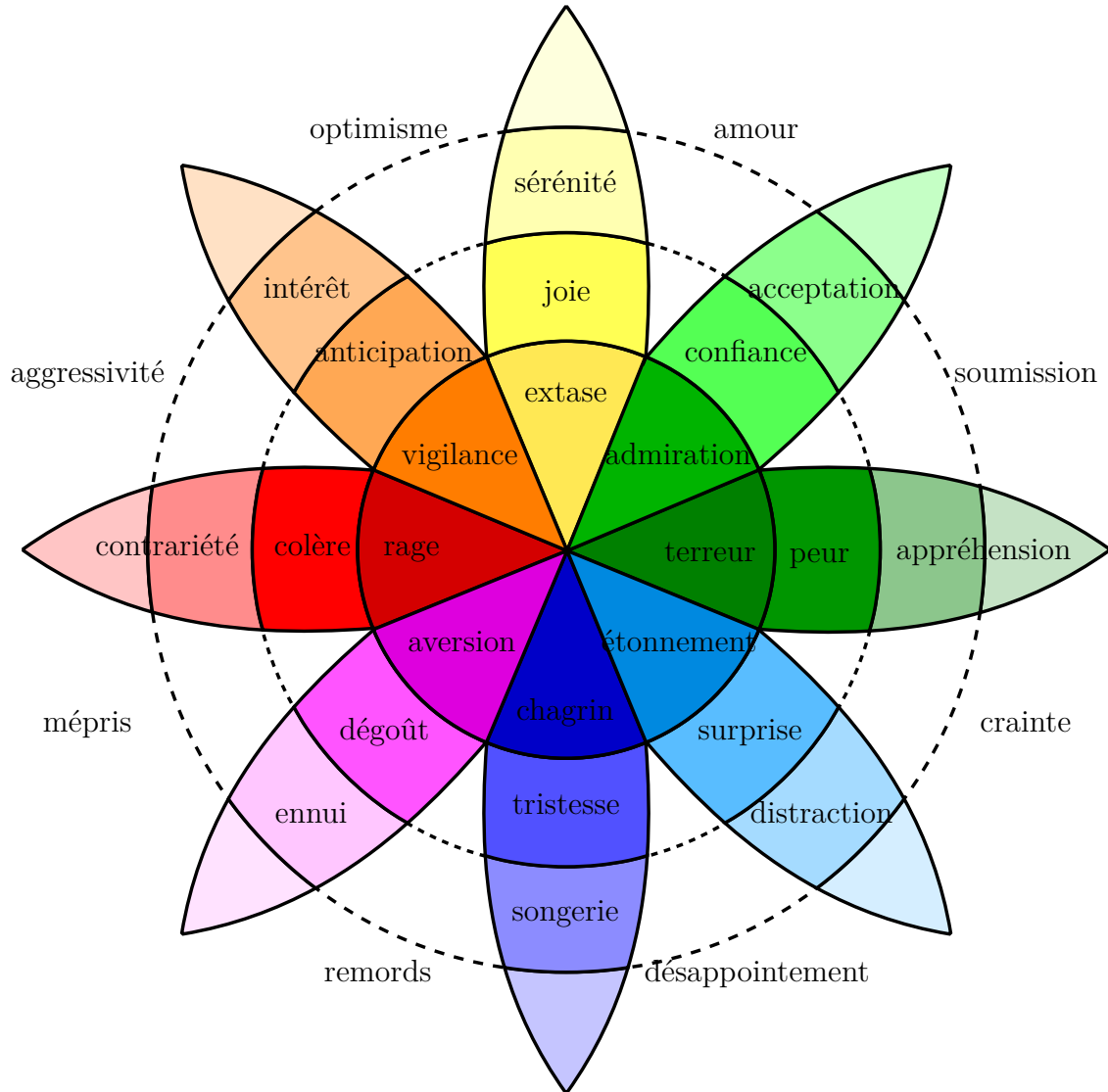


FIGURE 2.1 – La roue des émotions de Plutchik

Comme illustré dans la figure 2.1, le modèle repose sur huit émotions fondamentales : la joie, la confiance, la peur, la surprise, la tristesse, l'anticipation, la colère

et le dégoût. Ces émotions sont représentées sous forme de roue pour mettre en évidence leurs relations mutuelles. Certaines sont opposées tandis que d'autres sont proches et peuvent se combiner ou évoluer l'une vers l'autre selon les situations.

La roue permet également de représenter l'intensité : plus une émotion est proche du centre, plus elle est intense, et plus elle se situe vers l'extérieur, plus elle est modérée. Ce double ancrage, catégoriel et intensif, rend le modèle particulièrement utile pour l'étude des émotions dans le langage. Il ne permet pas seulement d'identifier une catégorie mais aussi de saisir les nuances et les relations entre émotions proches.

D'un point de vue méthodologique, travailler avec un référentiel commun comme la roue de Plutchik facilite la comparaison entre modèles. En imposant les mêmes catégories émotionnelles à ChatGPT, Claude, Gemini et Piccolo, on peut observer s'ils attribuent les mêmes émotions aux mêmes titres ou si des divergences apparaissent. La roue joue ainsi à la fois le rôle de cadre théorique et d'outil de standardisation. Comme le rappellent Nfaoui et Elfaik, il s'agit d'un modèle simple et facilement mobilisable dans les travaux de reconnaissance des émotions [19].

2.4 Capacités émotionnelles des grands modèles de langage

Les travaux récents sur les capacités émotionnelles des grands modèles de langage dressent un tableau nuancé. Ces systèmes peuvent accomplir un certain nombre de tâches affectives de façon pertinente même sans avoir été conçus spécifiquement pour cela. Les résultats restent cependant contrastés et dépendent fortement de la nature de la tâche, du cadre théorique mobilisé, du protocole de prompting, de la langue étudiée et du type de sortie attendu.

Selon Zhang et al., l'arrivée des grands modèles de langage a transformé les recherches en informatique affective en ouvrant de nouvelles possibilités pour la compréhension et la génération affectives notamment grâce à l'apprentissage en contexte, au raisonnement de sens commun et aux méthodes d'adaptation comme le fine-tuning ou LoRA [30]. Ces auteurs soulignent toutefois que les performances restent tributaires des tâches, des jeux de données et des méthodes d'évaluation ce qui rejoint directement la question posée dans ce mémoire.

2.4.1 Capacités affectives fines et émergence de compétences émotionnelles

Une première série de travaux suggère que certains grands modèles de langage traitent les informations affectives de manière relativement fine, même sans fine-tuning. Broekens et al. montrent que ChatGPT peut produire une analyse des sentiments significative selon les dimensions de valence, d'arousal et de dominance, mobiliser des représentations émotionnelles pertinentes et accomplir certaines tâches d'élicitation émotionnelle à partir du prompting seul [5]. Le modèle établit notamment des correspondances cohérentes entre situations, dimensions affectives et catégories émotionnelles, ce qui suggère l'émergence de capacités affectives à partir d'un entraînement généraliste.

Ces résultats sont importants car ils indiquent que le traitement émotionnel ne suppose pas nécessairement un système conçu à cet effet. Certaines compétences affectives semblent pouvoir émerger de modèles généralistes, à condition que les tâches soient correctement formulées par prompting [5]. Cela dit, cela ne signifie pas que les performances soient uniformes sur l'ensemble des phénomènes émotionnels.

2.4.2 Des performances variables selon les tâches affectives

Cette variabilité est bien documentée dans l'étude d'Amin et al., qui évaluent GPT-3.5 et GPT-4 sur treize tâches d'informatique affective [1]. Les performances sont

relativement bonnes sur les tâches liées aux sentiments, aux émotions explicites ou à la toxicité. En revanche, elles se dégradent quand les signaux émotionnels deviennent plus implicites ou difficiles à formaliser.

Les tâches où les modèles semblent le plus en difficulté sont notamment la mesure de l’engagement, l’évaluation de la personnalité, la détection du sarcasme et la subjectivité [1]. Cela montre qu’on ne peut pas parler d’une capacité émotionnelle unique et homogène : selon le type de signal à traiter, les performances varient considérablement. Les phénomènes explicites sont plus accessibles que les phénomènes implicites, ambigus ou fortement contextuels.

Cette distinction a une importance directe pour ce mémoire. Notre corpus contient à la fois des titres où l’émotion est clairement exprimée et des titres qui reposent sur la métaphore, l’ironie ou le sarcasme. On peut donc s’attendre à des disparités entre modèles sur ces deux types de titres.

2.4.3 Comparaisons entre familles de modèles de langage

Une partie plus récente de la littérature cherche à dépasser l’évaluation d’un seul modèle pour comparer plusieurs familles de grands modèles de langage. Cette évolution est importante : les résultats obtenus sur ChatGPT ne permettent pas de conclure que tous les modèles conversationnels reconnaissent les émotions de la même façon. Les différences d’architecture, d’entraînement, d’alignement et de politique de sécurité peuvent produire des écarts significatifs dans les réponses.

Lecourt et al. proposent une comparaison empirique entre GPT et d’autres grands modèles de langage pour la détection d’émotions dans le texte [16]. Leurs résultats montrent que les LLM possèdent de réelles capacités pour identifier certaines émotions, mais qu’ils ne dépassent pas nécessairement les modèles entraînés spécifiquement pour ce type de tâche. Ce constat invite à évaluer les modèles généralistes avec précaution et dans un protocole commun.

Buscemi et Proverbio comparent pour leur part ChatGPT, Gemini et LLaMA sur une tâche d’analyse des sentiments multilingue [7]. Ils utilisent des scénarios ambigus ou ironiques, les traduisent dans plusieurs langues et confrontent ensuite les réponses des modèles à des jugements humains. Les résultats varient selon les modèles, les langues et les formes d’ambiguïté. Ce type de comparaison est directement pertinent pour notre mémoire, puisque nos titres peuvent eux aussi comporter des formulations implicites ou métaphoriques.

Ces travaux confirment que ChatGPT, Claude, Gemini et Piccolo ne sont pas de simples variantes interchangeables d’un même outil. Gemini se distingue par son orientation multimodale [12]. Claude est associé à une approche d’alignement

particulière [2, 3]. LLaMA et Mixtral montrent qu’une architecture différente peut aussi produire de bons résultats dans de nombreux contextes [14, 15]. Comparer plusieurs modèles sur le même corpus et avec les mêmes catégories émotionnelles permet donc d’observer non seulement leurs performances relatives, mais aussi leurs différences d’interprétation.

L’enjeu n’est pas uniquement de déterminer quel modèle est le meilleur. Il s’agit aussi d’identifier les convergences et les divergences. Si plusieurs modèles attribuent la même émotion à un titre, cela peut signaler une certaine stabilité interprétative. À l’inverse, des divergences peuvent révéler une ambiguïté du titre, une sensibilité différente à certains mots ou simplement le style propre au modèle.

2.4.4 Comparaison avec la référence humaine et emotional awareness

Certains travaux ne comparent pas seulement les modèles entre eux mais les confrontent à des jugements humains. L’étude d’Elyoseph et al. mobilise la *Levels of Emotional Awareness Scale* (LEAS) pour comparer ChatGPT à des normes issues de la population générale [11]. Les résultats montrent que ChatGPT obtient des scores supérieurs à ceux de la population de référence sur plusieurs dimensions de l’*emotional awareness* et que ses réponses sont jugées très adaptées au contexte par deux psychologues indépendants [11].

Ces résultats doivent toutefois être interprétés avec prudence. Les auteurs reconnaissent eux-mêmes que l’étude porte sur un seul modèle, une seule langue et une période délimitée [11]. La question de savoir si une bonne performance sur une telle échelle équivaut à une véritable compréhension émotionnelle reste ouverte. Ce que cette étude apporte, c’est une confirmation que la comparaison avec une base humaine est une approche valide pour examiner les capacités émotionnelles des modèles de langage ce qui justifie le dispositif mis en place dans ce mémoire.

2.4.5 L’influence du protocole : prompting, langue et adaptation du modèle

Les performances observées dans la littérature dépendent aussi fortement des choix méthodologiques. L’article de Nfaoui et Elfaik, consacré à la reconnaissance des émotions en arabe, l’illustre bien [19]. Les auteurs comparent plusieurs approches : *in-context learning*, *emotional stimuli prompt* et fine-tuning et montrent que les résultats varient sensiblement selon la méthode. Dans leur étude, le fine-tuning de GPT-3.5 obtient les meilleurs résultats ce qui montre que les performances ne dépendent pas seulement de l’architecture du modèle, mais aussi de la manière

dont il est sollicité.

On retrouve ce constat dans les travaux de Belkhir et Sadat, qui montrent que l'ajout d'un classifieur émotionnel externe ou de consignes émotionnelles spécifiques peut modifier les réponses produites par ChatGPT dans des contextes conversationnels [4]. Certaines formes de prompt engineering renforcent l'expression de certaines émotions ou produisent des réponses jugées plus naturelles [4]. La capacité émotionnelle d'un modèle ne peut donc pas être évaluée indépendamment de la façon dont la tâche lui est soumise.

Les grands modèles de langage sont aussi de plus en plus mobilisés comme outils d'annotation. Gilardi et al. montrent que ChatGPT peut obtenir de très bons résultats dans certaines tâches d'annotation textuelle parfois supérieurs à ceux de travailleurs de foule [13]. Törnberg montre de même que GPT-4 peut produire des annotations compétitives dans certains contextes spécialisés [26]. Ces travaux montrent que les LLM peuvent formuler des jugements catégoriels, mais rappellent aussi la nécessité de confronter ces jugements à une référence humaine pour ne pas confondre cohérence apparente et validité réelle.

2.4.6 De la reconnaissance émotionnelle à l'empathie : une frontière à nuancer

Les capacités émotionnelles des grands modèles de langage ne doivent pas être confondues avec l'empathie au sens strict. Plusieurs travaux montrent qu'un modèle peut reconnaître ou reformuler des contenus émotionnellement pertinents sans pour autant manifester une empathie comparable à celle des humains. Les recherches sur l'empathie de ChatGPT soulignent que son évaluation dans les chatbots reste difficile à standardiser et qu'il n'existe pas encore de méthode unifiée pour la mesurer de façon fiable [25]. Certaines approches se concentrent sur des évaluations de type *feature-level*, d'autres adoptent une perspective plus globale fondée sur des questionnaires d'empathie [25].

Les résultats montrent que ChatGPT peut produire des réponses émotionnellement cohérentes mais qu'il obtient souvent des scores inférieurs à ceux d'humains sains sur plusieurs dimensions standardisées de l'empathie [25]. Cela rappelle que reconnaître l'émotion dominante dans un texte n'est pas équivalent à comprendre en profondeur l'expérience affective d'autrui. Ce mémoire ne porte pas directement sur l'empathie mais cette littérature aide à mieux situer ce que les modèles de langage font réellement quand ils « reconnaissent » des émotions.

2.4.7 Apports et limites des modèles généralistes

Les travaux récents confirment que les modèles généralistes obtiennent des résultats intéressants sur certaines tâches affectives mais que leurs performances restent inégales face à des tâches spécialisées. L'évaluation de GPT-4V sur des tâches de *Generalized Emotion Recognition* montre par exemple que ce système surpasse des baselines heuristiques tout en restant inférieur aux systèmes supervisés spécialisés [17]. Les auteurs notent en particulier des difficultés sur des tâches nécessitant un savoir spécialisé comme la reconnaissance des micro-expressions ou l'intégration de signaux issus de plusieurs modalités [17]. Même si cette étude ne porte pas sur un corpus purement textuel, le constat qu'elle formule rejoint ce qu'on observe par ailleurs sur ChatGPT : les modèles généralistes ne remplacent pas des systèmes spécialisés pour les phénomènes émotionnels complexes ou fins.

Un autre point de vigilance concerne les biais. Wake et al. montrent que les tâches de reconnaissance émotionnelle avec ChatGPT peuvent être exposées à des biais dans l'interprétation des émotions [29]. Les modèles conversationnels alignés sont entraînés pour produire des réponses sûres et acceptables ce qui peut influencer leurs interprétations [3]. Dans une tâche de classification émotionnelle, cette prudence peut avoir des effets ambivalents : elle peut réduire les réponses problématiques, mais rendre aussi le modèle plus hésitant face aux titres ambigus.

2.4.8 Synthèse

Au total, la littérature montre que les grands modèles de langage possèdent des capacités émotionnelles réelles mais partielles. Ils peuvent produire des évaluations affectives fines, reconnaître des émotions explicites, générer des réponses adaptées à certains contextes et, parfois, se rapprocher de performances humaines sur des tâches précises [5, 1, 11]. Ces capacités restent cependant fortement dépendantes des tâches, des données, de la langue, du protocole et du cadre théorique [1, 19, 4]. La littérature demeure par ailleurs très centrée sur ChatGPT, laissant encore ouverte la question d'une comparaison structurée entre plusieurs grands modèles de langage dans un même cadre d'analyse.

Chapitre 3

Méthodologie

Dans ce chapitre, nous présentons la méthodologie mise en œuvre pour comparer les capacités de reconnaissance émotionnelle de Claude, Gemini, Piccolo et ChatGPT à celles d’une base humaine de référence. Après avoir exposé l’approche générale et le dispositif d’étude, nous détaillons la méthode d’analyse retenue, organisée en deux niveaux de comparaison. Le chapitre se termine par une discussion des principales limites méthodologiques à prendre en compte dans l’interprétation des résultats.

3.1 Approche méthodologique

Cette recherche repose sur une démarche comparative. Elle vise à observer les convergences et les écarts entre les réponses produites par plusieurs grands modèles de langage et celles fournies par des répondants humains face à un même ensemble de données. L’étude ne cherche pas à établir une relation causale mais à comparer des réponses émotionnelles produites dans un contexte d’évaluation contrôlé. Cette démarche permet d’examiner la manière dont les émotions sont reconnues, nuancées et hiérarchisées selon les différents répondants. Elle s’organise autour de trois éléments principaux : la constitution d’un corpus de titres d’articles, la collecte parallèle de réponses humaines et de réponses issues des grands modèles de langage, puis la comparaison systématique de ces deux ensembles de réponses.

3.2 Dispositif d'étude

3.2.1 Constitution du corpus

Le questionnaire utilisé pour recueillir les réponses analysées dans ce mémoire a été élaboré à partir de titres d'articles. Ces titres ont été sélectionnés au sein d'un ensemble plus large d'articles trouvés en ligne. Les 28 titres retenus ont été choisis de façon à permettre une diversité de réponses et à éviter que l'identification des émotions ne soit trop simple, tant pour les participants humains que pour les modèles de langage. Lors de la constitution de ce corpus, des titres à la fois explicites et implicites ont été intégrés au questionnaire. Certains reposaient sur des formulations directes, tandis que d'autres faisaient intervenir des formes plus nuancées telles que l'ironie, le sarcasme ou la métaphore. Ce choix permet de couvrir un large éventail de situations émotionnelles et d'observer plus finement les différences d'interprétation entre répondants humains et grands modèles de langage. La liste complète des titres utilisés est disponible en annexe.

3.2.2 Collecte des réponses humaines

Une enquête a été réalisée afin de recueillir les réponses d'une quarantaine de participants et de constituer ainsi une base humaine de référence. Les réponses humaines ont été collectées au moyen d'un questionnaire en ligne, diffusé principalement sur les réseaux sociaux ainsi qu'auprès de l'entourage de l'auteur. Les participants ont donc été recrutés selon une méthode d'échantillonnage non probabiliste, fondée sur la participation volontaire. Le nombre de répondants a également été déterminé par les contraintes matérielles et budgétaires propres à ce mémoire, la collecte reposant sur une diffusion gratuite du questionnaire, sans dispositif de recrutement rémunéré ou statistiquement contrôlé.

Les répondants étaient âgés de 21 à 80 ans. Les informations sociodémographiques disponibles restent toutefois limitées, car seules les données relatives à l'âge global des participants et à leur localisation approximative ont pu être exploitées. Afin de ne pas surcharger le corps du texte, un tableau descriptif présentant la répartition géographique approximative des répondants est proposé en annexe, voir le tableau B.1. Ces informations permettent de contextualiser la base humaine de référence, sans pour autant prétendre à une représentativité statistique de l'ensemble de la population.

Les répondants ont été soumis au même questionnaire, fondé sur les huit émotions de base de la roue de Plutchik : la joie, la tristesse, la colère, la peur, l'anticipation, la surprise, le dégoût et la confiance. Pour chaque titre, les participants pouvaient nuancer leur réponse à l'aide d'un curseur allant de 0 à 100 pour chacune de ces

émotions, permettant ainsi d’exprimer à la fois la nature et l’intensité du ressenti.

3.2.3 Collecte des réponses des modèles de langage

La même base de titres a ensuite été soumise à chacun des quatre grands modèles de langage retenus, selon un protocole identique à celui des participants humains. Les titres ont été présentés en anglais et en français afin de limiter un éventuel biais lié à la langue d’entraînement des modèles et de rendre les conditions d’évaluation plus comparables à celles des répondants humains. Un prompt commun a été utilisé pour l’ensemble des modèles, de manière à garantir la comparabilité des conditions d’évaluation tout en laissant aux modèles une certaine liberté de réponse.

Bonjour, tu es un expert en étude des émotions en particulier concernant les émotions de la roue des émotions de Plutchik. Nous allons nous baser uniquement sur les émotions de base suivantes : Joie, Tristesse, Colère, Peur, Anticipation, Surprise, Dégoût et Confiance. Je vais te demander d’évaluer 28 titres d’articles. Pour chaque titre, je veux que tu te bases uniquement sur le titre que je te donne et que tu précises les émotions que tu ressens à partir de celui-ci. Chaque émotion est évaluée de 0 à 100 et plusieurs émotions peuvent être évaluées simultanément. Cela signifie qu’il est possible, par exemple, d’attribuer pour un titre : 42/100 pour la joie, 4/100 pour la tristesse, 16/100 pour la colère, 2/100 pour la peur, 0/100 pour l’anticipation, 32/100 pour la surprise, 17/100 pour le dégoût et 0/100 pour la confiance. Je veux que tu reportes toutes tes réponses dans un tableau structuré, facilement exportable vers Excel afin que je puisse récupérer facilement toutes les données.

Afin de préciser les conditions de collecte des réponses des modèles, les versions utilisées ont été relevées lorsque cette information était disponible. Cette précision est importante car les grands modèles de langage évoluent rapidement : une même requête peut produire des résultats différents selon la version du modèle, sa date d’utilisation, son interface d’accès ou ses paramètres internes. Le tableau suivant présente les modèles mobilisés dans ce mémoire ainsi que les informations disponibles concernant leur version ou leur mode d’accès.

TABLE 3.1 – Modèles de langage utilisés dans l’étude

Grand modèle de langage	Mode d’accès / version utilisée
GPT, OpenAI	Version utilisée : [GPT-5.5 Thinking]
Claude, Anthropic	Version utilisée : [Claude Sonnet 4.6]
Gemini, Google DeepMind	Version utilisée : [Gemini 3.5 Flash]
Piccolo, Mistral	Version utilisée : [Mistral Large 3]

Chaque modèle a été interrogé une seule fois pour chaque titre à l'aide du même prompt standardisé. Les réponses obtenues constituent donc une mesure ponctuelle du comportement des modèles au moment de la collecte.

3.3 Méthode d'analyse

Les données collectées sont analysées à deux niveaux de lecture complémentaires. La première analyse porte sur l'identification de l'émotion dominante et vise à vérifier si les modèles repèrent correctement l'émotion principale associée à chaque titre. La seconde analyse approfondit cette comparaison en examinant l'ensemble du profil émotionnel, à l'aide d'indicateurs d'écart et de tests d'inférence statistique. Ces deux niveaux sont présentés successivement dans les sous-sections qui suivent.

3.3.1 Première analyse : identification de l'émotion dominante

La première analyse constitue le niveau de lecture le plus direct de la comparaison. Pour chaque titre, l'émotion dont la moyenne humaine est la plus élevée est retenue comme émotion dominante de référence. Cette émotion dominante humaine est ensuite comparée à l'émotion ayant obtenu le score le plus élevé dans la réponse produite par chaque modèle pour ce même titre.

Deux indicateurs sont utilisés dans cette première analyse. Le premier est le taux de correspondance exacte. Il correspond à la proportion de titres pour lesquels l'émotion dominante du modèle est identique à l'émotion dominante issue de la référence humaine. Lorsque ces deux émotions coïncident, le modèle est considéré comme ayant correctement identifié l'émotion principale associée au titre.

Le second indicateur est le rang moyen de l'émotion dominante humaine dans la hiérarchie émotionnelle produite par le modèle. Lorsqu'un modèle ne place pas l'émotion humaine dominante en première position, l'analyse examine le rang occupé par cette émotion dans son classement. Une émotion dominante humaine classée en deuxième ou troisième position indique une proximité partielle avec la référence humaine, différente d'une absence totale de correspondance. Plus le rang moyen est proche de 1, plus le modèle place régulièrement l'émotion humaine dominante en haut de sa hiérarchie émotionnelle.

Cette première analyse permet donc d'évaluer la capacité des modèles à identifier l'émotion principale associée à chaque titre, tout en tenant compte des cas où cette émotion n'est pas placée en première position mais reste néanmoins proche

du sommet du classement.

3.3.2 Deuxième analyse : comparaison du profil émotionnel complet

La seconde analyse approfondit la comparaison en prenant en compte l'ensemble du profil émotionnel associé à chaque titre et non plus uniquement l'émotion dominante. Pour chaque titre, les scores attribués aux huit émotions par les modèles sont comparés aux scores moyens obtenus auprès des participants humains.

Les réponses humaines de référence sont construites à partir de la moyenne des scores attribués par les participants pour chaque combinaison titre $\times$ émotion. Chaque modèle est ensuite comparé à cette référence humaine afin d'évaluer dans quelle mesure son profil émotionnel se rapproche ou s'éloigne de celui observé chez les participants.

Cette analyse repose sur deux types d'outils complémentaires. Le premier ensemble regroupe des indicateurs descriptifs d'écart, tels que le MAE, le nMAE, le taux de couverture par intervalle de confiance à 95 % et le taux de déviations statistiquement significatives. Le second ensemble regroupe des tests d'inférence statistique permettant de comparer les modèles entre eux et d'examiner si leurs performances varient selon les émotions étudiées.

Indicateurs d'écart par rapport à la référence humaine

Le premier indicateur retenu est l'écart moyen absolu, ou *Mean Absolute Error* (MAE). Il mesure l'écart moyen entre les scores attribués par un modèle et les scores moyens issus de la référence humaine. Pour chaque titre, le MAE est calculé en prenant la moyenne des écarts absolus observés sur les huit émotions :

$$\text{MAE}_t = \frac{1}{n} \sum_{i=1}^n |M_{t,i} - \bar{H}_{t,i}|$$

Dans cette formule, $M_{t,i}$ correspond au score attribué par le modèle à l'émotion i pour le titre t , $\bar{H}_{t,i}$ correspond à la moyenne des scores humains pour cette même émotion et ce même titre, et $n = 8$ correspond au nombre d'émotions de base retenues. Le MAE global d'un modèle est ensuite obtenu en moyennant les MAE calculés sur l'ensemble des titres.

Toutefois, le MAE brut présente une limite importante : il traite de manière identique un écart survenant dans un contexte de fort consensus humain et un écart

survenant dans un contexte de forte dispersion des réponses. Or, si les participants humains divergent fortement sur une émotion donnée, un écart du modèle par rapport à la moyenne humaine est moins révélateur que si ce même écart survient dans un contexte de consensus.

Pour tenir compte de cette dispersion, un second indicateur est introduit : le MAE normalisé par l'écart-type humain, ou nMAE. Il consiste à diviser chaque écart absolu par l'écart-type des réponses humaines pour la cellule considérée. L'écart du modèle est ainsi exprimé en unités de variation interindividuelle humaine.

$$\text{nMAE}_t = \frac{1}{n} \sum_{i=1}^n \frac{|M_{t,i} - \bar{H}_{t,i}|}{\sigma_{t,i}}$$

Dans cette formule, $\sigma_{t,i}$ désigne l'écart-type des réponses humaines pour l'émotion i et le titre t . Un nMAE inférieur à 1 indique qu'en moyenne, l'écart du modèle reste inférieur à un écart-type humain. Un nMAE supérieur à 1 indique au contraire que l'écart moyen du modèle dépasse la dispersion observée entre les répondants humains.

Un troisième indicateur est le taux de couverture par intervalle de confiance à 95 %. Pour chaque cellule, c'est-à-dire pour chaque combinaison titre $\times$ émotion, un intervalle de confiance est construit autour de la moyenne humaine :

$$\text{IC}_{95\%} = \left[\bar{H}_{t,i} - 1,96 \cdot \frac{\sigma_{t,i}}{\sqrt{n_{t,i}}}, \bar{H}_{t,i} + 1,96 \cdot \frac{\sigma_{t,i}}{\sqrt{n_{t,i}}} \right]$$

Dans cette formule, $\bar{H}_{t,i}$ désigne la moyenne humaine pour le titre t et l'émotion i , $\sigma_{t,i}$ l'écart-type des réponses humaines, et $n_{t,i}$ le nombre de répondants humains pour cette cellule.

Le taux de couverture correspond à la proportion de cellules pour lesquelles le score du modèle se situe à l'intérieur de cet intervalle. Cet indicateur permet d'évaluer dans quelle mesure les réponses du modèle sont compatibles avec la moyenne humaine estimée. Il ne doit toutefois pas être interprété comme une comparaison directe avec un répondant humain individuel, car l'intervalle de confiance porte sur l'estimation de la moyenne humaine et non sur la dispersion complète des réponses individuelles.

Tests d'inférence statistique

Au-delà des indicateurs descriptifs, plusieurs tests d'inférence statistique sont mobilisés afin d'évaluer la significativité de certains écarts observés et de comparer les modèles entre eux. Ces tests ne poursuivent pas tous le même objectif. Le test t à un échantillon permet d'examiner, cellule par cellule, si le score d'un modèle est compatible avec la moyenne humaine. Le test de Wilcoxon signé apparié permet de comparer les modèles entre eux à partir de leurs MAE calculés titre par titre. Enfin, le test de Friedman permet d'examiner si les différences entre modèles varient selon les émotions étudiées.

Test t à un échantillon par cellule

Un test t à un échantillon est appliqué pour chaque cellule, c'est-à-dire pour chaque combinaison titre $\times$ émotion. Il permet de tester si le score attribué par un modèle est statistiquement compatible avec la moyenne des réponses humaines pour cette même cellule.

La statistique de test est calculée comme suit :

$$t = \frac{\bar{H}_{t,i} - M_{t,i}}{\sigma_{t,i} / \sqrt{n_{t,i}}}$$

Dans cette formule, $\bar{H}_{t,i}$ désigne la moyenne humaine pour le titre t et l'émotion i , $M_{t,i}$ le score attribué par le modèle, $\sigma_{t,i}$ l'écart-type des réponses humaines et $n_{t,i}$ le nombre de répondants humains pour cette cellule. Le numérateur mesure l'écart brut entre le modèle et la moyenne humaine. Le dénominateur, appelé erreur standard, exprime la précision avec laquelle la moyenne humaine est estimée.

Lorsque la valeur p associée au test est inférieure au seuil de 5 %, l'écart est considéré comme statistiquement significatif. Le taux global de déviations significatives, calculé sur l'ensemble des cellules d'un modèle, constitue alors un indicateur synthétique de la fréquence à laquelle le modèle produit des évaluations statistiquement différentes de la moyenne humaine.

Dans cette recherche, ces tests cellule par cellule sont utilisés comme un indicateur exploratoire agrégé plutôt que comme une série de conclusions indépendantes sur chaque titre et chaque émotion. Les proportions de déviations significatives sont donc rapportées au seuil nominal de 5 %. Elles doivent être interprétées avec prudence, car aucune correction stricte pour comparaisons multiples n'est appliquée à ce niveau.

Test de Wilcoxon signé apparié

Afin de comparer les modèles entre eux, un test de Wilcoxon signé apparié est appliqué à chaque paire de modèles. Pour une paire donnée, le test compare titre par titre les MAE obtenus par les deux modèles. Pour chaque titre, la différence de MAE est calculée de la manière suivante :

$$d_i = \text{MAE}_{m_1,i} - \text{MAE}_{m_2,i}$$

Les différences nulles sont ignorées, puis les valeurs absolues des différences restantes sont classées par rang croissant. Le test compare ensuite la somme des rangs associés aux différences positives, notée W^+ , et la somme des rangs associés aux différences négatives, notée W^- . Dans un test bilatéral, la statistique de test peut être définie comme :

$$W = \min(W^+, W^-)$$

Une valeur faible indique que les différences vont majoritairement dans le même sens, ce qui suggère qu'un modèle obtient des MAE plus faibles que l'autre de manière relativement systématique. Ce test non paramétrique est préféré au test t de Student en raison de la taille limitée de l'échantillon, soit 28 titres, et de l'absence d'hypothèse de normalité.

Dans la mesure où six paires de modèles sont comparées simultanément, une correction de Bonferroni est appliquée. Le seuil de significativité est donc ramené à :

$$\alpha = \frac{0,05}{6} \approx 0,0083$$

Cette correction permet de limiter le risque d'erreur de type I lié aux comparaisons multiples.

Test de Friedman par émotion

Enfin, un test de Friedman est conduit pour chacune des huit émotions de la roue de Plutchik. Ce test non paramétrique à mesures répétées permet de déterminer si les quatre modèles se comportent de manière significativement différente pour une émotion donnée. Il est appliqué séparément à chaque émotion, à partir des écarts observés entre les scores des modèles et la référence humaine sur les 28 titres.

La statistique de Friedman, notée ici χ_F^2 , est calculée comme suit :

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1)$$

Dans cette formule, n correspond au nombre de titres, soit 28, k au nombre de modèles comparés, soit 4, et R_j à la somme des rangs attribués au modèle j sur l'ensemble des titres pour l'émotion considérée. Pour chaque titre et pour une émotion donnée, les écarts des quatre modèles à la référence humaine sont classés par ordre croissant : le modèle qui s'écarte le moins reçoit le rang 1, tandis que celui qui s'en écarte le plus reçoit le rang 4.

Si les rangs sont répartis de manière aléatoire entre les modèles, la statistique χ_F^2 reste faible et la valeur p est élevée. À l'inverse, si certains modèles obtiennent systématiquement des rangs plus faibles ou plus élevés, la statistique augmente et la valeur p diminue. Un résultat significatif au seuil de 5 % indique donc que les modèles ne se comportent pas de manière équivalente pour l'émotion étudiée.

Comme le test est répété pour les huit émotions, les résultats doivent être interprétés avec prudence. Une correction de Bonferroni peut être mobilisée à titre indicatif afin de tenir compte de ces comparaisons multiples, ce qui ramène le seuil de significativité à :

$$\alpha = \frac{0,05}{8} = 0,00625$$

Ce test complète les comparaisons globales en révélant d'éventuelles différences de performance selon la nature de l'émotion évaluée.

3.4 Limites méthodologiques

Toute démarche empirique repose sur des choix qui délimitent son champ d'application et conditionnent l'interprétation de ses résultats. Les limites présentées ici ne remettent pas en cause la validité du dispositif retenu, mais invitent à une lecture prudente des conclusions qui en découlent. Elles s'articulent autour de plusieurs dimensions : la représentativité de l'échantillon humain, le format de réponse par curseur, la réduction inhérente au cadre théorique de Plutchik, les contraintes propres au corpus de titres, les spécificités des modèles de langage évalués et la variabilité de leurs réponses.

Représentativité de l'échantillon humain.

La base humaine de référence repose sur les réponses d'une quarantaine de participants, ce qui permet de dégager des tendances générales mais ne suffit pas à représenter l'ensemble des perceptions émotionnelles possibles. La taille de cet échantillon s'explique également par les contraintes matérielles et budgétaires propres à ce mémoire. La collecte ayant reposé sur une diffusion gratuite du questionnaire, principalement par les réseaux sociaux et auprès de l'entourage de l'auteur, il n'était pas possible de mettre en place un recrutement plus large, rémunéré ou statistiquement contrôlé.

Il s'agit donc d'un échantillon de convenance, fondé sur la participation volontaire des répondants. La sensibilité émotionnelle est en effet influencée par des facteurs personnels, culturels, linguistiques et contextuels qui introduisent une variabilité difficilement réductible. Le recours à une moyenne atténue une partie de cette hétérogénéité, mais ne la supprime pas. Cette contrainte est en partie prise en compte dans la méthode d'analyse elle-même, notamment à travers le nMAE et les intervalles de confiance, qui intègrent explicitement la dispersion des réponses humaines plutôt que de se limiter à la seule valeur centrale.

Format de réponse par curseur

Une autre limite concerne l'utilisation d'un curseur allant de 0 à 100 pour chaque émotion. Ce format permet aux participants de nuancer leurs réponses et de ne pas choisir une seule émotion, ce qui est utile pour rendre compte de ressentis mixtes ou ambigus.

Cependant, il complique aussi l'interprétation des résultats. Plusieurs émotions peuvent obtenir des scores proches pour un même titre, ce qui rend parfois difficile l'identification d'une émotion dominante. Cette limite est surtout importante pour la première analyse, qui repose sur la comparaison entre l'émotion dominante humaine et celle attribuée par les modèles.

Enfin, les participants peuvent utiliser l'échelle différemment. Certains mobilisent toute l'échelle, tandis que d'autres restent dans des valeurs plus modérées. Les résultats doivent donc être interprétés comme des tendances moyennes plutôt que comme des catégories émotionnelles parfaitement tranchées.

Réduction liée au cadre de Plutchik.

L'analyse repose exclusivement sur les huit émotions de base de la roue de Plutchik. Ce cadre présente l'avantage d'offrir une structure claire et directement opérationnalisable pour la comparaison entre humains et modèles. En contrepartie,

il introduit une simplification de la complexité émotionnelle : des ressentis plus nuancés, composites ou difficiles à catégoriser peuvent se trouver résumés à une ou plusieurs émotions fondamentales, au risque d'appauvrir la richesse des réponses collectées.

Contraintes du corpus de titres.

Le questionnaire porte sur 28 titres d'articles ce qui permet une comparaison structurée mais constitue un échantillon relativement restreint. Par ailleurs, l'analyse repose uniquement sur les titres et non sur le contenu complet des articles. Un titre pris isolément peut être ambigu, incomplet ou interprétable de plusieurs manières, ce qui peut influencer les réponses aussi bien des participants humains que des modèles. Cette contrainte est inhérente au choix de se concentrer sur des stimuli courts qui constituent pourtant un terrain d'évaluation pertinent pour comparer des capacités de reconnaissance émotionnelle dans des conditions identiques.

Spécificités des modèles de langage.

Les réponses produites par les grands modèles de langage peuvent être influencées par la formulation du prompt, les conditions d'interrogation ou les mécanismes internes propres à chaque modèle. Bien que le même cadre de consigne ait été appliqué à l'ensemble des modèles étudiés, des variations de réponse peuvent refléter non seulement des différences de capacité de reconnaissance émotionnelle, mais aussi des différences dans la manière dont chaque modèle traite et interprète la consigne. La comparaison doit donc être comprise comme une évaluation réalisée dans un contexte précis, et non comme une mesure absolue des performances émotionnelles de ces systèmes.

Variabilité des réponses des modèles de langage

Une autre limite concerne la nature probabiliste des réponses produites par les grands modèles de langage. Les réponses de ChatGPT, Claude, Gemini et Piccolo ont été collectées à partir d'une seule exécution du prompt pour chaque titre. Or, les grands modèles de langage ne produisent pas toujours une réponse parfaitement stable lorsqu'ils sont interrogés plusieurs fois avec une même consigne. Les résultats peuvent varier selon la version du modèle, l'interface utilisée, les paramètres de génération ou encore certains éléments internes au modèle.

Dans une perspective méthodologique plus robuste, il aurait été préférable de répéter plusieurs fois la même procédure pour chaque modèle et pour chaque titre. Cela aurait permis d'observer la stabilité des réponses produites et de mesurer

une forme de dispersion propre aux modèles. Une telle approche aurait évité de considérer chaque réponse comme une valeur unique et définitive. Elle aurait plutôt permis de l'interpréter comme une observation parmi plusieurs réponses possibles.

Cette limite est importante car les résultats obtenus dans ce mémoire doivent être compris comme une comparaison située. Ils reposent sur les réponses générées au moment de la collecte et ne constituent donc pas une mesure exhaustive ou parfaitement stable du comportement de chaque modèle.

Chapitre 4

Résultats

Dans ce chapitre, nous présentons les résultats issus de la comparaison entre la base humaine de référence et les réponses produites par les différents grands modèles de langage. Chaque modèle est analysé selon deux méthodes de comparaison. La première repose sur une approche plus générale et consiste à comparer uniquement l'émotion dominante pour chaque titre afin de vérifier si le modèle identifie bien l'émotion principale. La seconde propose une analyse plus approfondie en prenant en compte l'ensemble des réponses émotionnelles afin d'obtenir une vue plus complète des écarts et des rapprochements entre les évaluations humaines et celles des grands modèles de langage.

4.1 Résultats de la première analyse

Cette première analyse vise à examiner dans quelle mesure les différents grands modèles de langage parviennent à détecter l'émotion dominante associée à chaque titre à partir de la base humaine de référence construite précédemment. Pour ce faire, deux indicateurs ont été retenus. Le premier consiste à vérifier s'il existe une correspondance exacte entre l'émotion dominante humaine et l'émotion dominante proposée par le grand modèle de langage. Lorsque cette correspondance n'est pas observée, un second indicateur est mobilisé : il s'agit d'examiner le rang occupé par l'émotion dominante humaine dans la hiérarchie émotionnelle produite par le modèle.

Le tableau ci-dessous présente les résultats obtenus pour chacun des modèles.

Ces premiers résultats montrent des écarts nets entre les différents modèles. Piccolo est celui qui obtient le meilleur taux de correspondance exacte avec la réf-

Modèle	Taux de correspondance exacte (%)	Rang moyen
Gemini	32,14	2,21
ChatGPT	50,00	2,14
Claude	42,86	2,32
Piccolo	67,86	1,68

TABLE 4.1 – Résultats globaux de la première analyse

rence humaine : il identifie correctement l’émotion dominante dans 67,86 % des cas. ChatGPT occupe une position intermédiaire avec un taux de correspondance de 50,00 %, tandis que Claude obtient un résultat légèrement inférieur avec 42,86 %. Gemini est le modèle qui s’éloigne le plus de la base humaine dans cette première analyse, avec un taux de correspondance de 32,14 %.

Le rang moyen apporte une information complémentaire. Plus ce rang est proche de 1, plus l’émotion dominante humaine est placée haut dans la hiérarchie émotionnelle produite par le modèle. Là encore, Piccolo obtient le résultat le plus favorable avec un rang moyen de 1,68. ChatGPT et Gemini présentent des rangs moyens relativement proches tandis que Claude affiche le rang moyen le plus élevé avec 2,32.

4.1.1 Résultats par modèle

Les résultats de cette première analyse montrent que Piccolo est le modèle le plus proche de la référence humaine pour l’identification de l’émotion dominante. Il obtient non seulement le meilleur taux de correspondance exacte, mais il maintient également l’émotion dominante humaine à un rang relativement élevé lorsqu’il ne la place pas en première position.

ChatGPT se situe dans une position intermédiaire. Il retrouve correctement l’émotion dominante humaine dans un cas sur deux, ce qui traduit une capacité de reconnaissance partielle mais moins marquée que celle observée pour Piccolo.

Claude présente un profil proche de ChatGPT, mais légèrement moins favorable. Son taux de correspondance exacte est plus faible et l’émotion dominante humaine tend à être classée un peu plus loin dans la hiérarchie émotionnelle produite par le modèle.

Enfin, Gemini est le modèle qui obtient les résultats les plus faibles dans cette première analyse. Il retrouve moins souvent l’émotion dominante humaine et tend à la classer à une position plus éloignée que les autres modèles.

4.1.2 Analyse du rang de l'émotion dominante humaine

L'examen du rang occupé par l'émotion dominante humaine permet de nuancer les cas de non-correspondance exacte. En effet, lorsqu'un modèle ne place pas l'émotion humaine en première position, il peut néanmoins la classer en deuxième ou en troisième position ce qui traduit une forme de proximité partielle avec la référence humaine.

Dans le cas de Piccolo, l'émotion dominante humaine est placée en première position dans 19 cas sur 28. Lorsqu'elle n'est pas première, elle reste généralement proche du sommet de la hiérarchie puisqu'elle apparaît deux fois en deuxième position, quatre fois en troisième position et trois fois en quatrième position. Ce profil confirme la relative proximité de ce modèle avec la base humaine.

ChatGPT présente une distribution plus dispersée. L'émotion dominante humaine apparaît en première position dans 14 cas mais elle est également reléguée à des rangs plus éloignés dans plusieurs situations. Claude présente une configuration comparable avec une dispersion légèrement plus marquée.

Gemini se distingue par une distribution moins favorable. L'émotion dominante humaine n'apparaît en première position que dans 9 cas. Elle est fréquemment reléguée en deuxième ou troisième position qui indique que le modèle capte parfois partiellement la logique émotionnelle humaine sans toutefois identifier correctement l'émotion principale.

Dans l'ensemble, cette première analyse montre que la reconnaissance de l'émotion dominante varie sensiblement selon les modèles. Piccolo apparaît comme le plus proche de la référence humaine sur ce premier niveau de lecture, tandis que Gemini s'en éloigne davantage. ChatGPT et Claude occupent une position intermédiaire. Ces résultats justifient de poursuivre l'analyse en prenant en compte non plus seulement l'émotion dominante, mais l'ensemble du profil émotionnel attribué à chaque titre.

4.2 Résultats de la deuxième analyse

Dans cette seconde analyse, nous comparons l'ensemble du profil émotionnel attribué à chaque titre. Contrairement à la première analyse, qui se limitait à l'identification de l'émotion dominante, cette approche prend en compte les scores attribués aux huit émotions de la roue de Plutchik. Elle permet donc d'évaluer plus finement la proximité entre les réponses produites par les modèles de langage et la base humaine de référence.

Cette comparaison repose d'abord sur plusieurs indicateurs complémentaires : le

MAE, le nMAE, le taux de couverture par intervalle de confiance à 95 % et le taux de déviations statistiquement significatives. Elle est ensuite complétée par des tests de Wilcoxon appariés entre paires de modèles et par des tests de Friedman conduits séparément pour chaque émotion.

4.2.1 Résultats globaux

Le tableau ci-dessous présente les résultats globaux obtenus pour chaque modèle pour la deuxième analyse. Les meilleures valeurs sont indiquées en gras.

Modèle	MAE ↓	nMAE ↓	Couverture IC95 % ↑	Dév. signif. t-test (%) ↓
Gemini	17,575	0,801	39,29	57,14
ChatGPT	13,887	0,629	42,41	52,68
Claude	12,061	0,552	47,77	48,21
Piccolo	12,380	0,566	46,43	50,45

TABLE 4.2 – Indicateurs d’écart par rapport à la référence humaine. ↓ = valeur plus faible préférable ; ↑ = valeur plus élevée préférable.

Les résultats de la deuxième analyse font apparaître des différences assez nettes entre les modèles. Dans l’ensemble, Claude est celui qui se rapproche le plus de la référence humaine. Il obtient le MAE le plus faible (12,061), le nMAE le plus faible (0,552), le meilleur taux de couverture par intervalle de confiance à 95 % (47,77 %) et le taux de déviations significatives le plus bas (48,21 %). À l’autre extrémité, Gemini est le modèle le plus éloigné de la base humaine. Il présente les valeurs les moins favorables sur tous les indicateurs retenus : un MAE de 17,575, un nMAE de 0,801, une couverture de 39,29 % et un taux de déviations significatives de 57,14 %. ChatGPT se situe entre ces deux extrêmes. Piccolo, de son côté, obtient des résultats très proches de ceux de Claude.

MAE brut

Le MAE donne une première lecture simple des écarts entre les modèles et la moyenne humaine. Sur ce plan, Claude et Piccolo se distinguent assez clairement, avec des écarts moyens un peu supérieurs à 12 points. ChatGPT est légèrement plus éloigné, avec un MAE de 13,887. Gemini se détache davantage, avec un écart moyen supérieur à 17 points. Cela signifie que ses réponses s’écartent plus fortement du profil émotionnel moyen humain que celles des autres modèles.

MAE normalisé (nMAE)

Le nMAE permet d’aller plus loin que le MAE brut, puisqu’il tient compte de la dispersion des réponses humaines. Cet indicateur est important dans cette recherche, car un même écart ne doit pas être interprété de la même manière selon que les répondants humains sont globalement d’accord ou, au contraire, très partagés.

Sur ce point, Claude reste le modèle le plus proche de la référence humaine, suivi de très près par Piccolo. ChatGPT se situe à un niveau intermédiaire, tandis que Gemini reste le plus éloigné. Il faut toutefois souligner que les quatre modèles obtiennent un nMAE inférieur à 1. En moyenne, leurs écarts restent donc inférieurs à une unité d’écart-type humain. Cela suggère que les écarts moyens observés restent globalement contenus par rapport à la dispersion des réponses humaines même si certaines cellules peuvent présenter des écarts plus marqués.

Taux de couverture par intervalle de confiance à 95 %

Le taux de couverture par intervalle de confiance à 95 % propose une lecture plus stricte. Il ne s’agit plus seulement de regarder si le modèle est proche de la moyenne humaine, mais de vérifier si ses réponses se situent dans une zone statistiquement compatible avec la moyenne humaine estimée.

De ce point de vue, les résultats sont plus mitigés. Claude obtient la meilleure couverture, mais celle-ci reste inférieure à 50 %. Piccolo suit de près, puis ChatGPT, tandis que Gemini ferme la marche. Autrement dit, même lorsque les écarts moyens paraissent modérés, les modèles restent souvent en dehors de l’intervalle construit autour de la moyenne humaine.

Ce résultat doit être interprété avec prudence. L’intervalle de confiance à 95 % porte sur l’estimation de la moyenne humaine et non sur la dispersion complète des réponses individuelles. Il constitue donc un critère exigeant, qui évalue la compatibilité des scores des modèles avec la moyenne humaine plutôt qu’avec un répondant humain pris individuellement.

Taux de déviations statistiquement significatives

Le taux de déviations statistiquement significatives constitue la synthèse des tests t à un échantillon réalisés cellule par cellule. Pour chaque combinaison titre $\times$ émotion, le score attribué par le modèle a été comparé à la moyenne humaine de référence. Une cellule est considérée comme significativement déviante lorsque l’écart entre le score du modèle et la moyenne humaine est associé à une valeur p inférieure à 0,05.

Les résultats vont dans le même sens que ceux obtenus avec le taux de couverture par intervalle de confiance. Une part importante des réponses produites par les modèles demeure statistiquement différente de la moyenne humaine. Claude obtient ici encore le résultat le plus favorable, mais près d'une cellule sur deux présente tout de même une déviation significative. Piccolo et ChatGPT présentent des résultats légèrement moins favorables, tandis que Gemini affiche de nouveau la valeur la plus élevée.

Ces proportions doivent toutefois être interprétées avec prudence. Les tests t cellule par cellule sont nombreux et sont utilisés ici comme un indicateur exploratoire agrégé. Ils permettent d'évaluer la fréquence des écarts statistiquement significatifs, mais ne doivent pas être compris comme une conclusion indépendante et définitive pour chaque cellule prise isolément. Ce résultat invite donc à relativiser l'idée d'une forte proximité entre modèles et répondants humains : même les meilleurs résultats observés ici restent imparfaits.

Pris ensemble, ces indicateurs dessinent une tendance assez claire. Claude apparaît comme le modèle le plus proche de la référence humaine dans cette deuxième analyse. Piccolo se situe à un niveau très proche et confirme une bonne cohérence globale avec le profil émotionnel moyen humain. ChatGPT obtient des résultats corrects, mais plus éloignés que ceux de Claude et de Piccolo. Gemini, en revanche, se distingue par une distance plus marquée à la base humaine. Il faut également noter que l'écart entre Claude et Piccolo reste limité, alors que la différence avec Gemini est beaucoup plus nette.

Ces résultats complètent utilement ceux de la première analyse. Dans l'identification de l'émotion dominante, Piccolo apparaissait comme le modèle le plus performant. En revanche, lorsque l'on prend en compte l'ensemble du profil émotionnel, Claude passe légèrement devant. Ce décalage montre qu'un modèle peut repérer correctement l'émotion principale sans nécessairement reproduire avec la même précision la répartition complète des scores émotionnels observée dans la base humaine de référence.

Visualisation des profils émotionnels pour des titres à formulation implicite

Afin de compléter les indicateurs quantitatifs présentés précédemment, plusieurs graphiques radar ont été réalisés. Ces visualisations permettent de comparer, pour un titre donné, le profil émotionnel moyen humain aux profils produits par les différents modèles de langage. Chaque axe correspond à l'une des huit émotions de la roue de Plutchik, et les valeurs sont exprimées sur une échelle allant de 0 à 100.

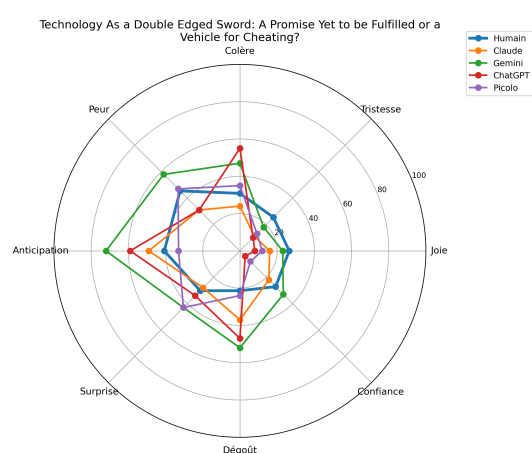
Ces graphiques doivent être compris comme des outils descriptifs. Ils ne remplacent pas les indicateurs statistiques, mais permettent de visualiser plus directement la forme des profils émotionnels et les écarts entre les modèles et la base humaine de référence.

Une attention particulière est portée aux titres à formulation implicite, métaphorique ou ambivalente. Ce choix est cohérent avec la revue de littérature, qui souligne que les modèles peuvent rencontrer davantage de difficultés lorsque les signaux affectifs sont moins directement exprimés. Les exemples sélectionnés permettent donc d'observer si les modèles reproduisent seulement l'émotion dominante ou s'ils parviennent également à restituer la structure complète du profil émotionnel humain.

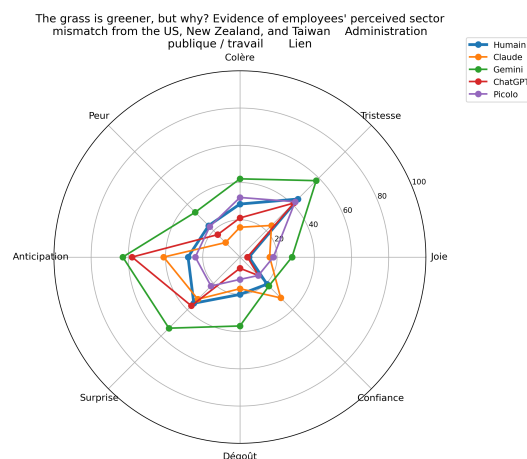
La figure 4.1 illustre quatre titres dont la formulation repose sur une expression métaphorique, idiomatique ou ambivalente. Ces exemples permettent de compléter l'analyse statistique par une lecture plus qualitative des profils émotionnels.

Dans ces cas, l'enjeu n'est pas uniquement d'identifier l'émotion dominante, mais aussi de restituer la structure générale du profil émotionnel humain. Un modèle peut en effet se rapprocher de la base humaine sur l'émotion principale tout en surestimant ou sous-estimant fortement certaines émotions secondaires. À l'inverse, certains modèles peuvent ne pas reproduire exactement la hiérarchie humaine, mais proposer un profil globalement proche dans sa forme.

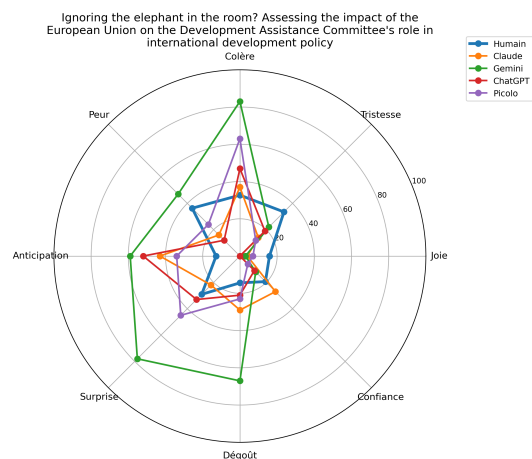
Ces visualisations montrent ainsi que les titres implicites constituent un terrain d'observation particulièrement intéressant. Ils obligent les modèles à interpréter des formulations qui ne portent pas toujours directement l'émotion dans des mots explicites, mais dans une structure figurative, une opposition ou une expression idiomatique. Les graphiques radar rendent visibles ces différences d'interprétation entre les modèles et la base humaine de référence.



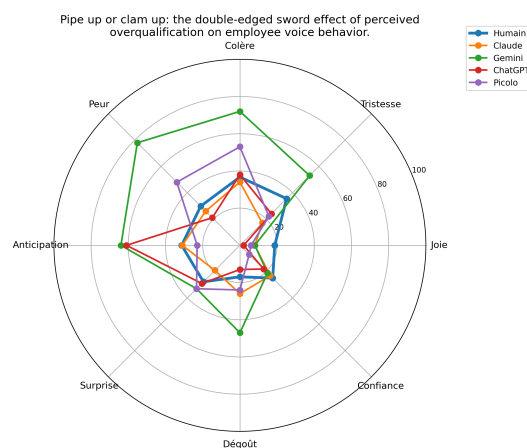
(a) Titre 23 : *Technology As a Double Edged Sword*



(b) Titre 25 : *The grass is greener, but why ?*



(c) Titre 26 : *Ignoring the elephant in the room ?*



(d) Titre 28 : *Pipe up or clam up*

FIGURE 4.1 – Profils émotionnels comparés pour des titres à formulation implicite, métaphorique ou ambivalente

Comparaisons statistiques entre modèles : tests de Wilcoxon appariés

Les indicateurs précédents permettent de comparer les performances moyennes des modèles mais ils ne suffisent pas à déterminer si les différences observées sont statistiquement significatives. Des tests de Wilcoxon signés appariés ont donc été appliqués à chaque paire de modèles, en comparant titre par titre les MAE obtenus. Ce test non paramétrique est adapté à la taille limitée de l'échantillon, soit 28 titres, et ne suppose pas la normalité des différences observées.

Comme six paires de modèles sont comparées, une correction de Bonferroni est appliquée. Le seuil de significativité est donc ramené à $\alpha = 0,05/6 \approx 0,0083$.

Modèle 1	Modèle 2	W	p -value	Modèle le plus proche
Gemini	ChatGPT	49,0	0,0002	ChatGPT
Gemini	Claude	7,0	<0,001	Claude
Gemini	Piccolo	15,0	<0,001	Piccolo
ChatGPT	Claude	88,0	0,0076	Claude
ChatGPT	Piccolo	114,5	0,0439	–
Claude	Piccolo	191,0	0,7966	–

TABLE 4.3 – Tests de Wilcoxon appariés entre paires de modèles

Les résultats permettent d’établir trois constats principaux.

Premièrement, Gemini se distingue significativement des trois autres modèles dans toutes les comparaisons réalisées. Les différences avec Claude et Piccolo sont associées à des valeurs $p < 0,001$, tandis que la différence avec ChatGPT est également significative ($p = 0,0002$). Ces résultats restent significatifs après correction de Bonferroni, ce qui confirme que Gemini présente des MAE plus élevés que les trois autres modèles de manière suffisamment systématique.

Deuxièmement, Claude apparaît significativement plus proche de la référence humaine que ChatGPT ($p = 0,0076$). Cette différence reste inférieure au seuil corrigé de Bonferroni ($\alpha = 0,0083$), mais elle demeure proche de ce seuil. La hiérarchie observée dans les indicateurs descriptifs se confirme donc au niveau statistique, tout en appelant une interprétation prudente.

Troisièmement, les comparaisons entre ChatGPT et Piccolo ($p = 0,0439$) et entre Claude et Piccolo ($p = 0,7966$) ne sont pas significatives après correction de Bonferroni. Il n’est donc pas possible de conclure à une différence statistiquement significative entre ces paires de modèles sur la base des données disponibles. En particulier, même si Claude présente des indicateurs descriptifs légèrement meilleurs que Piccolo, cette différence n’est pas confirmée par le test de Wilcoxon.

Ces tests mettent donc en évidence une structure assez claire. Gemini forme le modèle le plus éloigné de la base humaine. Claude et Piccolo apparaissent comme les deux modèles les plus proches de la référence humaine, sans qu’il soit possible d’établir une différence significative entre eux. ChatGPT occupe une position intermédiaire : il reste significativement moins proche de la base humaine que Claude, mais ne se distingue pas de manière significative de Piccolo.

Comparaison par émotion : tests de Friedman

Les tests de Friedman permettent d'examiner si les différences entre modèles varient selon l'émotion considérée. Ils sont appliqués séparément pour chacune des huit émotions de la roue de Plutchik. Pour chaque émotion, les écarts des quatre modèles à la référence humaine sont comparés sur les 28 titres.

Émotion	χ_F^2	p	Gemini	ChatGPT	Claude	Piccolo
Joie	2,64	0,451	10,72	9,12	11,75	9,17
Tristesse	3,31	0,347	14,23	10,33	10,86	12,63
Colère	2,20	0,532	15,89	13,94	14,04	13,27
Peur	0,44	0,932	14,92	13,24	12,61	11,34
Anticipation	26,22	<0,001	30,88	23,06	16,01	14,49
Surprise	7,97	0,047	18,42	12,58	12,14	10,99
Dégoût	14,29	0,003	15,23	13,44	10,31	12,88
Confiance	20,78	<0,001	20,31	15,38	8,77	14,27

TABLE 4.4 – Tests de Friedman par émotion et MAE moyens par modèle

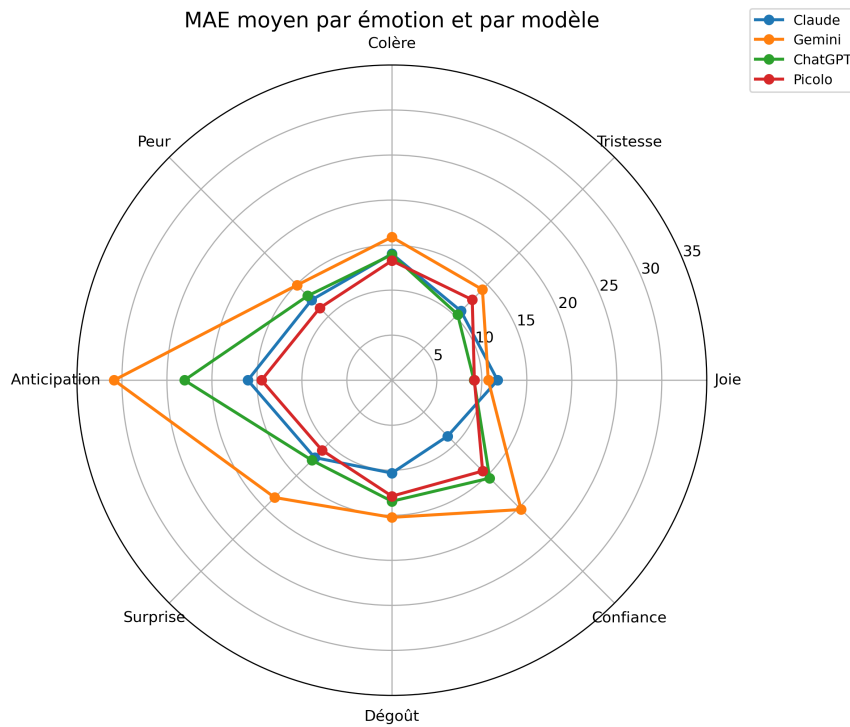


FIGURE 4.2 – MAE moyen par émotion et par modèle

Les résultats du tableau 4.4 peuvent être complétés par la figure 4.2, qui propose

une visualisation radar des MAE moyens par émotion.

Pour la joie, la tristesse, la colère et la peur, aucune différence statistiquement significative n'apparaît entre les quatre modèles. Sur ces émotions, les performances restent donc relativement comparables quel que soit le système évalué. Cette homogénéité peut s'expliquer par le fait que ces émotions sont souvent plus directement identifiables dans les formulations textuelles, même si cette interprétation doit rester prudente.

Au seuil nominal de 5 %, des différences significatives apparaissent pour quatre émotions : l'anticipation, la surprise, le dégoût et la confiance. Toutefois, cette conclusion doit être nuancée, car huit tests de Friedman sont réalisés, soit un par émotion. Si l'on applique une correction stricte de Bonferroni, le seuil de significativité devient $\alpha = 0,05/8 = 0,00625$. Dans ce cas, seules l'anticipation, le dégoût et la confiance demeurent significatives. La surprise, dont la valeur p est égale à 0,047, doit donc être interprétée avec prudence.

L'anticipation et la confiance produisent les écarts les plus marqués. Pour l'anticipation, Gemini obtient un MAE de 30,88, contre 14,49 pour Piccolo, soit un écart de plus de 16 points. Pour la confiance, Claude se distingue avec le MAE le plus faible de cette analyse par émotion (8,77), alors que Gemini atteint 20,31. Ces deux émotions renvoient à des dimensions moins immédiatement lisibles dans un titre court : l'anticipation suppose une projection temporelle, tandis que la confiance implique souvent une relation de fiabilité ou de crédibilité plus implicite. Cela peut expliquer qu'elles révèlent davantage les différences entre modèles.

Dans l'ensemble, cette analyse par émotion nuance les résultats globaux. Aucun modèle n'est systématiquement supérieur sur toutes les émotions. Certaines catégories, comme la joie ou la peur, donnent lieu à des performances relativement proches entre modèles. D'autres, comme l'anticipation ou la confiance, mettent davantage en évidence leurs différences. Cela montre que la qualité de reconnaissance émotionnelle d'un modèle ne dépend pas seulement de sa performance moyenne mais aussi de sa capacité à traiter certaines émotions plus difficiles ou plus ambiguës.

4.3 Synthèse des principaux résultats

Le tableau 4.5 résume les performances des quatre modèles sur l'ensemble des indicateurs retenus dans ce chapitre.

Les résultats présentés dans ce chapitre permettent de dégager plusieurs constats principaux, selon le niveau d'analyse retenu.

Modèle	Analyse 1			Analyse 2			Synthèse	
	Exact match	Rang moy.	MAE	nMAE	IC95 %	Dév. signif.	Position statistique	Position globale
Gemini	32,14 %	2,21	17,58	0,801	39,29 %	57,14 %	Moins proche	4
ChatGPT	50,00 %	2,14	13,89	0,629	42,41 %	52,68 %	Intermédiaire	3
Claude	42,86 %	2,32	12,06	0,552	47,77 %	48,21 %	Proche de la référence	1
Piccolo	67,86 %	1,68	12,38	0,566	46,43 %	50,45 %	Proche de Claude	2

TABLE 4.5 – Synthèse comparative des résultats obtenus par les différents modèles

La première analyse, centrée sur l’identification de l’émotion dominante, montre que les modèles n’obtiennent pas tous le même degré de proximité avec la référence humaine. Sur ce premier niveau de lecture, Piccolo apparaît comme le modèle le plus proche de la référence humaine : il retrouve correctement l’émotion dominante humaine dans 67,86 % des cas et maintient, lorsque la correspondance n’est pas exacte, cette émotion à un rang moyen de 1,68 dans sa hiérarchie émotionnelle. ChatGPT et Claude présentent des résultats intermédiaires, tandis que Gemini se distingue par le taux d’identification correcte le plus faible (32,14 %).

La seconde analyse, fondée sur la comparaison du profil émotionnel complet, apporte une lecture différente. Lorsque l’on prend en compte l’ensemble des scores attribués aux huit émotions plutôt que la seule émotion principale, Claude apparaît comme le modèle le plus proche de la référence humaine sur l’ensemble des indicateurs retenus. Piccolo obtient des résultats très proches, confirmant sa bonne cohérence globale avec le profil émotionnel moyen humain. ChatGPT se situe à un niveau intermédiaire, alors que Gemini demeure le modèle le plus éloigné de la base humaine. Ce décalage entre les deux analyses illustre qu’un modèle peut identifier correctement l’émotion principale sans reproduire avec la même précision la répartition complète des scores émotionnels.

Les tests statistiques confirment et précisent en partie cette hiérarchie. Les tests de Wilcoxon appariés montrent que Gemini se distingue significativement des trois autres modèles, ce qui confirme qu’il est le système le moins proche de la référence humaine dans le cadre de cette étude. Claude apparaît également significativement plus proche de la référence humaine que ChatGPT, même si cette différence reste proche du seuil corrigé de Bonferroni. En revanche, les différences entre Claude et Piccolo, ainsi qu’entre ChatGPT et Piccolo, ne sont pas statistiquement significatives après correction. Il n’est donc pas possible de conclure à une différence statistiquement établie entre ces paires de modèles sur la base des données disponibles.

L’analyse par émotion complète encore ce tableau. Les écarts entre modèles ne se manifestent pas de manière uniforme selon les émotions étudiées. Pour la joie, la tristesse, la colère et la peur, les performances des quatre modèles restent comparables et ne donnent lieu à aucune différence significative. Au seuil nominal de

5 %, des écarts significatifs apparaissent pour l'anticipation, la surprise, le dégoût et la confiance. Toutefois, après correction de Bonferroni pour les huit tests réalisés, seules l'anticipation, le dégoût et la confiance demeurent significatives. Ces émotions, moins directement lisibles dans un titre court, semblent constituer un terrain plus discriminant pour évaluer les capacités de reconnaissance émotionnelle des modèles.

Au vu des résultats, deux enseignements principaux se dégagent. D'une part, les grands modèles de langage étudiés sont capables de produire des évaluations émotionnelles présentant un certain degré de proximité avec les réponses humaines. D'autre part, cette proximité reste partielle : même les meilleures performances observées témoignent d'écarts fréquents, dont une part importante apparaît statistiquement significative au seuil retenu. Ces constats servent de base à la discussion finale du mémoire.

Chapitre 5

Conclusion

Ce mémoire avait pour objectif d'examiner dans quelle mesure plusieurs grands modèles de langage sont capables de reconnaître les émotions de manière comparable à des répondants humains lorsqu'ils sont confrontés aux mêmes stimuli textuels. Pour répondre à cette question, nous avons comparé les réponses de Claude, Gemini, Piccolo et ChatGPT à une base humaine de référence construite à partir d'un questionnaire portant sur 28 titres d'articles. Ces titres comprenaient des formulations explicites, implicites, métaphoriques ou ambiguës. La roue des émotions de Plutchik a été utilisée comme cadre théorique commun afin de comparer les réponses autour de huit émotions : la joie, la tristesse, la colère, la peur, l'anticipation, la surprise, le dégoût et la confiance.

La revue de littérature a montré que les grands modèles de langage disposent de capacités affectives réelles, mais que celles-ci restent variables selon les tâches, les protocoles de prompting, la langue utilisée et le type de signal émotionnel étudié. Les travaux existants soulignent notamment que les émotions explicites sont généralement plus faciles à traiter que les phénomènes plus implicites comme l'ironie, le sarcasme ou la subjectivité. Dans ce contexte, l'un des apports de ce mémoire a été de proposer une comparaison directe entre plusieurs modèles conversationnels à partir d'un même corpus, d'un même cadre théorique et d'une même base humaine de référence.

Les résultats obtenus permettent de dégager plusieurs enseignements. La première analyse montre que les modèles n'atteignent pas tous le même niveau de proximité avec les réponses humaines lorsqu'il s'agit d'identifier l'émotion dominante. Sur ce premier niveau de lecture, Piccolo apparaît comme le modèle le plus proche de la référence humaine. Il retrouve correctement l'émotion dominante humaine dans 67,86 % des cas et maintient généralement cette émotion à un rang élevé lorsqu'elle

n'apparaît pas en première position. À l'inverse, Gemini est le modèle qui s'éloigne le plus de la base humaine, tandis que ChatGPT et Claude occupent une position intermédiaire.

La seconde analyse apporte une image plus nuancée. Lorsqu'on ne se limite plus à l'émotion dominante mais que l'on prend en compte l'ensemble du profil émotionnel, Claude apparaît comme le modèle le plus proche de la référence humaine. Piccolo obtient des résultats très proches, ChatGPT présente des performances correctes mais moins solides, et Gemini reste le modèle le plus éloigné. Les tests statistiques confirment que Gemini se distingue significativement des autres modèles et que Claude est significativement plus proche de la base humaine que ChatGPT. En revanche, les écarts entre Claude et Piccolo ne sont pas suffisamment marqués pour être considérés comme statistiquement établis dans le cadre de cette étude.

L'analyse par émotion montre aussi que les différences entre modèles ne se manifestent pas de manière uniforme. Pour certaines émotions, comme la joie, la tristesse, la colère ou la peur, les écarts restent limités. En revanche, des différences plus nettes apparaissent pour l'anticipation, le dégoût et la confiance. La surprise présente également un écart au seuil nominal, mais celui-ci doit être interprété avec prudence après correction de Bonferroni. Ces résultats suggèrent que les différences entre modèles se révèlent surtout lorsque l'émotion est moins directement lisible dans le titre ou lorsqu'elle demande une forme d'interprétation plus implicite.

Dans l'ensemble, les résultats montrent que les grands modèles de langage étudiés sont capables de produire des évaluations émotionnelles présentant un certain degré de proximité avec les réponses humaines. Il ne serait toutefois pas juste d'en conclure qu'ils reconnaissent les émotions comme des humains. Cette proximité reste partielle et varie selon le modèle, le niveau d'analyse retenu et l'émotion examinée. Un modèle peut ainsi bien repérer l'émotion principale d'un titre sans reproduire précisément la répartition complète des scores émotionnels observée chez les répondants humains.

Ce travail présente néanmoins plusieurs limites. La base humaine de référence repose sur un échantillon restreint d'environ quarante participants, recrutés selon une méthode de convenance. Le corpus se limite à 28 titres d'articles et l'analyse porte uniquement sur les titres, sans tenir compte du contenu complet des articles. Le choix de la roue des émotions de Plutchik fournit un cadre clair, mais réduit nécessairement la complexité émotionnelle à huit catégories de base. À cela s'ajoute le fait que les modèles de langage ont été interrogés une seule fois pour chaque titre. Les réponses obtenues doivent donc être comprises comme une me-

sure ponctuelle, susceptible de varier selon la version du modèle, l'interface utilisée ou les paramètres de génération. Enfin, le format de réponse par curseur de 0 à 100 permet une certaine nuance, mais peut aussi compliquer l'identification d'une émotion dominante lorsque plusieurs émotions obtiennent des scores proches. Ces limites n'annulent pas les résultats obtenus, mais invitent à les interpréter avec prudence.

Plusieurs prolongements peuvent être envisagés. Il serait d'abord intéressant d'élargir l'échantillon humain afin de renforcer la robustesse de la base de référence. Il serait aussi pertinent de travailler sur un corpus plus vaste et plus diversifié, en incluant davantage de titres ironiques, métaphoriques ou ambigus. Une autre piste consisterait à comparer les réponses des modèles non plus seulement sur des titres, mais sur des textes complets, afin de mesurer l'effet du contexte sur la reconnaissance émotionnelle. Il serait également utile de répéter plusieurs fois le même prompt auprès de chaque modèle, afin d'étudier la stabilité des réponses produites et de ne pas dépendre d'une seule génération.

Une perspective de recherche serait également d'intégrer une baseline de comparaison. Dans ce mémoire, les modèles sont comparés entre eux et à une base humaine de référence, mais ils ne sont pas confrontés à une méthode simple ou générique. Une future étude pourrait par exemple ajouter une attribution aléatoire des émotions, une distribution uniforme des scores, un modèle classique d'analyse de sentiment ou un modèle spécialisé en reconnaissance des émotions. Cette baseline permettrait de mieux mesurer la valeur ajoutée réelle des grands modèles de langage et de vérifier si leurs réponses s'éloignent réellement de ce qu'une méthode plus simple pourrait produire.

En conclusion, ce mémoire montre que les grands modèles de langage peuvent approcher certaines logiques de reconnaissance émotionnelle humaine, sans les reproduire parfaitement ni uniformément. Parmi les modèles étudiés, Piccolo se distingue dans l'identification de l'émotion dominante, tandis que Claude apparaît légèrement plus proche de la référence humaine lorsqu'on considère l'ensemble du profil émotionnel. Gemini est le modèle qui s'éloigne le plus de cette base humaine. Ces résultats confirment que la reconnaissance émotionnelle par les grands modèles de langage constitue un objet de recherche important, à la fois prometteur et encore largement ouvert. Ils invitent aussi à poursuivre les comparaisons dans des dispositifs plus larges, plus répétés et mieux contrôlés.

Bibliographie

- [1] Mostafa M. AMIN et al. « A Wide Evaluation of ChatGPT on Affective Computing Tasks ». In : *arXiv preprint arXiv :2308.13911* (2023). arXiv : 2308.13911 [cs.AI].
- [2] ANTHROPIC. *The Claude 3 Model Family : Opus, Sonnet, Haiku*. Model card. 2024. URL : <https://www.anthropic.com/news/claude-3-family>.
- [3] Yuntao BAI et al. « Constitutional AI : Harmlessness from AI Feedback ». In : *arXiv preprint arXiv :2212.08073* (2022).
- [4] Ahmed BELKHIR et Fatiha SADAT. « Beyond Information : Is ChatGPT Empathetic Enough ? ». In : *Proceedings of Recent Advances in Natural Language Processing*. 2023, p. 159-169. DOI : 10.26615/978-954-452-092-2_018.
- [5] Joost BROEKENS et al. « Fine-grained Affective Processing Capabilities Emerging from Large Language Models ». In : *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2023.
- [6] Tom B. BROWN et al. « Language Models are Few-Shot Learners ». In : *Advances in Neural Information Processing Systems*. T. 33. 2020, p. 1877-1901.
- [7] Alessio BUSCEMI et Daniele PROVERBIO. « ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis ». In : *arXiv preprint arXiv :2402.01715* (2024).
- [8] CONTRIBUTEURS DE WIKIPÉDIA. *Grand modèle de langage*. Consulté le 21 avril 2026. 2026. URL : https://fr.wikipedia.org/wiki/Grand_mod%C3%A8le_de_langage.
- [9] CONTRIBUTEURS DE WIKIPÉDIA. *Ingenierie de prompt*. Consulté le 21 avril 2026. 2026. URL : https://fr.wikipedia.org/wiki/Ing%C3%A9nierie_de_prompt.

- [10] Dorottya DEMSZKY et al. « GoEmotions : A Dataset of Fine-Grained Emotions ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, p. 4040-4054. DOI : 10.18653/v1/2020.acl-main.372.
- [11] Zohar ELYOSEPH et al. « ChatGPT Outperforms Humans in Emotional Awareness Evaluations ». In : *Frontiers in Psychology* 14 (2023), p. 1199058. DOI : 10.3389/fpsyg.2023.1199058.
- [12] GEMINI TEAM. « Gemini : A Family of Highly Capable Multimodal Models ». In : *arXiv preprint arXiv :2312.11805* (2024).
- [13] Fabrizio GILARDI, Meysam ALIZADEH et Maël KUBLI. « ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks ». In : *Proceedings of the National Academy of Sciences* 120.30 (2023), e2305016120. DOI : 10.1073/pnas.2305016120.
- [14] Aaron GRATTAFFIORI et al. « The Llama 3 Herd of Models ». In : *arXiv preprint arXiv :2407.21783* (2024).
- [15] Albert Q. JIANG et al. « Mixtral of Experts ». In : *arXiv preprint arXiv :2401.04088* (2024).
- [16] Florian LECOURT, Madalina CROITORU et Konstantin TODOROV. « “Only ChatGPT Gets Me” : An Empirical Analysis of GPT versus Other Large Language Models for Emotion Detection in Text ». In : *Companion Proceedings of the ACM Web Conference 2025*. 2025, p. 2603-2611. DOI : 10.1145/3701716.3718375.
- [17] Zheng LIAN et al. « GPT-4V with Emotion : A Zero-shot Benchmark for Generalized Emotion Recognition ». In : *arXiv preprint arXiv :2312.04293* (2024). arXiv : 2312.04293 [cs.CV].
- [18] Saif M. MOHAMMAD et al. « SemEval-2018 Task 1 : Affect in Tweets ». In : *Proceedings of the 12th International Workshop on Semantic Evaluation*. 2018, p. 1-17. DOI : 10.18653/v1/S18-1001.
- [19] El Habib NFAOUI et Hanane ELFAIK. « Evaluating Arabic Emotion Recognition Task Using ChatGPT Models : A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning ». In : *Journal of Theoretical and Applied Electronic Commerce Research* 19 (2024), p. 1118-1141. DOI : 10.3390/jtaer19020058.
- [20] Long OUYANG et al. « Training Language Models to Follow Instructions with Human Feedback ». In : *arXiv preprint arXiv :2203.02155* (2022).
- [21] Rosalind W. PICARD. *Affective Computing*. Cambridge, MA : MIT Press, 1997.

- [22] Robert PLUTCHIK. *Emotion : A Psychoevolutionary Synthesis*. New York : Harper & Row, 1980.
- [23] Robert PLUTCHIK. « The Nature of Emotions ». In : *American Scientist* 89.4 (2001), p. 344-350.
- [24] James A. RUSSELL. « A Circumplex Model of Affect ». In : *Journal of Personality and Social Psychology* 39.6 (1980), p. 1161-1178. DOI : 10.1037/h0077714.
- [25] Kristina SCHAAFF, Caroline REINIG et Tim SCHLIPPE. « Exploring ChatGPT’s Empathic Abilities ». In : *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2023.
- [26] Petter TÖRNBERG. « ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning ». In : *PLOS ONE* 18.11 (2023), e0291980. DOI : 10.1371/journal.pone.0291980.
- [27] UCLouvain. *Piccolo*. Mistral Medium, Large Language Model. 2026. URL : <https://piccolo.uclouvain.be>.
- [28] Ashish VASWANI et al. « Attention Is All You Need ». In : *Advances in Neural Information Processing Systems*. T. 30. 2017.
- [29] Naoya WAKE et al. « Bias in Emotion Recognition with ChatGPT ». In : *arXiv preprint arXiv :2310.11753* (2023).
- [30] Yiqun ZHANG et al. « Affective Computing in the Era of Large Language Models : A Survey from the NLP Perspective ». In : *arXiv preprint arXiv :2408.04638* (2024).

Annexe A

Titre du questionnaire

Cette annexe présente la liste complète des titres d'articles utilisés dans le cadre du questionnaire soumis aux répondants humains et aux grands modèles de langage.

1. *Solastalgia : the distress caused by environmental change*
2. *Optimism and Hope in Chronic Disease*
3. *Evidence of prehistoric human in the Falkland Islands*
4. *Workplace bullying : a tale of adverse consequences*
5. *Positive Affect Evoked by Music and the Arts*
6. *Loss, trauma, and human resilience : have we underestimated the human capacity to thrive after extremely aversive events ?*
7. *Psychological Adjustment to Spousal Bereavement in Older Adults*
8. *Power Outages and Community Health*
9. *Mental health response to community disasters*
10. *Loneliness and social isolation as risk factors for mortality*
11. *Quantifying energy transition vulnerability helps more just and effective climate policy*
12. *Health Care Professionals' Experience of Using AI*
13. *Fruit and Vegetable Intake and Mental Health in Adults*
14. *Benefits of wellness interventions for persons with chronic and disabling conditions*
15. *Identifying and understanding the factors that influence the functioning of integrated healthcare systems in the NHS*
16. *Bodily maps of musical sensations across cultures*
17. *The Cosmic History of Star Formation*

18. *Resilience in families of children with developmental disabilities*
19. *The evolution and future of research on Nature-based Solutions to address societal challenges*
20. *Riding the waves of family firm internationalization*
21. *Citizen scientists report global rapid reductions in the visibility of stars from 2011 to 2022*
22. *Blind trust ? The importance and interplay of parasocial relationships and advertising disclosures in explaining influencers' persuasive effects on their followers*
23. *Technology As a Double Edged Sword : A Promise Yet to be Fulfilled or a Vehicle for Cheating ?*
24. *Should I stay or should I go ? A dependence model of breakups*
25. *The grass is greener, but why ? Evidence of employees' perceived sector mismatch from the US, New Zealand, and Taiwan*
26. *Ignoring the elephant in the room ? Assessing the impact of the European Union on the Development Assistance Committee's role in international development policy*
27. *Too much of a good thing ? Visitor Congestion Management Issues for Popular World Heritage Tourist Attractions*
28. *Pipe up or clam up : the double-edged sword effect of perceived over qualification on employee voice behavior*

Annexe B

Tableau démographique

TABLE B.1 – Répartition géographique approximative des répondants par zone après imputation des coordonnées non exploitables

Zone géographique estimée	Nombre de répondants	Pourcentage
Bruxelles-Capitale	12	33,3 %
Brabant wallon	7	19,4 %
Hainaut	7	19,4 %
Liège	6	16,7 %
Namur	3	8,3 %
Flandre orientale	1	2,8 %
Total	36	100,0 %

Annexe C

Résultats des grands modèle de langage

Voici les résultats obtenus pour chaque grand modèle de langage.

N°	Titre	Jolie	Tristesse	Colère	Peur	Anticipation	Surprise	Dégoût	Confiance	Émotion dominante	Intensité dominante
1	Solastalgia: the distress caused by environmental change	2	82	28	61	18	22	12	4	Tristesse	82
2	Optimism and Hope in Chronic Disease	74	18	3	12	68	14	0	64	Jolie	74
3	Evidence of prehistoric human in the Falkland Islands	38	2	0	4	56	72	0	22	Surprise	72
4	Workplace bullying: a tale of adverse consequences	0	67	79	36	12	9	58	1	Colère	79
5	Positive Affect Evoked by Music and the Arts	88	2	0	0	34	26	0	41	Jolie	88
6	Loss, trauma, and human resilience: have we underestimated the human capacity to thrive after extremely aversive events?	22	71	14	29	58	18	5	63	Tristesse	71
7	Psychological Adjustment to Spousal Bereavement in Older Adults	8	84	4	26	31	6	0	28	Tristesse	84
8	Power Outages and Community Health	0	39	23	67	44	21	6	12	Peur	67
9	Mental health response to community disasters	1	49	18	72	51	17	4	21	Peur	72
10	Loneliness and social isolation as risk factors for mortality	0	86	19	63	11	8	5	2	Tristesse	86
11	Quantifying energy transition vulnerability helps more just and effective climate policy	26	24	16	28	71	19	3	69	Anticipation	71
12	Health Care Professionals' Experience of Using AI	19	6	4	28	62	34	2	51	Anticipation	62
13	Fruit and Vegetable Intake and Mental Health in Adults	57	4	0	0	42	9	0	54	Jolie	57
14	Benefits of wellness interventions for persons with chronic and disabling conditions	63	12	0	6	55	10	0	66	Confiance	66
15	Identifying and understanding the factors that influence the functioning of integrated healthcare systems in the NHS	12	6	2	10	58	11	0	61	Confiance	61
16	Bodily maps of musical sensations across cultures	46	0	0	0	49	68	0	29	Surprise	68
17	The Cosmic History of Star Formation	41	0	0	3	64	73	0	24	Surprise	73
18	Resilience in families of children with developmental disabilities	34	43	4	18	46	7	0	72	Confiance	72
19	The evolution and future of research on Nature-based Solutions to address societal challenges	31	14	5	12	76	18	0	74	Anticipation	76
20	Riding the waves of family firm internationalization	24	2	0	8	39	16	0	28	Anticipation	39
21	Citizen scientists report global rapid reductions in the visibility of stars from 2011 to 2022	6	62	27	44	22	57	8	10	Tristesse	62
22	Blind trust? The importance and interplay of parasocial relationships and advertising disclosures in explaining influencers' persuasive effects on their followers	2	9	37	28	54	48	42	7	Anticipation	54
23	Technology As a Double Edged Sword: A Promise Yet to be Fulfilled or a Vehicle for Cheating?	8	10	55	31	59	34	47	4	Anticipation	59
25	Should I stay or should I go? A dependence model of breakups	7	69	14	46	57	16	4	5	Tristesse	69
26	The grass is greener, but why? Evidence of employees' perceived sector mismatch from the US, New Zealand, and Taiwan	4	41	21	17	58	37	6	14	Anticipation	58
27	Ignoring the elephant in the room? Assessing the impact of the European Union on the Development Assistance Committee's role in international development policy	0	19	47	12	52	33	21	11	Anticipation	52
28	Too much of a good thing? Visitor Congestion Management Issues for Popular World Heritage Tourist Attractions	3	28	42	15	49	27	18	9	Anticipation	49
29	Pipe up or clam up: the double-edged sword effect of perceived overqualification on employee voice behavior.	2	24	38	21	61	29	13	18	Anticipation	61

FIGURE C.1 – Résultats ChatGPT

	Jolie	Tristesse	Colère	Peur	Anticipation	Surprise	Dégout	Confiance	Emotion dominante	Intensité dominante	Commentaire
The distress caused by environmental change in the Falkland Islands	2	71	24	41	9	11	22	4	Tristesse	71	La détresse environnementale fait surtout ressortir la tristesse, avec une peur marquée et une légère colère liée à la dégradation du milieu.
	44	27	3	29	56	6	2	47	Anticipation	56	Soutenue par une confiance positive face à l'avenir.
	23	2	0	4	49	74	1	38	Surprise	74	La découverte préhistorique suscite surtout la surprise, avec une forte curiosité tournée vers l'exploration et l'explication.
	1	43	61	27	6	4	63	4	Dégout	63	Le harcèlement au travail évoque fortement le dégoût et la colère, sur fond de tristesse liée aux conséquences subies.
The Effect Evoked by Music and the Arts	74	3	0	0	26	21	0	47	Jolie	74	Le titre active clairement la joie, avec une confiance positive et une légère surprise associées à l'expérience artistique.
	31	53	7	34	39	27	6	42	Tristesse	53	La perte et le trauma installent une tristesse nette, mais la résilience introduit aussi de la confiance et un léger élan anticipateur.
	6	76	4	21	14	3	3	24	Tristesse	76	Le deuil conjugal produit une tristesse très dominante, accompagnée d'une inquiétude diffuse et d'une confiance affaiblie.
	2	24	19	56	28	13	9	18	Peur	56	Les coupures de courant font surtout émerger la peur, avec une anticipation des effets concrets sur la santé et la vie collective.
Psychological Adjustment to Spousal Bereavement in Older Adults	4	37	13	61	31	16	8	23	Peur	61	Le registre des catastrophes communautaires active d'abord la peur, puis une tristesse importante face aux dommages psychiques.
	1	78	8	46	7	11	14	7	Tristesse	78	La solitude et le risque de mortalité entraînent une tristesse très forte, doublée d'une peur sensible des conséquences vitales.
	12	13	19	36	61	7	12	44	Anticipation	61	La vulnérabilité alimente l'anticipation, tandis que l'idée d'une politique plus juste renforce une confiance modérée.
	18	5	8	27	58	36	7	39	Anticipation	58	L'usage de l'IA par les soignants fait surtout naître de l'anticipation, avec de la curiosité et une légère prudence face à l'inconnu.
Vegetable Intake and Mental Health in Adults	32	6	0	6	38	22	0	54	Confiance	54	Le lien entre alimentation et santé mentale inspire avant tout la confiance, avec une tonalité globalement positive et constructive.
	47	14	0	7	49	9	0	63	Confiance	63	Les bénéfices des interventions de bien-être font dominer la confiance, accompagnée d'anticipation et d'une joie mesurée.
	7	9	14	12	46	5	8	49	Confiance	49	Le titre analytique évoque surtout la confiance dans la compréhension du système, avec une anticipation tournée vers l'amélioration.
	58	6	0	0	24	44	0	46	Jolie	58	Les sensations musicales et leur dimension interculturelle suscitent principalement la joie, enrichie d'une surprise notable.
Cosmic History of Star Formation	41	6	0	13	56	64	0	39	Surprise	64	L'histoire cosmique de la formation stellaire produit surtout de la surprise, avec une forte anticipation intellectuelle et un émerveillement joyeux.
	27	43	6	24	38	7	4	51	Confiance	51	La résilience familiale place la confiance au centre, tout en laissant paraître une tristesse de fond liée au handicap développemental.
	28	9	16	17	64	14	8	49	Anticipation	64	L'idée d'évolution et de futur donne une forte place à l'anticipation, soutenue par une confiance relative dans les solutions proposées.
	31	6	7	24	57	18	2	43	Anticipation	57	La métaphore du mouvement et de l'internationalisation active l'anticipation, avec une confiance mesurée et une énergie positive.
The waves of family firm internationalization	6	51	26	18	13	67	31	14	Surprise	67	La baisse rapide de visibilité des étoiles combine surprise et tristesse, avec une coloration de dégoût face à l'ampleur du constat.
	7	8	26	27	29	21	41	16	Dégout	41	La notion de blind trust et d'influence persuasive fait ressortir le dégoût, avec méfiance, peur et irritation sous-jacentes.
	16	11	24	31	49	28	37	22	Anticipation	49	L'image du double tranchant nourrit l'anticipation, mais elle s'accompagne d'un dégoût et d'une peur liés aux dérives possibles.
	11	54	13	41	23	16	8	14	Tristesse	54	Le thème de la rupture fait dominer la tristesse, soutenue par la peur de la séparation et de l'incertitude affective.
The elephant in the room? Evidence of employees' perceived sector mismatch from the US, New Zealand, and the European Union on the Development Assistance for International Development policy	16	24	16	11	41	32	17	31	Anticipation	41	Le contraste entre herbe plus verte et mismatch alimente surtout l'anticipation, avec une confiance modérée et une légère surprise.
	4	14	37	16	43	22	29	27	Anticipation	43	L'éléphant dans la pièce fait monter l'anticipation critique, avec de la colère et du dégoût envers ce qui est laissé de côté.
	26	13	22	8	36	24	21	38	Confiance	38	La congestion touristique évoque surtout la confiance dans la gestion, mêlée d'anticipation et d'une irritation modérée.
	8	17	34	26	31	19	26	23	Colère	34	L'effet double tranchant de la surqualification fait émerger surtout la colère, avec une anticipation tendue et une peur modérée.

Joie		Tristesse	Colère	Peur	Anticipation	Surprise	Dégoût	Confiance	Émotion dominante	Intensité dominante	Commentaire	
	2	82	34	67	42	13	21	7	Tristesse	82	La détresse environnementale active surtout la tristesse, avec une peur nette et une fonction anticipation.	
	68	37	4	22	86	2	1	89	Confiance	89	Le vocabulaire d'espoir et d'optimisme fait dominer la confiance et l'anticipation, avec une forte tristesse négative.	
	14	0	0	3	38	87	0	44	Surprise	87	La découverte préhistorique suscite avant tout la surprise, accompagnée de tristesse et d'insécurité négative.	
	0	63	88	72	21	18	79	3	Colère	88	Le harcèlement au travail évoque fortement la colère, la peur et le dégoût face aux conséquences négatives.	
	92	2	0	1	44	33	0	81	Joie	92	Le registre positif lié à la musique et aux arts fait dominer la joie, soutenue par une forte confiance.	
	54	78	12	38	62	24	7	87	Confiance	87	La perte reste très présente, mais le lexique de résilience et d'épanouissement confère nettement la confiance.	
	1	96	7	53	14	3	2	19	Tristesse	96	Le deuil conjugal provoque une tristesse extrême, avec une peur secondaire liée à la mort et à la solitude.	
	0	28	42	68	57	17	11	14	Peur	68	Les coupures de courant liées à la santé activent surtout la peur et l'anticipation des effets négatifs.	
	9	74	23	79	48	12	14	62	Peur	79	Le thème des catastrophes communautaires combine fortement peur et tristesse, avec une attente autour de la réponse apportée.	
	0	92	13	82	27	6	18	4	Tristesse	92	La solitude associée à la mortalité fait dominer une tristesse très forte, accompagnée d'une peur élevée.	
	22	29	33	54	88	14	19	57	Anticipation	88	La vulnérabilité climatique nourrit l'anticipation et la peur, mais la perspective de transition maintient un peu de confiance.	
	32	6	13	38	77	42	7	53	Anticipation	77	L'usage de l'IA en santé inspire surtout l'anticipation, avec de la confiance mais aussi une légère tristesse.	
	58	4	1	3	64	11	1	78	Confiance	78	Le lien entre alimentation saine et santé mentale évoque surtout confiance, anticipation et amélioration du bien-être.	
	63	33	2	17	74	8	1	86	Confiance	86	Les bénéfices des interventions de bien-être font dominer la confiance, l'anticipation et une joie modérée.	
	17	12	28	24	69	7	12	64	Anticipation	69	Le fonctionnement des systèmes intégrés dans le NHS produit une lecture négative marquée par l'anticipation et la confiance.	
	73	4	1	3	49	78	2	67	Surprise	78	Les sensations musicales à travers les cultures provoquent surprise et joie, avec une forte tristesse négative.	
	42	2	0	1	58	83	0	52	Surprise	83	L'histoire cosmique de la formation des étoiles évoque l'émerveillement, avec une confiance abstraitement positive.	
	47	68	11	32	59	9	2	82	Confiance	82	La résilience familiale face au handicap mêle tristesse de fond et forte confiance dans les capacités d'adaptation.	
	53	14	7	28	87	18	3	83	Anticipation	87	Les solutions fondées sur la nature activent surtout anticipation et confiance, avec une confiance abstraitement positive.	
	48	4	6	32	78	23	2	69	Anticipation	78	L'internationalisation des entreprises familiales suggère surtout anticipation et confiance, avec une part de doute négatif.	
	1	83	37	58	42	64	22	8	Tristesse	83	La baisse de visibilité des étoiles fait dominer la tristesse, amplifiée par la tristesse due à l'absence de l'abandonner.	
	12	17	32	48	53	58	64	18	Dégoût	64	Le thème de la confiance aveugle et de l'influence publicitaire évoque dégoût, avec une part de la manipulation.	
	23	18	47	58	72	43	52	33	Anticipation	72	La technologie présentée comme ambivalente active surtout l'anticipation, avec une détresse due à la peur d'une dérive négative.	
	7	77	38	72	89	23	18	14	Anticipation	89	Le modèle de rupture est dominé par l'anticipation, avec une forte tristesse et une peur marquée de la récession.	
	28	58	42	34	63	54	37	22	Anticipation	63	L'idée d'un ailleurs meilleur mêlée au mismatch professionnel combine anticipation et tristesse.	
	3	22	83	47	59	78	67	12	Colère	83	L'éléphant dans la pièce suscite colère, surprise et dégoût, avec une attente quant aux effets négatifs.	
	22	43	68	37	48	53	73	18	Dégoût	73	La congestion touristique d'un lieu très prisé évoque surtout dégoût et colère, avec une attente devant les effets.	
	8	53	72	78	64	33	47	21	Peur	78	La surqualification et la prise de parole au travail activent peur, colère et anticipation face à l'insécurité.	

	Jolie	Tristesse	Colère	Peur	Anticipation	Surprise	Dégout	Confiance	Émotion dominante	Intensité dominante	Commentaire
	8	85	22	35	12	18	15	3	Tristesse	85	Distress domine (tristesse), colère/peur liées à l'environnement.
	72	12	0	7	55	8	0	63	Jolie	72	Hope et optimism génèrent joie/confiance, anticipation pour l'avenir.
	23	0	0	0	33	67	0	12	Surprise	67	Surprise archéologique, curiosité (anticipation).
	0	63	78	45	6	12	33	0	Colère	78	Bullying et adverse = colère/tristesse, peur des conséquences.
	88	0	0	0	22	19	0	25	Jolie	88	Positive affect = joie pure, légère surprise.
city to thrive after extremely aversive	42	75	7	23	35	14	6	52	Tristesse	75	Loss/trauma = tristesse, mais resilience/thrive = joie/confiance.
	14	82	6	28	11	9	0	33	Tristesse	82	Bereavement = tristesse dominante, peur de l'isolement.
	7	43	37	52	24	22	12	11	Peur	52	Peur des conséquences, colère contre les défaillances, surprise.
	12	65	24	48	32	17	10	22	Tristesse	65	Disasters = tristesse/peur, anticipation des réponses.
	3	92	11	58	5	6	7	0	Tristesse	92	Mortality + loneliness = tristesse/peur extrêmes.
re policy	34	22	18	43	52	13	6	45	Anticipation	52	Vulnerability = peur, mais just/effective = espoir (joie/confiance).
	23	6	12	25	53	42	7	34	Anticipation	53	AI = anticipation/surprise, légère peur de l'inconnu.
itions	54	6	0	0	23	11	0	62	Confiance	62	Lien santé/bien-être = joie/confiance.
	73	11	0	0	44	16	0	65	Jolie	73	Benefits = joie/confiance, anticipation des résultats.
egrated healthcare systems in the NHS	13	7	6	12	54	23	0	43	Anticipation	54	Recherche analytique = anticipation/confiance, légère surprise.
	62	0	0	0	25	44	0	23	Jolie	62	Musical sensations = joie, surprise interculturelle.
	33	0	0	0	42	55	0	14	Surprise	55	Émerveillement (joie) et surprise cosmique.
	53	32	6	14	43	12	0	64	Confiance	64	Resilience = joie/confiance, mais tristesse sous-jacente.
ietal challenges	44	6	7	13	62	22	0	53	Anticipation	62	Future = anticipation, solutions = espoir (joie/confiance).
	38	15	10	28	55	20	0	30	Anticipation	55	Métaphore dynamique = anticipation/joie, mais peur du risque.
11 to 2022	12	53	34	25	22	33	14	13	Tristesse	53	Reductions = tristesse/colère, surprise face à l'ampleur.
rtising disclosures in explaining	11	7	23	44	22	35	22	3	Peur	44	Blind trust = méfiance (peur/dégout), surprise des effets.
e for Cheating?	12	13	35	47	33	43	24	8	Peur	47	Double-edged = ambivalence (peur/colère), surprise des implications.
atch from the US, New Zealand, and	12	72	33	45	22	21	13	6	Tristesse	72	Breakups = tristesse/peur, colère face à l'incertitude.
n on the Development Assistance	18	42	32	23	24	22	12	14	Tristesse	42	Greener = envie (joie), mais mismatch = tristesse/colère.
World Heritage Tourist Attractions	7	12	63	24	34	45	23	6	Colère	63	Elephant in the room = colère/dégout, surprise de l'impact.
on on employee voice behavior	13	8	44	35	23	32	22	12	Colère	44	Too much = frustration (colère), peur des conséquences.
	6	22	53	48	23	33	24	7	Colère	53	Double-edged = colère/peur, dégoût face à l'injustice.

Annexe D

Graphiques radar par titre

Cette annexe présente les graphiques radar réalisés pour chacun des titres du corpus. Pour chaque titre, le profil émotionnel moyen humain est comparé aux profils produits par Gemini, ChatGPT, Claude et Piccolo. Ces visualisations complètent les analyses quantitatives du chapitre des résultats et permettent d'observer plus finement les convergences et divergences entre les modèles et la base humaine de référence.

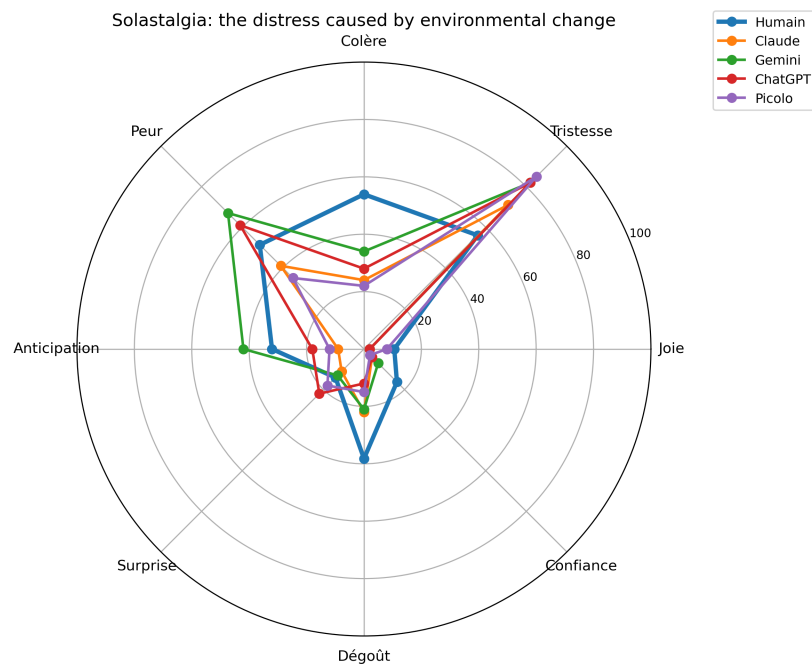
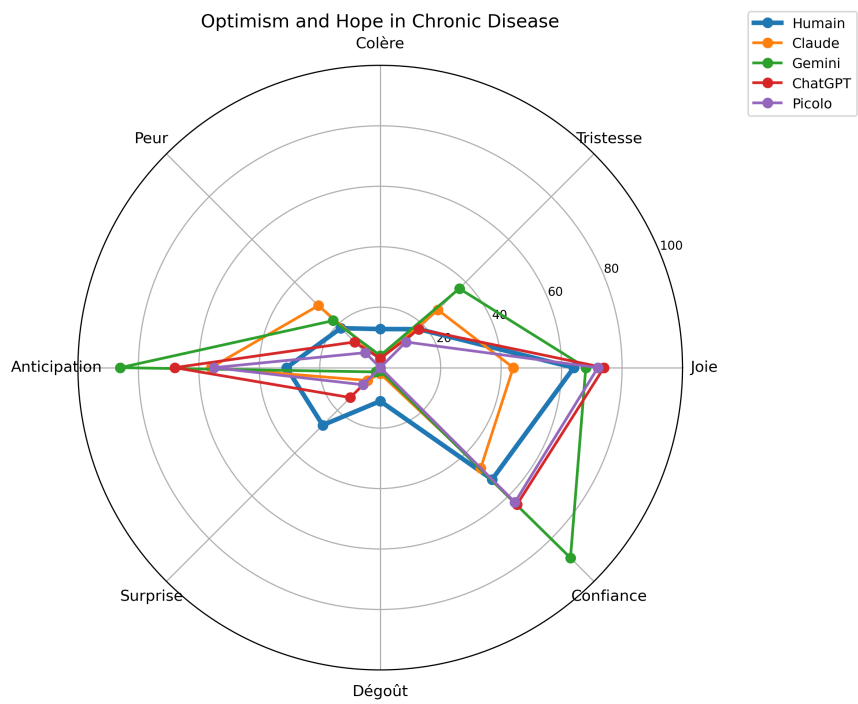
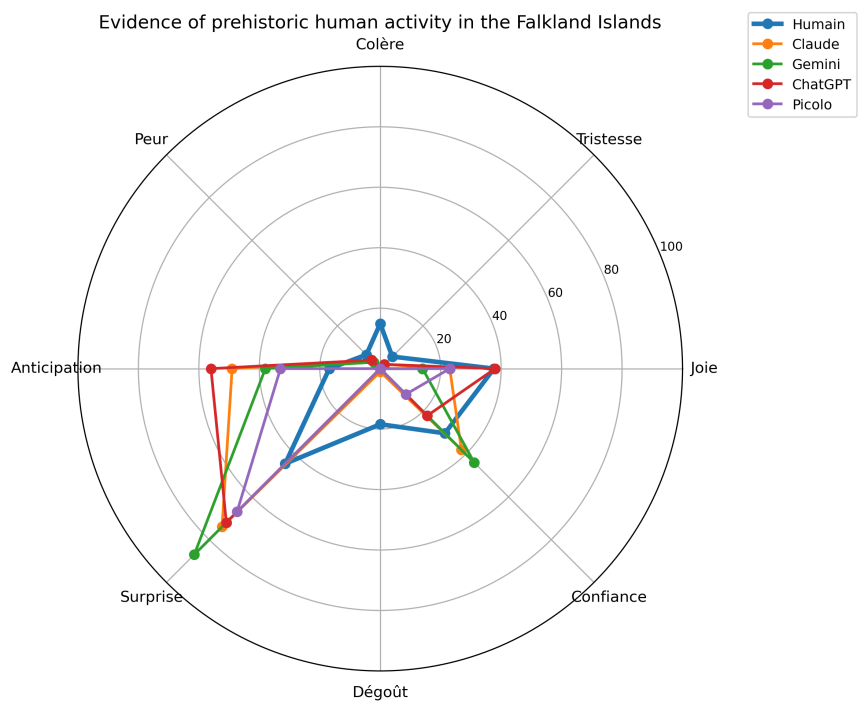


FIGURE D.1 – Graphique radar du titre 1



Titre 2

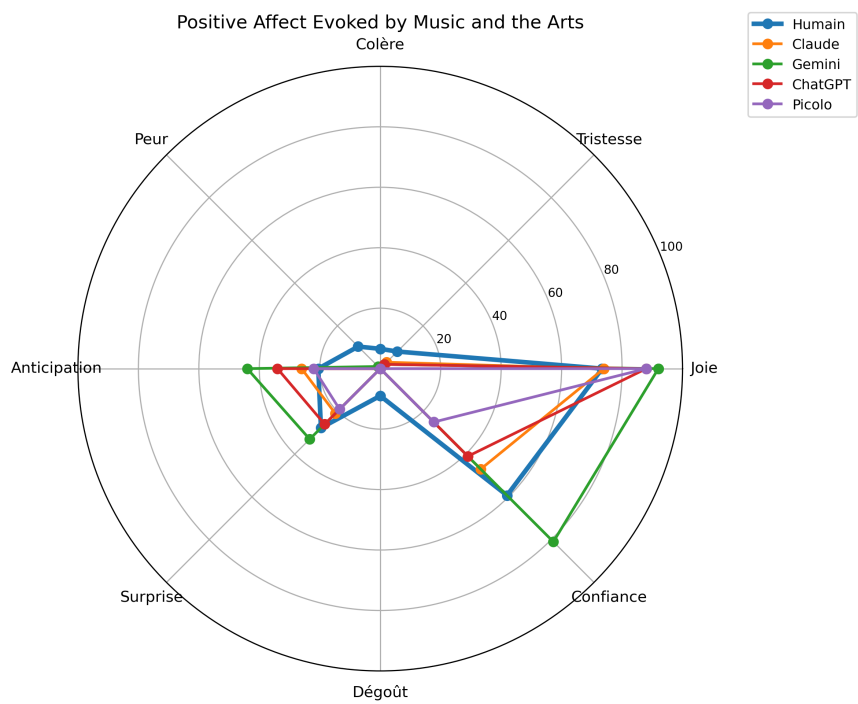


Titre 3

FIGURE D.2 – Graphiques radar des titres 2 et 3

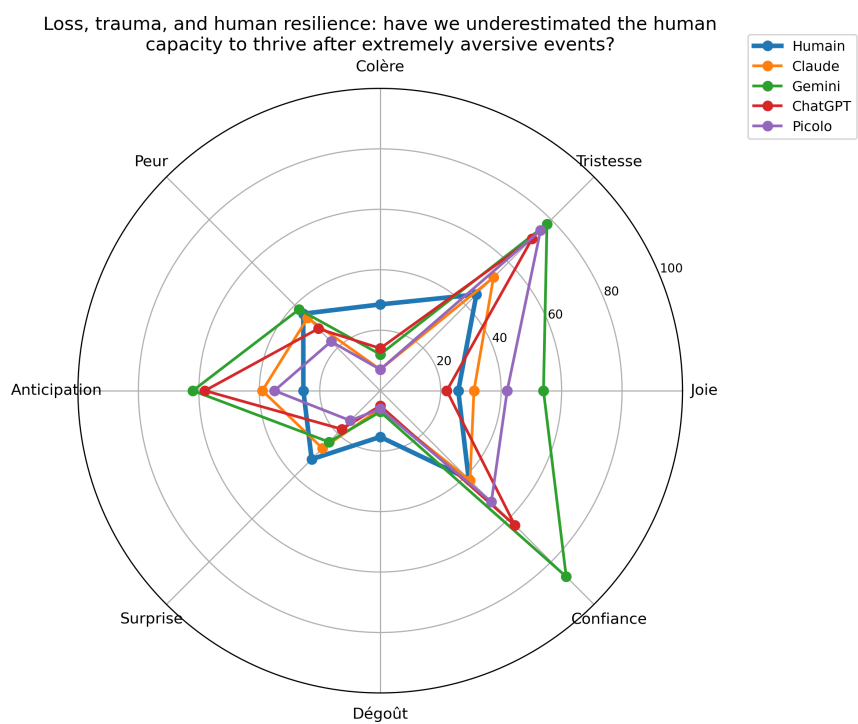


Titre 4

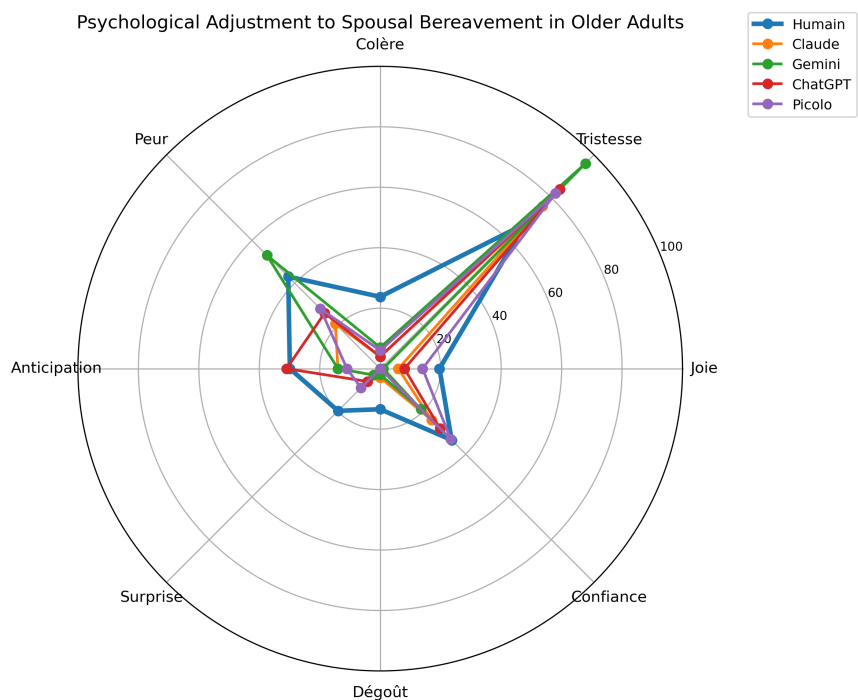


Titre 5

FIGURE D.3 – Graphiques radar des titres 4 et 5

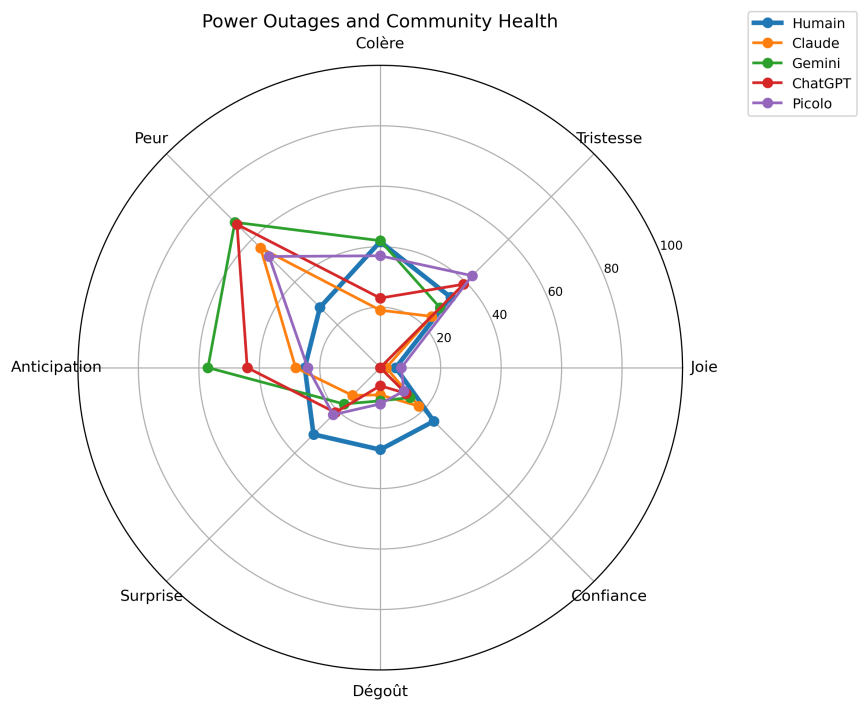


Titre 6

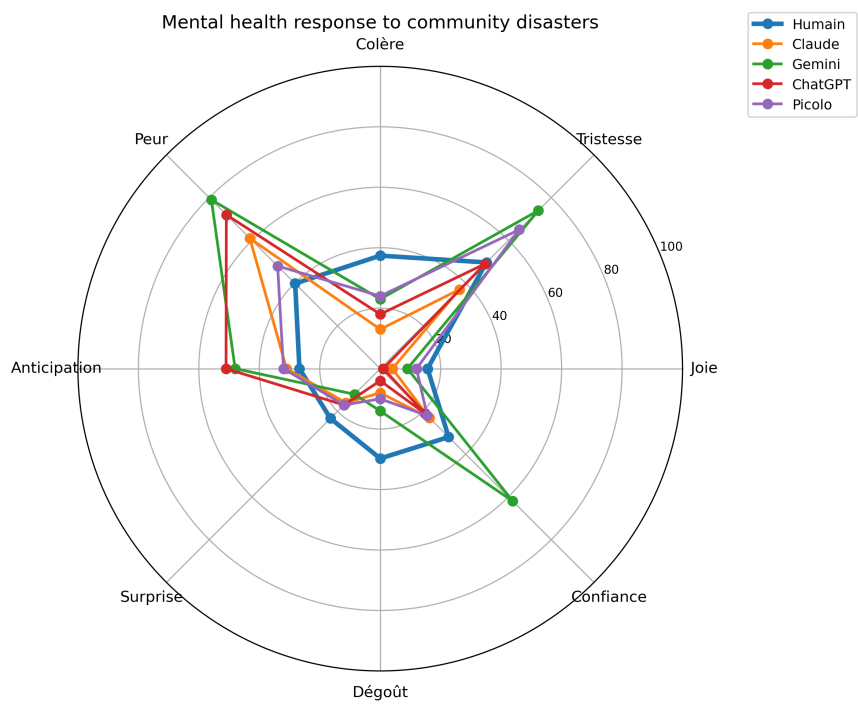


Titre 7

FIGURE D.4 – Graphiques radar des titres 6 et 7

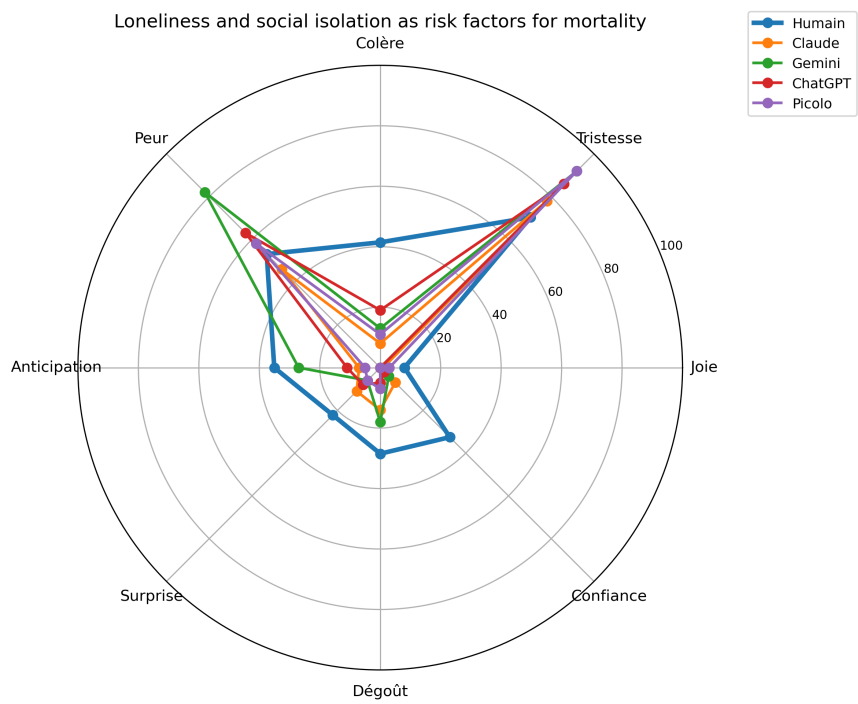


Titre 8

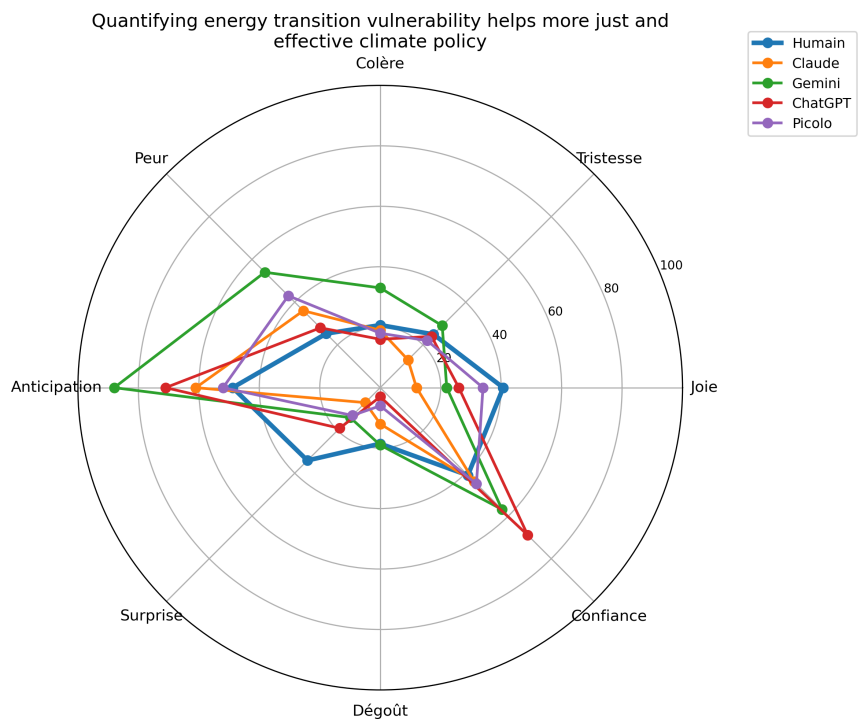


Titre 9

FIGURE D.5 – Graphiques radar des titres 8 et 9

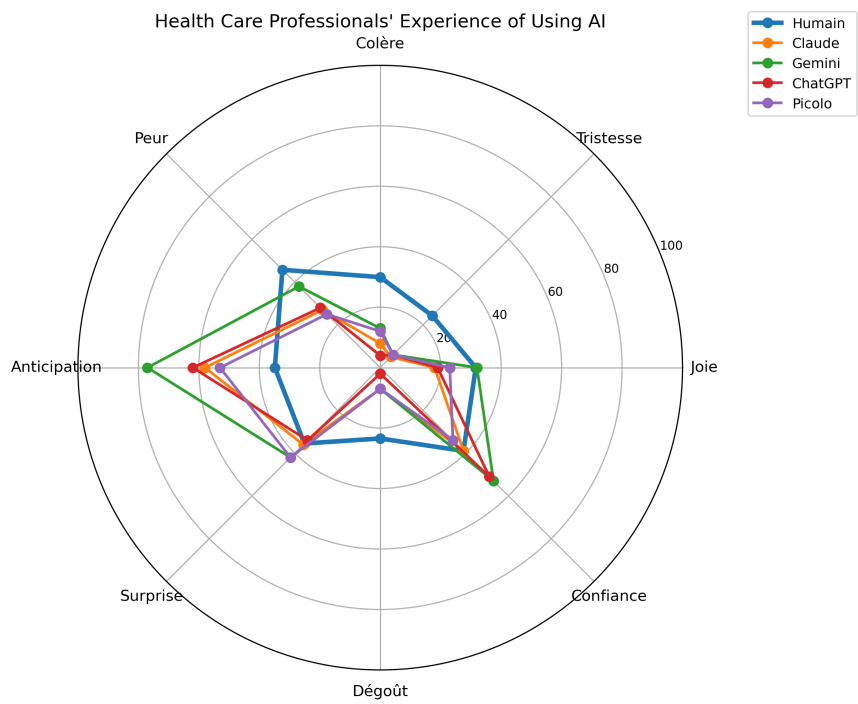


Titre 10

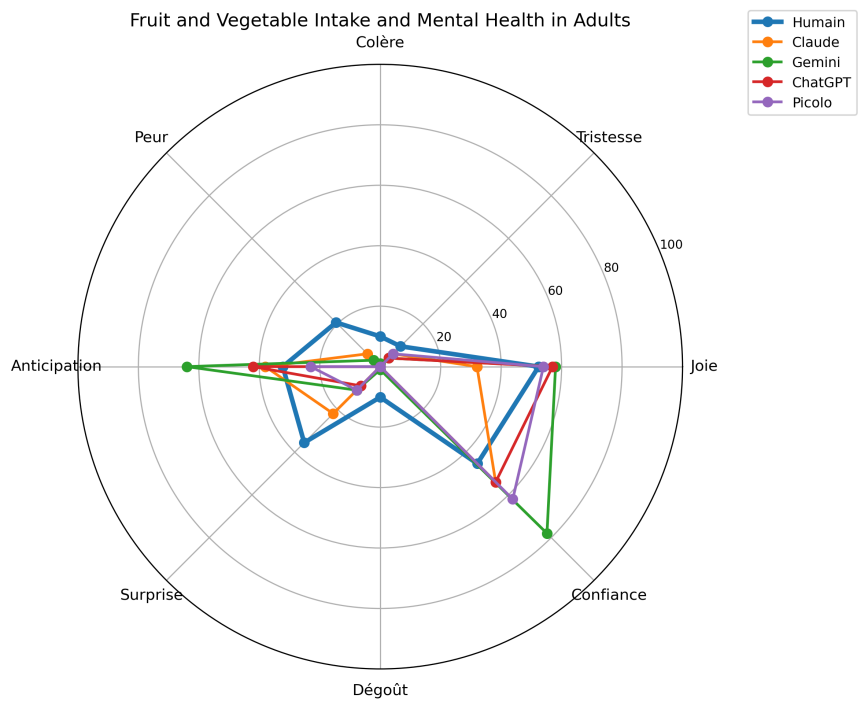


Titre 11

FIGURE D.6 – Graphiques radar des titres 10 et 11

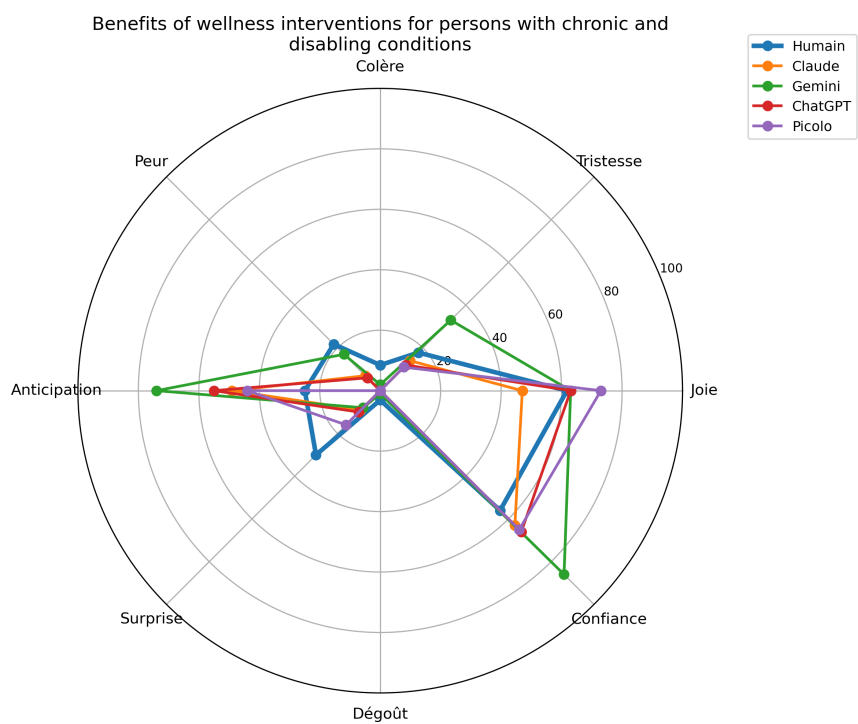


Titre 12

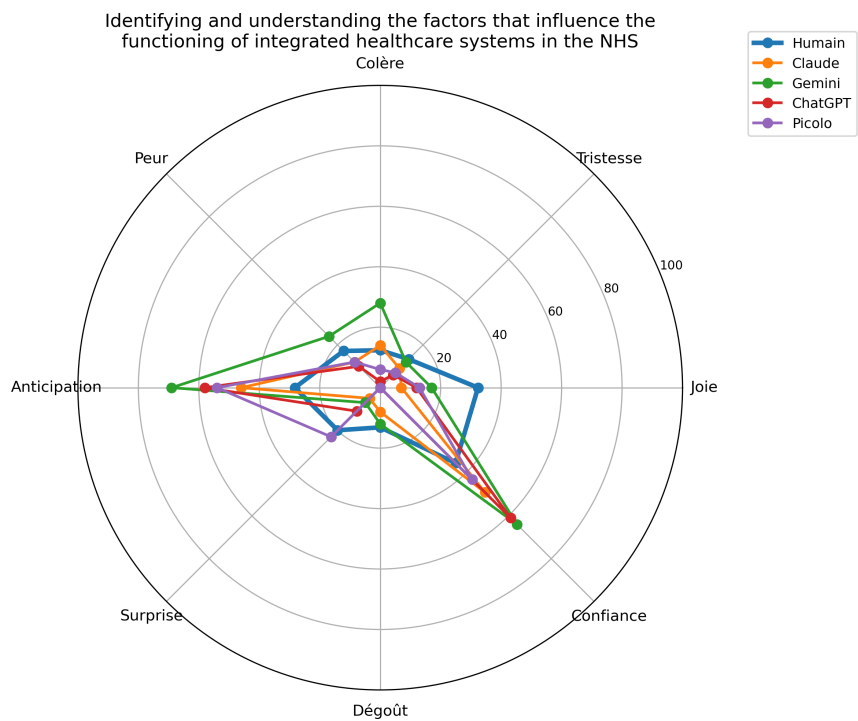


Titre 13

FIGURE D.7 – Graphiques radar des titres 12 et 13

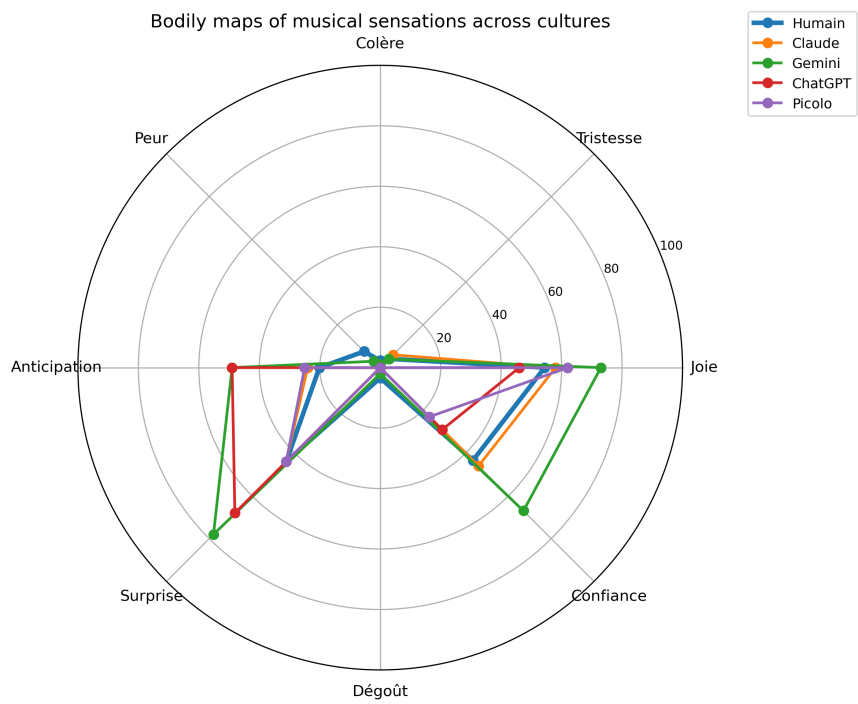


Titre 14

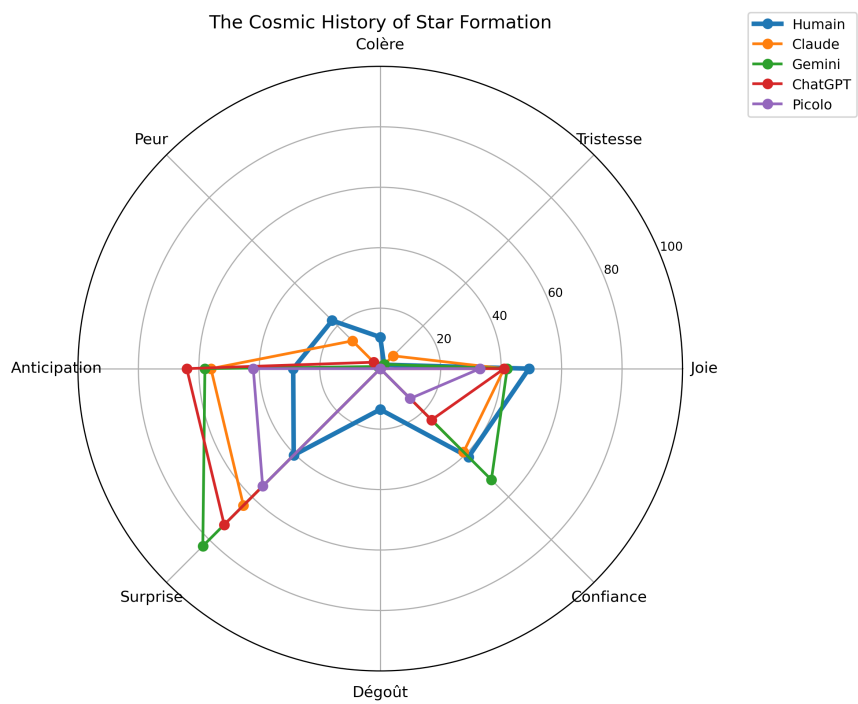


Titre 15

FIGURE D.8 – Graphiques radar des titres 14 et 15

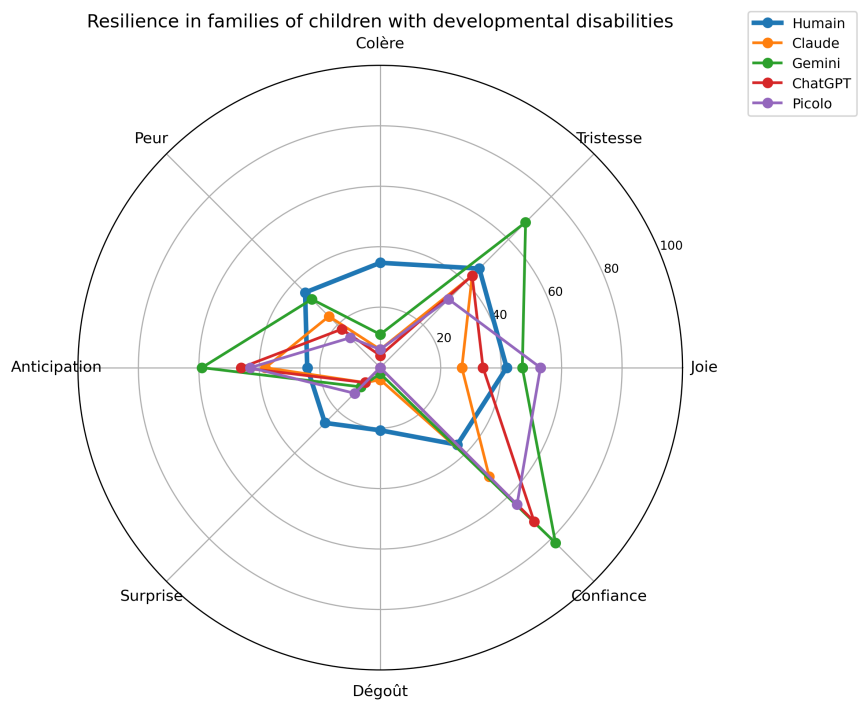


Titre 16

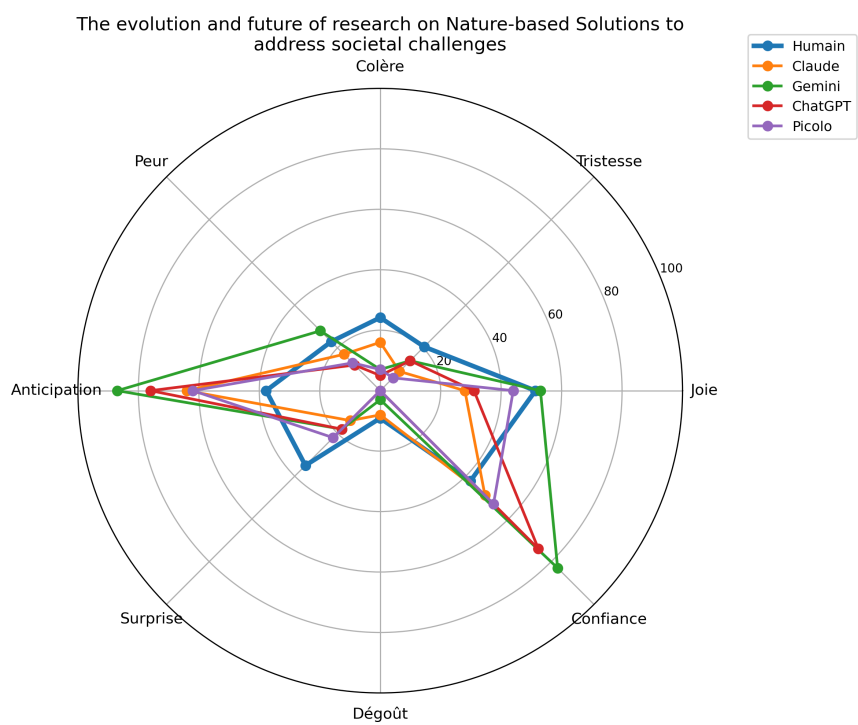


Titre 17

FIGURE D.9 – Graphiques radar des titres 16 et 17

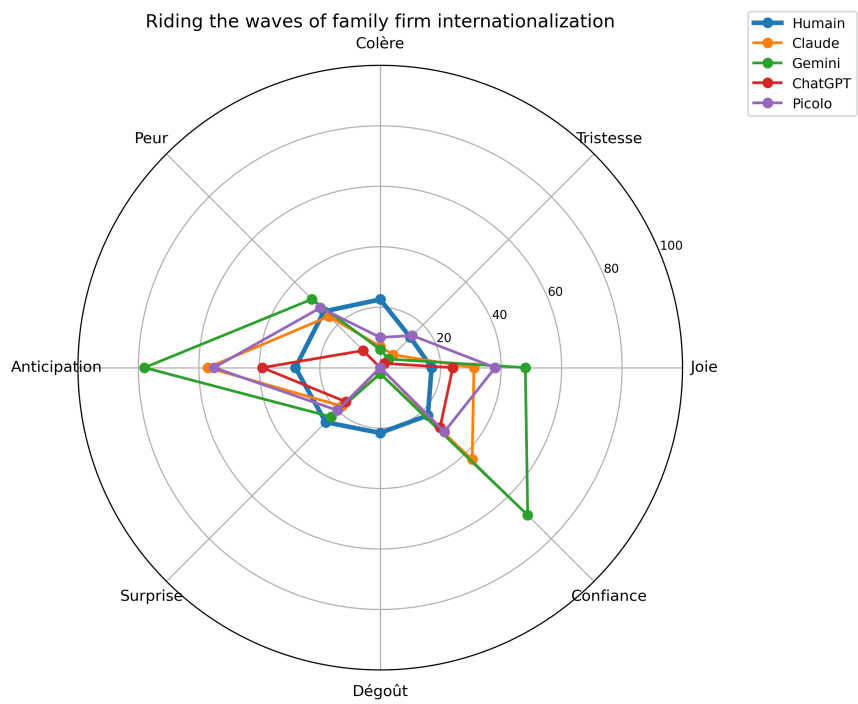


Titre 18

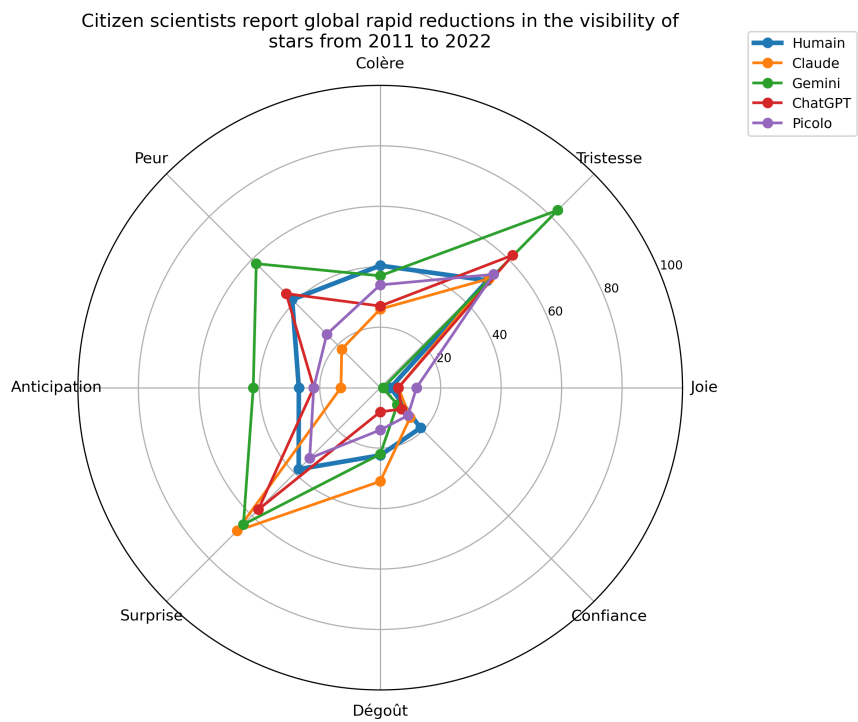


Titre 19

FIGURE D.10 – Graphiques radar des titres 18 et 19



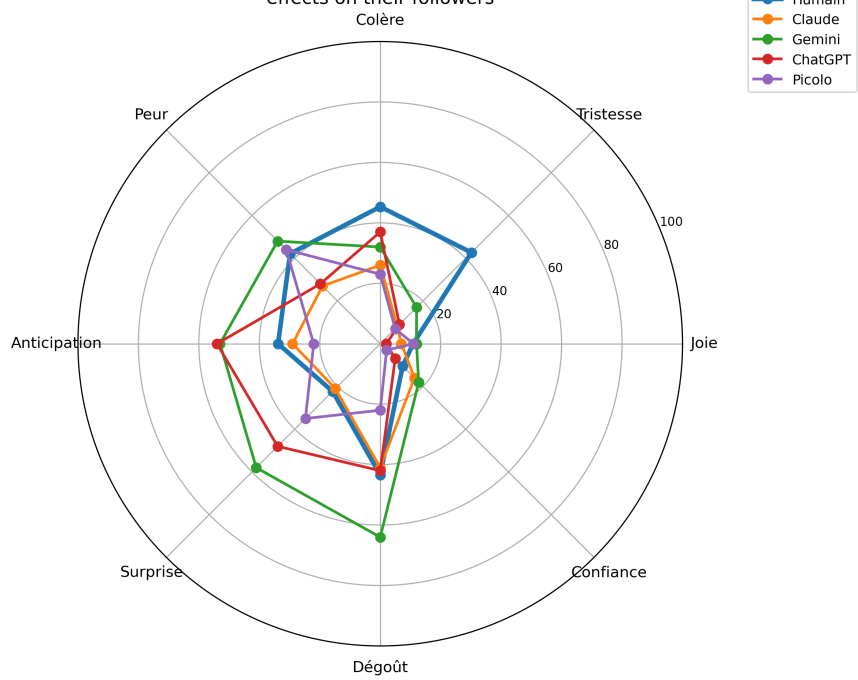
Titre 20



Titre 21

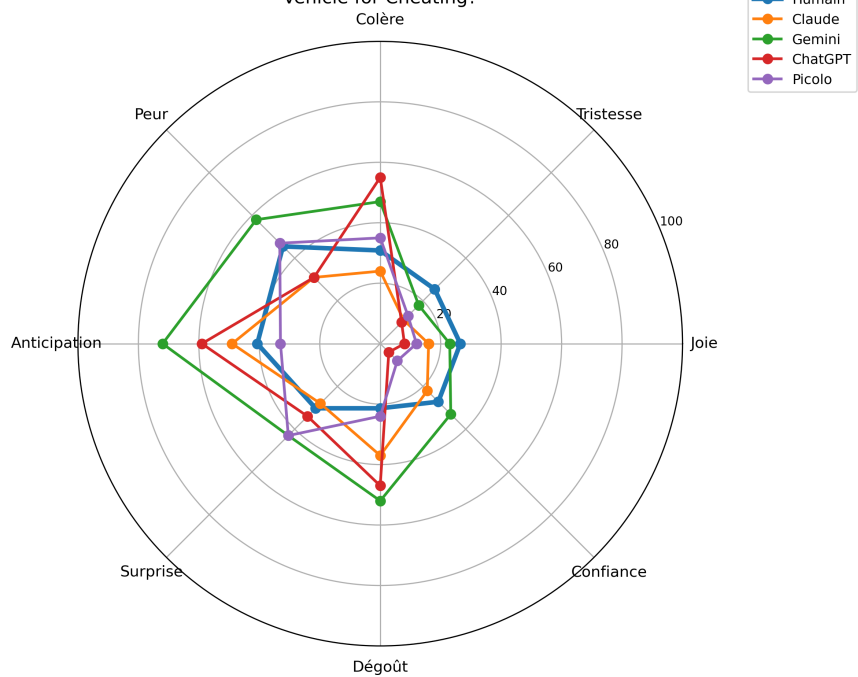
FIGURE D.11 – Graphiques radar des titres 20 et 21

Blind trust? The importance and interplay of parasocial relationships
and advertising disclosures in explaining influencers' persuasive
effects on their followers



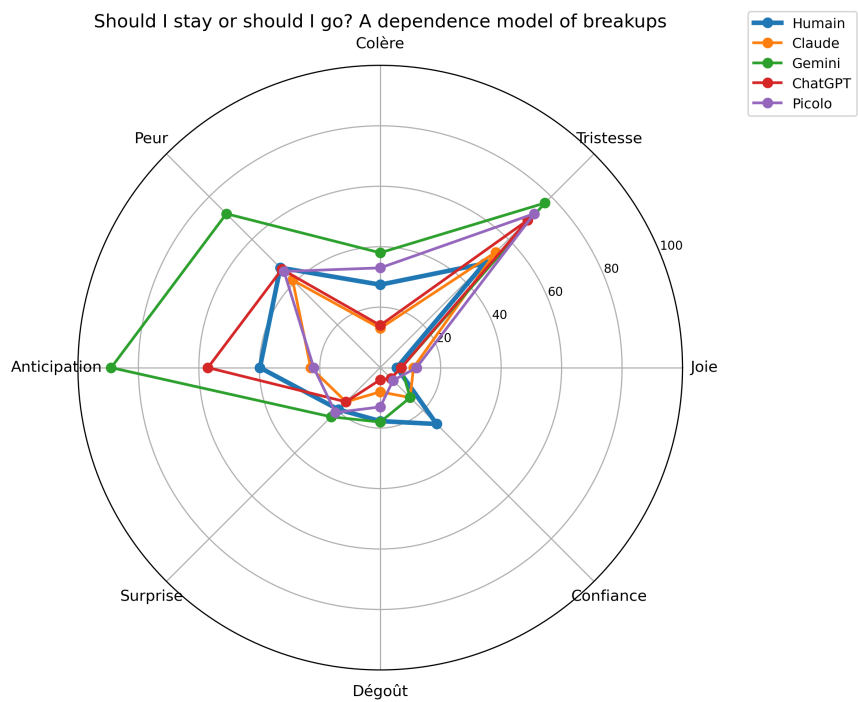
Titre 22

Technology As a Double Edged Sword: A Promise Yet to be Fulfilled or a
Vehicle for Cheating?

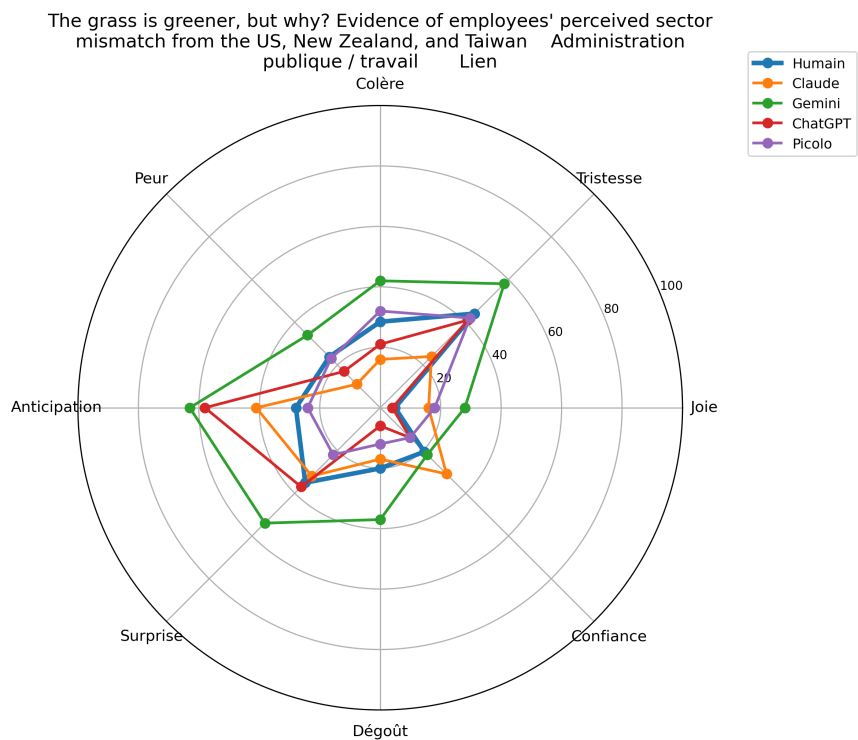


Titre 23

FIGURE D.12 – Graphiques radar des titres 22 et 23



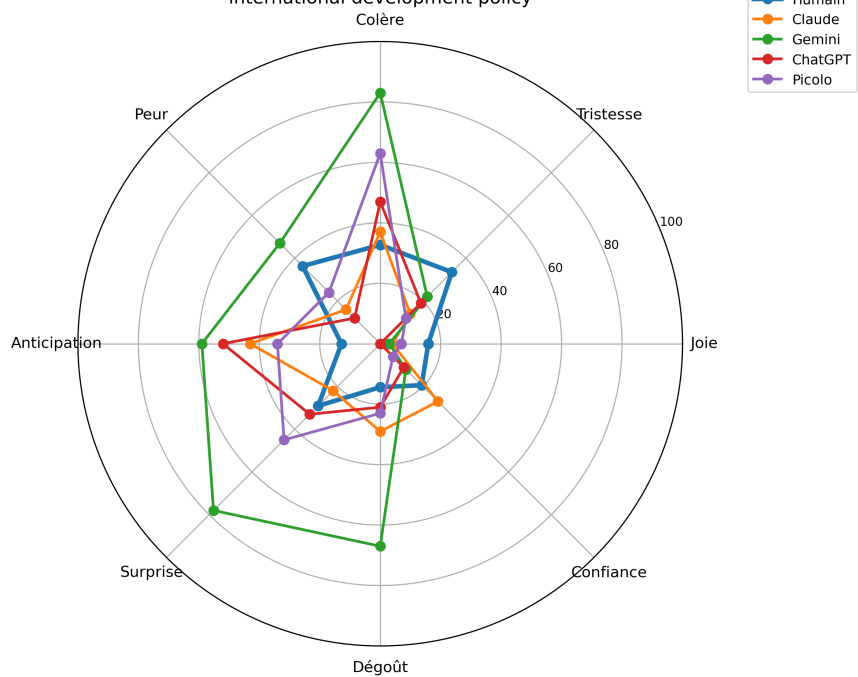
Titre 24



Titre 25

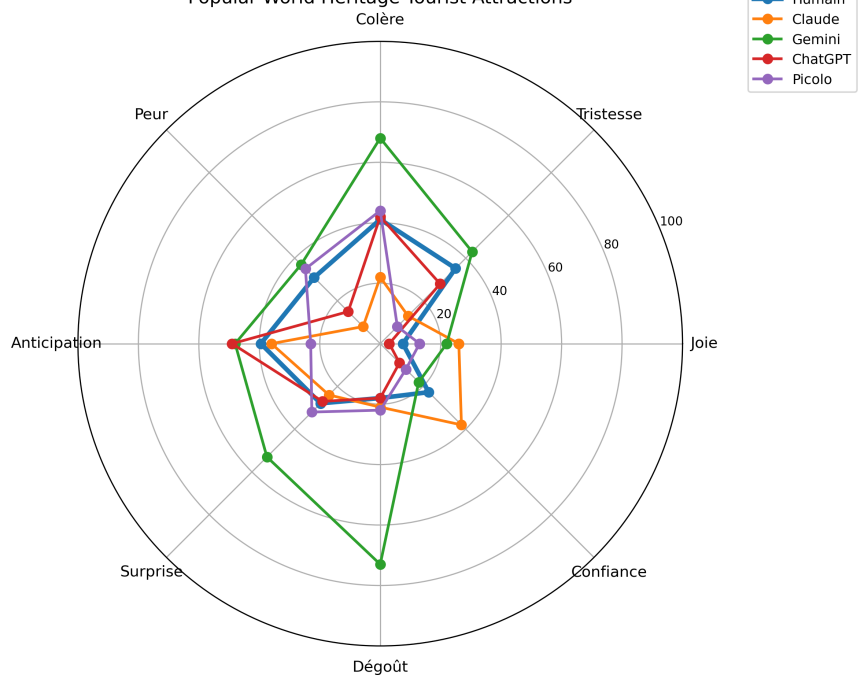
FIGURE D.13 – Graphiques radar des titres 24 et 25

Ignoring the elephant in the room? Assessing the impact of the European Union on the Development Assistance Committee's role in international development policy



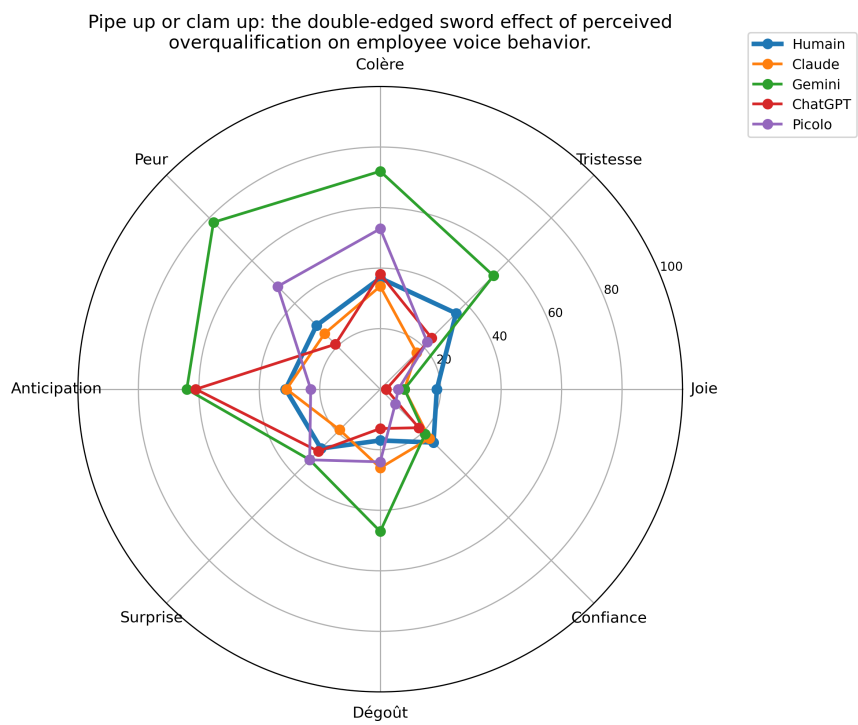
Titre 26

Too much of a good thing? Visitor Congestion Management Issues for Popular World Heritage Tourist Attractions



Titre 27

FIGURE D.14 – Graphiques radar des titres 26 et 27



Titre 28

FIGURE D.15 – Graphique radar du titre 28

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve

Boulevard Emile Devreux 6, 6000 Charleroi, Belgique

Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm