

**L'intelligence artificielle peut-elle
permettre d'améliorer l'évaluation
ESG ? Une analyse comparative des
modèles LLM**

Mémoire de recherche en Master II
Mention Monnaie Banque Finance Assurance, Economie
et Gestion des Risques Financiers

Présenté par :
Léonardo OSORIO

Sous la direction de :
Monsieur Vincent IEHLÉ

UFR Droit, Economie, Science Politique
Université de Rouen Normandie
2024-2025



Déclaration sur l'Honneur

Je, soussigné(e) Léonardo OSÓRIO PEREIRA
étudiant(e) en Master MBFA, Economie et Gestion des Risques Financiers, rédigeant un mémoire / projet /
rapport sous la direction de Monsieur Vincent IEHLÉ à l'Université de Rouen,

certifie que ce document est le fruit d'un travail original et que toutes les sources utilisées ont été clairement indiquées.

Je certifie de plus que je n'ai ni copié ni utilisé des idées, des formulations ou tout autre document tirés d'un ouvrage, article ou mémoire ou projet, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations sont expressément signalées entre guillemets.

Le non-respect de ces dispositions me rend passible de poursuites devant la section disciplinaire de l'Université et les tribunaux de la République française pour plagiat universitaire.

Fait à Rouen le 24/08/2025

Signature



REMERCIEMENTS

Au terme de ce travail de recherche, je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué à sa réalisation.

Je remercie tout d'abord mon directeur de mémoire, Monsieur Vincent Iehlé, pour sa bienveillance, sa disponibilité et ses conseils avisés. Sa rigueur et son expertise ont été un guide précieux tout au long de ce projet.

Je souhaite également exprimer ma reconnaissance à Monsieur Nicolas Abad pour l'évaluation de mon mémoire, ainsi qu'au responsable de ma formation, Monsieur Cyril Dell'Eva, et à l'ensemble des professeurs du Master II Monnaie Banque Finance Assurance, Economie et Gestion des Risques Financiers de l'Université de Rouen Normandie, pour la qualité de leur enseignement.

Enfin, je remercie l'Université de Rouen Normandie et le service de scolarité pour leur soutien administratif tout au long de mon parcours.

Résumé

L'évaluation des rapports de responsabilité sociétale des entreprises (RSE) est un défi majeur de la finance durable car elle implique d'analyser des performances sociales et environnementales variées, hétérogènes et difficiles à comparer. Ce mémoire évalue la capacité des modèles de langage de grande taille (LLM) à automatiser et fiabiliser cette tâche. Une étude comparative de cinq LLM de premier plan a été menée sur un corpus de 18 rapports RSE. La performance des modèles dans l'extraction et la classification d'informations ESG a été évaluée via la méthode « LLM-as-a-judge » et les résultats ont été classés à l'aide du modèle statistique de Bradley-Terry. Les résultats démontrent la nette supériorité du modèle Gemini-2.5-Pro-Preview-05-06 avec un taux de victoire de 83,3 % et un score de Bradley-Terry de 19.4886, principalement grâce à sa capacité à gérer de longs contextes et à son intégration de la recherche en temps réel.

SOMMAIRE

Introduction.....	1
I. Revue de littérature.....	3
1. La finance durable et l'évaluation ESG.....	3
1.1 Définition et évolution de la finance durable.....	3
1.2 Les instruments, les acteurs et le cadre réglementaire de la finance durable.....	4
1.2.1 Les instruments de la finance durable.....	4
1.2.2 Les acteurs de la finance durable.....	4
1.2.3 Le cadre réglementaire de la finance durable.....	5
1.3 Les défis de la finance durable.....	5
2. L'Intelligence Artificielle au service de la Finance.....	6
2.1 Fondements de l'IA et des LLM.....	6
2.2 Usages actuels de l'intelligence artificielle et des LLM dans le secteur financier...	6
2.3 Limites et enjeux éthiques de l'IA.....	7
3. L'IA au service de l'évaluation ESG.....	8
3.1 L'intelligence artificielle dans la collecte des données ESG.....	8
3.2 L'intelligence artificielle dans l'analyse de données ESG.....	9
3.3 L'intelligence artificielle dans le scoring des données ESG.....	11
II. Développement.....	12
4. LLM-as-a-judge.....	12
4.1 Arena Score.....	12
4.2 Biais.....	13
5. Méthodologie.....	14
5.1 Sélection des éléments à comparer.....	14
5.2 Choix des modèles.....	14
5.3 Prompts.....	15
5.4 Illustration du processus.....	16
6. Résultats.....	20
6.1 Classements.....	20
6.1.1 Système de classement Elo.....	21
6.1.2 Glicko.....	22
6.1.3 TrueSkill.....	22
6.1.4 Modèle de Bradley-Terry.....	22
6.2 Discussions.....	24
6.2.1 Pourquoi Gemini domine-t-il ?.....	24
6.2.2 Implications pour les parties prenantes.....	24
6.2.3 Limites.....	25
6.2.4 Travaux futurs.....	25
6.2.5 Éthique.....	26
Conclusion.....	26
Bibliographie.....	27

Introduction

La finance durable est en croissance depuis plusieurs années atteignant plus de 7 000 milliards de dollars d'investissements en 2023, soit une augmentation de 20 % par rapport à l'année précédente (UNCTAD, 2024). Les rapports de responsabilité sociétale des entreprises (RSE) sont devenus des outils de communication essentiels permettant aux entreprises de présenter leurs engagements et résultats extra-financiers.

Toutefois, l'évaluation de ces documents peut s'avérer être un exercice complexe, subjectif et chronophage nécessitant l'analyse d'une quantité importante d'informations. Récemment, l'intelligence artificielle et en particulier les modèles de langage de grande taille (LLM) sont apparus comme une solution prometteuse pour améliorer l'efficacité et la cohérence des analyses ESG. Ces modèles ont montré leur capacité à comprendre, résumer, classer des textes avec un jugement très proche de celui des humains.

L'application des LLM à l'évaluation de rapport ESG demeure récente et expérimentale. Malgré un intérêt croissant dans la littérature, de nombreuses questions fondamentales restent posées et peu étudiées :

- Les LLM peuvent-ils capter les nuances extra-financières des rapports RSE ?
- Peuvent-ils produire des évaluations comparables d'un rapport à l'autre ?
- Quels sont aujourd'hui les LLM les plus performants pour l'évaluation ESG ?

Ce mémoire étudie ces problématiques à travers une étude comparative de cinq LLM appliqués à l'analyse d'un corpus de 18 rapports RSE d'entreprises issues de secteurs variés. L'objectif est d'évaluer la capacité de ces modèles à extraire, structurer et évaluer l'information ESG de manière cohérente et fiable et d'identifier lequel de ces modèles est le plus performant.

La première partie de ce mémoire est une revue de littérature permettant d'offrir une vue d'ensemble des enjeux actuels de l'intelligence artificielle appliquée à la finance durable. Nous évoquons le contexte de la finance durable et plusieurs de ses défis tels que le risque de greenwashing, le manque d'uniformisation des rapports RSE ou encore les divergences de notation. Nous mettons en lumière l'essor récent de l'intelligence artificielle dans le paysage financier et plus spécifiquement des modèles d'analyse textuelle. L'intelligence artificielle est déjà fortement utilisée pour automatiser la collecte et l'analyse ESG, notamment pour identifier des incohérences dans certains discours d'entreprises et pour l'analyse de rapports RSE. Des modèles de NLP spécialisés tels que FinBERT ou ClimateBERT ont démontré leur efficacité sur des tâches spécifiques à l'évaluation extra-financière. Toutefois, les études comparatives rigoureuses sur les LLM les plus récents sont peu nombreuses. Il est aujourd'hui difficile pour un investisseur de savoir quel LLM utiliser, pourquoi et comment les choisir.

La deuxième partie de ce mémoire développe une méthodologie pour comparer plusieurs des meilleurs LLM actuels dans le cadre de l'évaluation des rapports ESG. La méthodologie choisie repose sur une confrontation directe des modèles via une approche nommée « LLM-as-a-judge » pour évaluer la capacité de ces modèles à extraire et structurer l'information ESG. Les résultats sont ensuite analysés à l'aide de systèmes de classement statistiques, notamment le modèle de Bradley-Terry, pour établir un classement de performance fiable.

Les résultats de cette étude indiquent une nette domination du modèle Gemini-2.5-Pro-Preview-05-06, qui s'est imposé comme le plus performant avec un taux de victoire de 83,3 % et un score de Bradley-Terry de 19.4886, loin devant ses concurrents. Cette supériorité s'explique principalement par sa capacité à traiter de très longues séquences textuelles et par l'intégration native de la recherche Google, lui permettant de vérifier et de compléter l'identification des références GRI.

Pour les parties prenantes de l'écosystème ESG, ces résultats démontrent que les LLM permettent une analyse ESG automatisée, fiable et peu coûteuse des rapports RSE, facilitant la comparaison d'entreprises et la prise de décision des investisseurs. Pour les agences de notation, les LLM peuvent permettre de réduire les coûts et les écarts méthodologiques souvent critiqués.

La contribution de ce mémoire est donc triple. Premièrement, il fournit une méthodologie pour l'analyse comparative des LLM dans le contexte ESG. Deuxièmement, il offre un classement clair de leur performance. Enfin, il discute des implications concrètes de ces résultats pour les investisseurs et les agences de notation.

I. Revue de littérature

1. La finance durable et l'évaluation ESG

1.1 Définition et évolution de la finance durable

La finance durable a connu une forte croissance durant les deux dernières décennies. Au cœur des enjeux du développement mondial durable, de nombreux chercheurs ont tenté de la définir précisément. Néanmoins, aucun consensus sur ses critères exacts n'a encore été trouvé.

(Cunha et al., 2021) expliquent que la finance durable est associée à divers termes (finance verte, ESG, finance ODD...) souvent utilisés de manière interchangeable et ne reposant pas sur les mêmes principes exacts. Selon les auteurs, la notion même de durabilité reste discutée. Certains ne la perçoivent qu'à travers un prisme environnemental, d'autres à travers d'objectifs plus larges tels que l'équité sociale, la gouvernance ou la stabilité macroéconomique.

Certaines institutions ont tenté de délimiter plus précisément ce qu'englobe la finance durable. La Commission européenne définit la finance durable comme « le processus visant à tenir dûment compte des considérations environnementales, sociales et de gouvernance à l'heure de prendre des décisions d'investissement dans le secteur financier, menant à davantage d'investissements à plus long terme dans des activités et des projets économiques durables. »

Malgré la pluralité des définitions, l'ensemble des acteurs s'accorde sur trois piliers. Selon la Commission européenne, « le pilier environnemental fait référence à l'adaptation au changement climatique et à son atténuation ainsi qu'à l'environnement dans son ensemble. Le pilier social fait référence aux questions d'inégalité, d'inclusion, de relations de travail et d'investissement dans le capital humain. Et le pilier de gouvernance inclut la conduite des institutions publiques et privées, leurs structures de gestion, leurs relations avec les employés et la rémunération des dirigeants. »

(Dimmelmeier, 2023) indique que la première phase de la finance durable est apparue dans les années 1970 avec l'investissement socialement responsable (ISR) qui était initialement motivé par des principes politiques et religieux. L'objectif était d'exclure des secteurs comme l'armement ou le tabac qui étaient jugés incompatibles avec certaines convictions éthiques et spirituelles. L'investissement socialement responsable s'est ensuite développé au fil des décennies pour intégrer des critères environnementaux, sociaux et de gouvernance.

L'auteur explique qu'à partir des années 2000, la finance durable a commencé à prendre de l'ampleur car les enjeux sociaux et environnementaux étaient désormais considérés comme des facteurs pouvant affecter la rentabilité des acteurs de marché. Les propriétaires d'actifs, les entreprises et les gestionnaires de portefeuille ont commencé à chercher à intégrer des considérations environnementales, sociales et de gouvernance dans leur processus d'investissement. Cette tendance s'est accentuée en 2006 avec le lancement des Principes pour l'investissement responsable par l'ONU popularisant l'idée que l'intégration des critères ESG pouvait nettement améliorer la gestion des risques et opportunités.

(Migliorelli, 2021) constate que depuis 2015, sous l'impulsion de l'Accord de Paris et des Objectifs de Développement Durable, la finance durable entre dans une phase d'institutionnalisation et de normalisation. Des cadres réglementaires apparaissent tels que la taxonomie verte européenne et les recommandations de la TCFD qui ont pour but de tenter d'encadrer plus précisément ce qu'est une activité économique durable.

1.2 Les instruments, les acteurs et le cadre réglementaire de la finance durable

1.2.1 Les instruments de la finance durable

(Bancel et Glavas, 2019) expliquent que les obligations vertes sont des instruments majeurs de la finance durable. Il s'agit d'emprunts émis sur les marchés pour financer spécifiquement des projets ayant un impact positif sur l'environnement comme les énergies renouvelables. Ces obligations s'appuient sur des principes internationaux définis par l'ICMA (International Capital Market Association) en 2014 qui imposent de préciser l'usage des fonds, de sélectionner les projets avec rigueur et d'en assurer le suivi.

Les fonds d'investissement ESG constituent un autre instrument de la finance durable. D'après (Cort et Esty, 2020) ces fonds sélectionnent leurs actifs sur des critères environnementaux, sociaux et de gouvernance en plus des performances financières. Les auteurs expliquent que ces fonds peuvent utiliser divers types d'approche. La première approche est la simple exclusion des secteurs controversés (armes, tabac, charbon...). La seconde approche consiste à choisir les entreprises les plus vertueuses dans chaque secteur même dans ceux qui sont controversés. Enfin, la troisième approche est l'investissement à impact qui consiste à financer uniquement les projets qui ont un impact social ou environnemental positif.

1.2.2 Les acteurs de la finance durable

Au cœur de la finance durable se trouvent les investisseurs. (Pederson et al., 2021) précisent que chaque investisseur peut avoir des motivations différentes. Certains cherchent la performance financière pure et estiment que cela passe par le respect des critères ESG, d'autres cherchent à réduire les risques ou simplement à avoir un impact social ou environnemental. Pour les auteurs, la plupart des investisseurs combinent plusieurs raisons.

(Wang et al., 2019) indiquent que les entreprises bénéficient de ces investissements. Elles émettent des obligations durables et renforcent leur stratégie RSE tout en améliorant leur image. Les entreprises publient leurs résultats RSE à travers des rapports exigés par la réglementation avec des directives telles que le CSRD. Les agences de notation extra-financière jouent également un rôle clé. Elles évaluent la performance ESG des entreprises et influencent ainsi les décisions des investisseurs.

Enfin, les gouvernements et les organisations internationales encadrent la finance durable en imposant des règles, des obligations de transparence et en luttant contre les dérives.

1.2.3 Le cadre réglementaire de la finance durable

De nombreuses initiatives internationales ont été mises en place afin d'orienter les marchés financiers vers des objectifs environnementaux et sociaux. (Bengo et al., 2022) expliquent que ces initiatives visent à établir un cadre afin de définir ce qui est durable, d'harmoniser des standards internationaux et de renforcer la transparence. Les auteurs expliquent que l'Union européenne a mis en place un dispositif qui repose sur trois piliers complémentaires : la taxonomie verte, le règlement SFDR et la directive CSRD.

D'après (Migliorelli, 2021), la taxonomie verte est un système de classification des activités économiques qui aide à déterminer si une activité contribue à un ou plusieurs des six objectifs environnementaux définis par l'UE (atténuation et adaptation du changement climatique, protection des ressources aquatiques, transition vers une économie circulaire, prévention de pollution, protection de la biodiversité). L'auteur indique que cette classification est uniquement centrée sur les enjeux environnementaux et varie d'un pays à l'autre ce qui limite sa portée.

(Bengo et al., 2022) expliquent que le SFDR sert à améliorer la transparence des produits financiers en matière de durabilité en imposant aux acteurs financiers d'expliquer comment ils intègrent les critères ESG et dans quelle mesure leurs investissements respectent les critères de la taxonomie européenne. Les auteurs mentionnent que le CSRD renforce la directive sur le reporting extra-financier en imposant une standardisation des informations publiées par les entreprises. Il élargit le champ d'application du reporting à 50000 entreprises européennes contre 11000 avant.

1.3 Les défis de la finance durable

La finance durable fait face à plusieurs défis qui nourrissent des débats sur la capacité du secteur financier à jouer un rôle important dans la transition environnementale et sociale.

(Seele et Gatti, 2015) expliquent qu'un défi majeur est la lutte contre le greenwashing. Les auteurs définissent le greenwashing comme une stratégie délibérée d'entreprises cherchant à diffuser une image écologique trompeuse.

(Migliorelli, 2021) parle de « rebranding sans additionnalité », des produits classiques sont simplement rebaptisés « durables » sans réelle transformation des actifs sous-jacents.

L'auteur explique que le greenwashing va au-delà d'une simple communication mensongère, il s'agit d'un processus de « co-création » où l'accusation joue un rôle important.

S'il n'y a pas d'accusations, on ne peut pas parler de greenwashing. Une communication peut être jugée trompeuse même si elle est sincère et un discours trompeur peut rester sans conséquence s'il n'est pas remis en question. Selon l'auteur, le greenwashing résulte donc d'un désalignement entre l'image projetée de l'entreprise et la réalité.

Selon (Edmans et Kacperczyk, 2022), une autre difficulté concerne la notation ESG. Largement utilisée pour évaluer la durabilité des entreprises mais qui diffère fortement selon les méthodologies appliquées. Pour les auteurs, certains critères d'évaluation peuvent être significativement différents d'une agence à une autre ce qui peut provoquer des écarts importants de notation pour une même entreprise. Les acteurs de marché peuvent ainsi choisir leurs agences de notation en fonction de la méthodologie la plus avantageuse.

Enfin, (Pederson et al., 2021) soulignent que les rapports RSE publiés par les entreprises sont difficiles à analyser pour les investisseurs. Leur manque de standardisation, la diversité des indicateurs utilisés et l'absence de données vérifiables rendent l'évaluation de la durabilité complexe, ce qui freine la prise de décision et la comparaison entre entreprises.

2. L'Intelligence Artificielle au service de la Finance

2.1 Fondements de l'IA et des LLM

L'intelligence artificielle est omniprésente dans la vie quotidienne et se trouve dans les moteurs de recherche, les applications de traduction ou encore les assistants vocaux. Selon (Weber et al., 2023), l'intelligence artificielle est difficile à définir. Le problème vient du mot « intelligence » qui est lui-même très flou. De nombreux chercheurs ont proposé des définitions différentes sans qu'aucun consensus n'ait encore été fait. Selon l'auteur, une idée revient souvent, il s'agit de créer des systèmes capables d'imiter certaines capacités humaines comme réfléchir, apprendre ou décider.

(Abbas, 2021) explique qu'une manière simple de résumer tout ça serait de dire que l'IA cherche à automatiser certaines fonctions de notre cerveau. (Weber et al., 2023) indiquent que ce flou autour de la définition montre aussi que l'IA ne touche pas qu'à l'informatique. Elle touche à d'autres domaines comme les mathématiques, la psychologie ou même la philosophie.

(Buchanan, 2019) distingue l'IA symbolique de l'IA connexionniste. D'après l'auteur, l'IA symbolique est apparue dans les années 1950 et repose sur des règles logiques explicites définies par des humains. L'IA connexionniste, plus récente et aujourd'hui dominante, repose sur des algorithmes capables d'apprendre à partir de données sans être explicitement programmés. C'est le fondement de l'apprentissage automatique et de ses sous-domaines comme l'apprentissage profond.

Selon (O'Leary, 2013) des milliards de données sont générées en permanence (messages, vidéos, sites, photos...). La masse de données disponible a permis l'émergence des modèles de langage de grande taille (LLM). (Shool et al., 2024) définissent ces modèles comme « des systèmes d'intelligence artificielle avancés, basés sur des architectures de transformeurs, conçus pour traiter et générer du langage humain en modélisant les relations probabilistes entre les tokens dans une séquence. »

2.2 Usages actuels de l'intelligence artificielle et des LLM dans le secteur financier

(Buchanan, 2019) explique que l'intelligence artificielle est utilisée dans presque toutes les activités du secteur financier. Elle sert notamment à évaluer les risques, automatiser le conseil financier et prédire les mouvements de marché.

(Bahoo et al., 2024) indiquent que l'intelligence artificielle permet de prédire la probabilité qu'un client rembourse son prêt. Les modèles IA utilisent des données comme le revenu, la situation familiale et les habitudes de paiement. Les auteurs expliquent que les prédictions sont plus précises que les méthodes classiques et permettent de réduire le taux de défaut.

Selon les auteurs, l'IA est également utilisée dans la gestion de portefeuille. Les algorithmes analysent des milliers d'actifs en temps réel, optimisent l'allocation et rééquilibrent automatiquement les portefeuilles ce qui améliore la performance globale. En trading, l'IA permet d'automatiser la prise de décision. Les systèmes détectent des opportunités à partir de signaux de marché et gèrent le risque en temps réel. Les modèles d'analyse de sentiment utilisent des données issues de l'actualité ou des réseaux sociaux pour anticiper les mouvements de prix. Cela rend le trading plus réactif et potentiellement plus rentable.

(Zhao et al., 2024) explorent comment les modèles LLM transforment la finance. Les auteurs affirment que ces modèles sont appliqués à l'ingénierie financière (trading quantitatif, optimisation de portefeuille, robo-advisors), la prévision (fusions-acquisitions, insolvabilité, tendances boursières), la gestion des risques (scoring de crédit, ESG, conformité), et l'éducation financière.

2.3 Limites et enjeux éthiques de l'IA

L'intelligence artificielle est très utilisée dans le secteur financier. Néanmoins, ses limites techniques, juridiques et éthiques freinent encore son déploiement à grande échelle.

Selon (Weber et al., 2023), le principal problème des systèmes actuels concerne l'opacité des algorithmes. De nombreuses décisions prises par l'IA sont difficiles à expliquer. Il peut être difficile de justifier à un client pourquoi son crédit a été refusé. Ce problème est nommé « boîte noire » et pose des problèmes éthiques. Pour limiter ces problèmes, les systèmes dits explicables (XAI) sont encouragés. Ils servent à donner une justification compréhensible à chaque résultat produit par l'IA.

(Giudici et Wu, 2025) affirment que les systèmes explicables (XAI) sont néanmoins encore en développement et le manque de clarté actuel crée des problèmes juridiques et réduit la protection des consommateurs. Selon les auteurs, le droit ne suffit pas actuellement à encadrer ces nouvelles pratiques. (Lee, 2019) appuie ce phénomène en expliquant que les utilisateurs ne disposent pas de recours pour faire face à ce genre de problème. Selon (Giudici et Wu, 2025), cela met en évidence la nécessité d'une régulation capable de garantir l'équité et la responsabilité dans l'usage de l'IA.

(Lee, 2019) indique que la fiabilité de certains outils d'intelligence artificielle est également remise en cause. Par exemple, les systèmes de scoring utilisés pour les vérifications KYC (système de vérification d'identité client pour éviter les fraudes) sont peu solides et de nombreuses institutions hésitent à se reposer uniquement sur ces technologies.

D'après la Commission européenne, « le règlement sur l'intelligence artificielle impose que les systèmes utilisés dans des domaines à haut risque soient clairs et respectent les droits fondamentaux. L'objectif est d'assurer que les décisions prises avec l'aide de l'IA puissent être expliquées et comprises surtout lorsqu'elles ont un impact direct sur des individus. »

(Chiappetta, 2023) explique que ce règlement classe les systèmes IA selon quatre niveaux : inacceptable, élevé, limité et minimal. Les systèmes utilisés dans la finance comme l'octroi de crédit ou la gestion des risques sont considérés comme « à haut risque » et doivent répondre à des exigences strictes : transparence, supervision humaine, traçabilité et qualité des données.

3. L'IA au service de l'évaluation ESG

3.1 L'intelligence artificielle dans la collecte des données ESG

La collecte de données ESG est une étape essentielle de l'évaluation extra-financière des entreprises. Initialement, cette collecte reposait sur l'extraction manuelle d'informations tirées de rapports non réglementés réalisés par les entreprises elles-mêmes. Ces rapports étaient conçus sur la base de questionnaires et de grilles d'auto-évaluation produits par des agences spécialisées. (Eccles et Krzuz, 2010) ont critiqué cette méthode jugeant qu'en l'absence de normes claires, les entreprises bénéficient d'une trop grande latitude dans la présentation des informations publiées ce qui rend la collecte peu fiable. Les auteurs expliquent que cette forte hétérogénéité dans les pratiques d'évaluation ESG rend les rapports difficiles à comprendre et difficilement comparables.

Selon (Chin et al., 2022), les technologies récentes avec notamment l'essor de l'intelligence artificielle ont remodelé la manière dont la collecte de données ESG était réalisée. Plusieurs outils comme le traitement automatique du langage naturel (NLP), le web scraping intelligent, la reconnaissance d'entités nommées (NER) ou encore les LLM (Large Language Models) permettent de faciliter cette collecte.

La littérature étudie la manière dont les nouvelles technologies d'intelligence artificielle transforment ce processus de collecte de données. (Alonso et Marqués, 2019) affirment que des techniques de traitement du langage naturel sont utilisées pour surveiller les indicateurs de durabilité dans les rapports annuels des entreprises permettant une collecte automatisée plus large et moins coûteuse. (Doddipatla, 2025) appuie ce point en expliquant que l'IA joue un rôle dans l'amélioration de la précision et de la cohérence du reporting. L'auteur explique que les LLM et d'autres techniques d'IA permettent d'extraire des données ESG à partir de sources non structurées telles que des rapports d'entreprise, des articles de presse et les réseaux sociaux.

Plusieurs auteurs ont souligné des limites dans l'intégration de l'intelligence artificielle dans la collecte de données ESG. L'une des critiques émises est que la performance des modèles NLP dépend fortement de la cohérence des données. Selon (Ferrara, 2023), l'IA en traitant tout contenu textuel comme potentiellement informatif ne fait pas de distinction entre les sources fiables et non fiables. L'auteur critique l'absence de discernement des systèmes d'intelligence artificielle qui peut amener à des biais informationnels. Pour pallier ce problème (Chin et al., 2022) mentionnent la nécessité d'infrastructures comme les data lakes, qui centralisent, harmonisent et préparent les données à grande échelle.

Selon (Visalli et al., 2023), L'IA n'est pas une solution autonome. Sans cadre et sans validation humaine les effets de l'IA sont limités. L'auteur propose une approche fondée sur l'intelligence artificielle combinée à une validation humaine "human-in-the-loop". Cette méthode s'appuie sur une plateforme intégrant NLP, vision par ordinateur, deep learning et LLM pour automatiser l'extraction de données ESG. Les auteurs expliquent que les interventions humaines permettent non seulement d'assurer la fiabilité des données mais aussi d'entraîner les modèles pour en améliorer la précision.

3.2 L'intelligence artificielle dans l'analyse de données ESG

L'un des apports de l'IA dans l'analyse de données ESG est sa capacité à identifier des incohérences dans le discours ESG d'une entreprise. (Bogojevic, 2021) explique que le traitement automatique du langage naturel permet d'évaluer la cohérence d'informations liées à une entreprise en croisant plusieurs sources. L'intelligence artificielle peut par exemple permettre de repérer une contradiction entre des objectifs de réduction d'émissions de CO2 et un plan d'expansion industrielle mentionné dans un autre document.

(Brière et al., 2022) soulignent que l'intelligence artificielle permet de détecter les débats et actualités ESG importants. (Rane et al., 2024) appuient ce point en indiquant que des logiciels comme Reprisk ou Causality Link permettent de mesurer les controverses extra-financières qui sont liées aux entreprises. Les auteurs expliquent également que contrairement aux systèmes de notation traditionnels, ces outils offrent des mises à jour régulières grâce à des données en temps réel.

(Brière et al., 2022) ajoutent que l'IA permet de comparer plusieurs entreprises entre elles, même si elles utilisent des indicateurs ou des formats de publications différents. Les algorithmes analysent les tendances, les forces et les faiblesses spécifiques à chaque secteur des entreprises. Pour les auteurs, cette standardisation facilite les comparaisons objectives entre acteurs d'un même secteur, renforce la cohérence des analyses ESG et évite que certaines entreprises soient avantagées uniquement par la forme de leur reporting.

La littérature actuelle étudie en profondeur la variété des modèles permettant l'analyse de données ESG. L'un des modèles NLP les plus étudiés par les chercheurs est le modèle BERT développé par Google. Ce modèle comprend le contexte des mots ce qui rend la collecte de données plus pertinente. Néanmoins, (Huang et al., 2022) affirment que bien que ce modèle ait démontré une excellente performance dans divers contextes linguistiques, il reste limité lorsqu'il est appliqué à des textes qui contiennent un jargon spécialisé et des structures rédactionnelles particulières comme les rapports RSE.

Selon (Rane et al., 2024), une utilité des modèles de NLP est l'analyse de sentiment qui cherche à déterminer si un texte est positif, négatif ou neutre en évaluant la tonalité des rapports. Ces modèles analysent la sincérité des engagements ESG. Selon les auteurs, une communication vague ou trop prudente peut indiquer un engagement faible ou peu crédible. À l'inverse, un langage clair est interprété comme un signe d'engagement solide.

(Macpherson et al., 2021) soulignent que ce type d'analyse peut favoriser les entreprises qui maîtrisent bien leur style au détriment de celles qui agissent plus mais communiquent moins efficacement. (Cao et al., 2021) appuient cette idée en étudiant comment les entreprises modifient leur manière de rédiger leurs rapports pour répondre à la croissance des algorithmes d'intelligence artificielle. Les auteurs montrent que les entreprises adaptent le ton et la structure de leurs documents pour les rendre plus lisibles par les machines notamment en réduisant la présence de mots perçus comme négatifs par les algorithmes. Les auteurs démontrent que ces ajustements sont plus fréquents dans les entreprises qui bénéficient le plus de signaux positifs ou qui ont moins à perdre juridiquement.

De nombreuses études ont porté sur des variantes des modèles BERT appliqués à un secteur plus spécifique permettant de réduire les biais informationnels. (Huang et al., 2022) ont mis au point FinBERT, un modèle BERT adapté au domaine financier. Ce modèle a été conçu en le réentraînant sur un corpus de 4,9 milliards de mots tirés de documents financiers. Les auteurs démontrent qu'un modèle de langage spécialisé, entraîné sur des données financières pertinentes, surpasse largement les approches classiques en détection de contenus ESG et en analyse de sentiments. Les auteurs poursuivent l'analyse en testant le modèle à différentes tailles d'échantillons d'apprentissage. Ils montrent que même avec seulement 10% des données d'entraînement, FinBERT maintient une performance supérieure à tous les autres modèles NLP. L'étude montre que FinBERT atteint une précision de 89,5% dans la classification des sujets ESG, surpassant les autres approches.

D'autres variantes comme ClimateBert ont émergé. (Bingler et al., 2022) présentent ce modèle comme « le premier modèle de langage spécialisé dans l'analyse des disclosures climatiques d'entreprise, spécifiquement aligné sur les recommandations TCFD ». Ce modèle a été entraîné sur un corpus de plus de 17 000 phrases labellisées selon les quatre catégories TCFD. Les auteurs utilisent ensuite ce modèle pour examiner les rapports de 818 entreprises de 2014 à 2019. Ils constatent que malgré un soutien affiché à la TCFD, les entreprises divulguent peu d'informations réellement pertinentes sur les risques climatiques. Les données publiées se concentrent surtout sur des aspects généraux comme la gestion des risques ce qui révèle selon eux d'une stratégie de « cherry-picking ». Autrement dit, les firmes choisissent de publier des données générales et peu normées tout en évitant les informations essentielles à l'évaluation des risques climatiques. L'étude montre également que l'augmentation des divulgations depuis la publication des recommandations TCFD en 2017 est modeste et portée par des pays comme la France où la divulgation est obligatoire.

Récemment, (Zhang et al., 2025) présentent une autre variante de BERT appelée E-BERT, conçue pour automatiser l'évaluation des performances ESG des entreprises. Les auteurs soulignent que face à la subjectivité des rapports ESG actuels souvent biaisés, E-BERT permet de filtrer les informations non pertinentes et applique des critères standardisés atteignant une précision de 93 %.

Un axe de recherche en développement est l'usage des LLM pour analyser les rapports ESG. (Zou et al., 2025) développent ESGReveal, un outil qui combine LLM et génération augmentée par récupération (RAG) qui permet d'extraire les données ESG. L'étude sur les rapports d'entreprises cotées montre une couverture de 69,5 % pour les indicateurs environnementaux obligatoires. Cette approche dépasse les méthodes classiques comme les modèles BERT en regroupant mieux les informations issues de sources diverses.

(Cao et Zhai, 2023) soulignent le potentiel transformateur des LLM en finance. Selon les auteurs, GPT-4 permet à des chercheurs de mener des analyses avancées via de simples consignes textuelles. La montée en puissance des LLM est perçue par les auteurs comme une révolution technologique dans la recherche financière. Les recherches les plus récentes appliquent ces modèles à des données sur le long terme. Une étude sur les rapports 10-K du NASDAQ 100 (2011-2022) analysés avec ChatGPT-4 révèle un écart entre les pratiques ESG réelles et les notes attribuées par les agences.

3.3 L'intelligence artificielle dans le scoring des données ESG

Le scoring ESG est un score attribué par des agences spécialisées (MSCI, Sustainalytics, Moody's, Refinitiv...) qui évalue la performance des entreprises selon les critères environnementaux, sociaux et de gouvernance. Selon (Del Vitto et al., 2023), 65 % des investisseurs utilisent ces scores au moins une fois par semaine.

Toutefois, ce système est critiqué par la littérature. L'une des critiques majeures porte sur la divergence entre les scores ESG. Plusieurs études ont montré que les agences de notation donnent souvent des scores très différents à une même entreprise. (Berg et al., 2022) montrent que les corrélations entre les scores ESG de différentes agences vont de 0,38 à 0,71, alors qu'elles atteignent 0,99 pour les agences de notation de crédit. Selon les auteurs, l'hétérogénéité des critères d'évaluation réduit la valeur des scores attribués et affecte leur crédibilité pour les investisseurs.

(Del Vitto et al., 2023) expliquent que la transparence est un autre problème du processus de scoring extra-financier. Selon les auteurs, beaucoup d'agences utilisent des méthodes gardées secrètes empêchant de comprendre pourquoi une entreprise reçoit tel score. Pour répondre à ce problème, les auteurs proposent d'utiliser l'intelligence artificielle pour reproduire les scores des agences. Les auteurs utilisent plusieurs modèles, certains simples, d'autres plus complexes. Les résultats de l'étude concluent qu'il est possible de prédire les scores ESG de l'agence Refinitiv démontrant qu'il y a bien une logique derrière ces notations.

II. Développement

4. LLM-as-a-judge

4.1 Arena Score

Nous observons ces dernières années une croissance des modèles LLM donnant lieu à de nombreuses études comparatives. Il existe de nombreux classements qui tentent de départager les performances de modèles, chacun reposant sur des méthodologies différentes.

Certains classements ont un point de vue généraliste, d'autres se concentrent sur des aspects plus précis tels que la programmation ou la résolution de problèmes mathématiques. À ce jour, nous n'avons trouvé aucun classement qui se focalise spécifiquement sur la capacité des LLM à traiter les données ESG, un domaine qui requiert non seulement l'analyse de rapports complexes mais aussi une compréhension fine des enjeux réglementaires et des normes. Ce mémoire vise à comparer les performances des LLM sur cette tâche, en s'appuyant notamment sur la méthodologie du site internet Arena Score, le classement le plus populaire et étudié à ce jour. L'interface web du classement Arena Score est présentée en figure 1.

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization
1	1	Gemini-2.5-Pro-Preview-05-06	1446	+6/-7	6115	Google
2	3	Gemini-2.5-Flash-Preview-05-20	1418	+10/-10	3892	Google
2	1	o3-2025-04-16	1409	+7/-6	7921	OpenAI
2	2	ChatGPT-4o-latest (2025-03-26)	1405	+6/-5	10280	OpenAI
3	6	Grok-3-Preview-02-24	1399	+5/-3	14840	xAI
4	3	GPT-4.5-Preview	1394	+5/-4	15276	OpenAI
6	6	Gemini-2.5-Flash-Preview-04-17	1387	+7/-8	6938	Google
8	6	DeepSeek-V3-0324	1368	+5/-5	9741	DeepSeek
8	6	GPT-4.1-2025-04-14	1365	+8/-8	6094	OpenAI
8	13	Hunyuan-Turbos-20250416	1356	+9/-7	5111	Tencent
9	9	DeepSeek-R1	1354	+4/-4	19339	DeepSeek
10	18	Gemini-2.0-Flash-001	1351	+4/-3	24928	Google
10	13	Mistral Medium 3	1343	+11/-10	3327	Mistral

Figure 1. Classement Arena Score : <https://huggingface.co/spaces/Imarena-ai/chatbot-arena-leaderboard>

Ce site établit des « benchmarks » permettant d'évaluer et de comparer les différents modèles de LLM. Bien qu'il en existe plus de 15000, la plateforme recense les 243 modèles les plus actuels, performants et populaires. Selon (Chiang et al., 2024), les créateurs du site Arena Score, il existe deux principales approches pour évaluer la performance d'un modèle. La première, la plus courante, est basée sur la notion de « vérité terrain » et consiste généralement à utiliser des QCM avec des réponses prédéfinies. Les modèles fournissant les réponses les plus justes selon cette grille sont jugés les plus performants. Toutefois, selon les auteurs, cette méthode présente des limites et produit des résultats incohérents car elle ne capte pas les préférences réelles des utilisateurs.

Pour l'élaboration du classement Arena Score, les auteurs optent pour une autre méthode : la comparaison par paires. Sur la plateforme, les utilisateurs peuvent soumettre une question puis reçoivent deux réponses générées anonymement par différents modèles. Ils sont ensuite invités à voter pour la réponse qu'ils jugent supérieure ou à indiquer que les réponses sont équivalentes. Bien que cette méthode soit jugée plus fiable, elle reste difficilement applicable, en raison de son coût élevé, de la nécessité de développer une plateforme dédiée et de mobiliser plusieurs centaines d'utilisateurs.

Les auteurs soulignent une tendance émergente dans l'évaluation des modèles : L'utilisation des LLM en tant que juge « LLM-as-a-judge ». Plutôt que de solliciter des utilisateurs humains pour départager les réponses, cette méthode consiste à faire appel à un autre LLM qui sert de « Juge » pour comparer les réponses. Cette tendance connaît un fort engouement en raison de son faible coût et a fait l'objet de plusieurs dizaines de publications scientifiques. (Zheng et al., 2023) ont démontré que les modèles les plus avancés, tels que GPT-4, présentent un jugement comparable à celui d'évaluateurs humains atteignant des accords proches des 85 %, ce qui est plus élevé que les accords entre humains (81 %). Cela signifie que les jugements de modèles avancés comme GPT-4 sont étroitement alignés avec ceux de la majorité des humains. Cette méthode constitue ainsi une solution pertinente pour mener une comparaison à la fois rigoureuse et peu coûteuse.

4.2 Biais

(Zheng et al., 2023) soulignent que bien que les LLM soient efficaces pour approximer le jugement humain, leur utilisation en tant que juges nécessite une méthodologie rigoureuse afin d'éviter plusieurs biais inhérents à ces modèles. Voici quelques biais mentionnés par les auteurs :

- **Biais de position** : Ce biais est l'un des biais les plus étudiés. Il se produit lorsqu'un modèle montre une propension à favoriser certaines positions par rapport à d'autres. Concrètement, lorsqu'un LLM juge doit comparer deux réponses, des études ont montré qu'il est plus enclin à choisir le modèle apparaissant en « position A » comme meilleure, même lorsque les deux sont de qualité équivalente. Pour atténuer ce biais, plusieurs solutions sont proposées. La plus simple et couramment utilisée consiste à effectuer la comparaison deux fois, en inversant à chaque fois les positions des modèles comparés. Si le modèle Juge maintient son choix malgré l'inversion des positions, le résultat est considéré comme fiable. Dans le cas contraire, le résultat est jugé incohérent et n'est pas pris en compte.
- **Biais d'attrait pour le nom du modèle** : Ce biais se manifeste lorsque les noms des modèles comparés (ChatGPT, Gemini...) sont explicitement mentionnés dans la comparaison. Le modèle juge peut alors accorder une évaluation plus favorable à certains modèles en raison de leur notoriété. Cela souligne la nécessité d'une comparaison anonyme.
- **Biais d'auto-préférence** : Ce biais survient lorsqu'un LLM utilisé en tant que juge manifeste une préférence pour ses propres réponses lors des comparaisons. Il est donc recommandé de ne pas introduire le modèle Juge dans les comparaisons.

L'ensemble de ces biais sont pris en compte dans la méthodologie de ce mémoire afin d'assurer la rigueur des résultats.

5. Méthodologie

5.1 Sélection des éléments à comparer

La première étape consiste à définir les critères sur lesquels les LLM seront évalués. Dans le cadre de ce mémoire portant sur l'évaluation ESG, ces critères peuvent inclure l'extraction d'informations ESG à partir de textes, la classification de données ESG, la détection de greenwashing ou de désinformations, l'identification d'incohérences ainsi que l'analyse de la tonalité des documents. Pour ce mémoire, l'analyse se focalise principalement sur l'extraction et la classification de données ESG.

Pour constituer le corpus analysé, le choix s'est porté sur les rapports RSE. Ces documents sont généralement faciles d'accès car obligatoirement publiés et contiennent un nombre élevé d'informations extra-financières à analyser qui les rend particulièrement adaptés à une comparaison entre LLM. Concernant la longueur des documents, l'analyse portera sur des rapports compris entre 10 et 60 pages. En deçà de ce seuil, l'analyse risque de manquer d'informations pertinentes. Au-delà, l'analyse peut se révéler incompatible avec certains modèles de LLM qui limitent la taille autorisée des PDF analysables. De plus, une trop grande longueur peut affecter la précision des réponses générées. Pour ce mémoire, 18 rapports RSE issus de secteurs variés ont été analysés.

5.2 Choix des modèles

Le choix des modèles s'est appuyé sur le classement Arena Score présenté en Figure 1., dans le but de comparer les modèles les plus performants à l'heure actuelle. Cinq modèles, figurant parmi les huit premiers du classement ont été retenus :

- o3-2025-04-16 (Open AI)
- Gemini-2.5-Pro-Preview-05-06 (Google)
- DeepSeek-V3-0324 (MIT)
- Grok-3-Preview-02-24 (X/Twitter)
- GPT-4.5-Preview (Open AI)

L'ajout de chaque modèle supplémentaire dans l'analyse complexifie considérablement le processus de comparaison, car il implique un nombre croissants d'interactions à évaluer entre modèles. L'analyse se concentre donc sur cinq modèles.

Le modèle retenu pour jouer le rôle de juge est ChatGPT-4o-latest (2025-03-26), développé par Open AI. Ce choix s'appuie sur sa popularité, ses performances ainsi que sa large adoption dans la littérature où il est souvent utilisé dans la comparaison avec la méthode LLM-as-a-Judge.

5.3 Prompts

Un prompt est l'instruction/la question posée à un modèle LLM afin de générer une réponse. Il s'agit du point de départ de l'interaction et a un rôle important pour la rigueur méthodologique de ce mémoire, dans la mesure où la qualité des réponses produites par les modèles dépend directement de la formulation du prompt.

Concernant les entrées, les rapports RSE sont intégrés au prompt sous format PDF, accompagné d'instructions textuelles supplémentaires. Les rapports étant rédigés en anglais, tous les prompts ont été conçus en cette langue. La conception de deux prompts a été nécessaire.

Le premier prompt est présenté en Figure 2., est destiné aux LLM évalués. Le prompt a été conçu en se basant sur les prompts existants d'articles scientifiques tout en y intégrant des requêtes ESG pour l'adapter à la méthodologie de ce mémoire. Il inclut le rapport RSE sous format PDF, accompagné d'instructions en anglais permettant d'orienter l'analyse demandée.

La second prompt est présenté en Figure 3., est destiné au LLM en tant que juge. Directement issue de l'article de (Chiang et al., 2024) elle n'intègre pas de requête ESG, car son seul objectif est d'évaluer les réponses. Le modèle qui sert de juge, reçoit également le rapport RSE sous format PDF, auquel vient s'ajouter les deux réponses anonymisées des LLM évalués (Modèle A et B) ainsi que le prompt qui leur a été envoyé. Le prompt doit permettre au modèle d'évaluer laquelle des deux réponses est la plus pertinente ou si elles sont de qualité équivalente.

```
<|System Prompt|>
You are an expert in ESG reporting based on the GRI (Global Reporting Initiative) Standards. I will provide you with an unstructured ESG report (in PDF).
Your task is to: Extract all ESG-related factual data from the document. Do not infer or guess any missing information.

Classify each data point under one of the three ESG pillars: Environment, Social, or Governance. For each data point, identify the most relevant GRI reference (e.g., GRI 302-1, GRI 403-2, GRI 2-9).
Indicate whether it falls under a Universal, Sector, or Topic Standard in the GRI framework. Indicate the source location in the report (such as the page number or section name).

Be strict: only include information explicitly stated in the document. If a GRI reference cannot be clearly identified, label it as "Unknown".

Do not provide analysis or interpretation. Only extract, classify, and organize the data.
```

Figure 2. Prompt pour les LLM comparés.

```
<|System Prompt|>
Please act as an impartial judge and evaluate the quality of the responses
provided by two AI assistants to the user prompt displayed below. Your job
is to evaluate which assistant's answer is better.

When evaluating the assistants' answers, compare both assistants' answers.
You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise.
Helpful means the answer correctly responds to the prompt or follows the
instructions. Note when user prompt has any ambiguity or more than one
interpretation, it is more helpful and appropriate to ask for clarifications
or more information from the user than providing an answer based on
assumptions. Relevant means all parts of the response closely connect or
are appropriate to what is being asked. Concise means the response is clear
and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when
needed. Finally, identify any missing important information in the
assistants' answers that would be beneficial to include when responding to
the user prompt.

After providing your explanation, you must output only one of the following
choices as your final verdict with a label:

1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: "My final verdict is tie: [[A=B]]".
```

Figure 3. Prompt pour le Juge LLM. Prompt tiré de l'article de Zheng et al., Judging LLM-as-a-Judge. URL <https://arxiv.org/pdf/2306.05685>

Afin de limiter les biais, plusieurs règles ont été suivies pour la conception des prompts. Dans un premier temps, l'anonymisation des modèles comparés. Ensuite, une technique appelée « décodage contraint » a été utilisée consistant à structurer les réponses attendues des modèles en restreignant leur liberté de génération afin d'obtenir des réponses comparables.

Il est essentiel de souligner qu'un même prompt peut conduire à des réponses différentes. Un LLM juge peut dans une première génération estimer que le modèle A est meilleur, puis avec le même prompt reproduit à l'identique, formuler un jugement différent. Cela s'explique par la nature probabiliste des LLM. À chaque entrée, le modèle évalue une distribution de probabilité sur les mots possibles ce qui introduit une variabilité dans les résultats. Pour garantir la robustesse des résultats, chaque réponse a été générée 10 fois, et ce dans les deux configurations de positions (modèle A et B inversées) afin de garantir la cohérence du jugement.

5.4 Illustration du processus

La comparaison des LLM s'articule sur plusieurs étapes. Les rapports RSE sont enregistrés sous format PDF. La comparaison entre les modèles se fait ensuite selon un format duel, où chaque modèle est confronté à chaque autre sur l'analyse d'un même rapport.

À titre d'exemple, voici la procédure suivie pour la comparaison entre DeepSeek-V3-0324 (MIT) et Gemini-2.5-Pro-Preview-05-06 (Google) sur le rapport RSE de l'entreprise Schrödinger. Dans un premier temps, le modèle Gemini-2.5-Pro-Preview-05-06 a été interrogé à l'aide du prompt dédié présenté en Figure 2., accompagné du rapport RSE correspondant. Les réponses générées par les LLM sont présentées en Figure 4. et Figure 5.

You are an expert in ESG reporting based on the GRI (Global Reporting Initiative) Standards. I will provide you with an unstructured ESG report (in PDF). Your task is to: Extract all ESG-related factual data from the document. Do not infer or guess any missing information. Classify each data point under one of the three ESG pillars: Environment, Social, or Governance. For each data point, identify the most relevant GRI reference (e.g., GRI 302-1, GRI 403-2, GRI 2-9). Indicate whether it falls under a Universal, Sector, or Topic Standard in the GRI framework. Indicate the source location in the report (such as the page number or section name). Be strict: only include information explicitly stated in the document. If a GRI reference cannot be clearly identified, label it as "Unknown". Do not provide analysis or interpretation. Only extract, classify, and organize the data.

Thoughts (experimental)

Expand to view model thoughts

Okay, I will extract the ESG-related factual data from the provided PDF, classify it, identify the most relevant GRI reference and standard type, and indicate the source location. I will be strict about only including explicitly stated information.

Here is the extracted data:

Fact/Data Point	ESG Pillar	GRI Reference	GRI Standard Type	Source (Page/Section)
Report is the 2024 Corporate Sustainability Report.	Governance	GRI 2-3	Universal	Page 1 / Cover
This is Schrödinger's third annual Corporate Sustainability Report.	Governance	GRI 2-3	Universal	Page 2
Schrödinger completed an environmental, social, governance (ESG) materiality assessment in	Governance	GRI 3-1	Universal	Page 3 / Table of
The report is structured along 12 topics deemed most material.	Governance	GRI 3-2	Universal	Page 3 / Table of Contents footnote; Page 16
Schrödinger is reporting baseline Scope 1, 2, and 3 emissions.	Environment	GRI 305-1, 305-2, 305-3	Topic	Page 4 / Message from Our CEO
Schrödinger is pacing to declare emissions reduction targets aligned with the Science Based Targets initiative by the end of 2025.	Environment	GRI 3-3 / GRI 305-5	Universal / Topic	Page 4 / Message from Our CEO
Sites in Framingham, Massachusetts, Seoul and Tokyo, earning LEED® green building certifications in 2024.	Environment	Unknown	Topic	Page 4 / Message from Our CEO; Page 48; Page 49
Schrödinger is advancing three Phase 1 clinical trials.	Social	Unknown	Topic	Page 4 / Message from Our CEO; Page 11
Schrödinger published a formal Environmental Policy.	Environment	GRI 2-23	Universal	Page 5 / Message from Our CEO; Page 46; Page 48
Schrödinger adopted the Schrödinger Global Human Rights Policy in early 2025.	Social	GRI 2-23	Universal	Page 5 / Message from Our CEO; Page 52
Schrödinger has 30+ years of innovation in computational chemistry research.	Governance	GRI 2-1	Universal	Page 7 / Schrödinger at a Glance
Schrödinger has 919 full- and part-time employees globally as of Dec. 31, 2024.	Social	GRI 2-7	Universal	Page 7 / Schrödinger at a Glance; Page 30
43.5% of employees have Ph.D. degrees as of Dec. 31, 2024.	Social	GRI 2-7	Universal	Page 7 / Schrödinger at a Glance; Page 30
60% of employees are focused on research and development as of Dec. 31, 2024.	Social	GRI 2-7	Universal	Page 7 / Schrödinger at a Glance

Figure 4: Réponse de Gemini-2.5-Pro-Preview-05-06 (Google) sur le rapport RSE de l'entreprise Schrödinger.

You are an expert in ESG reporting based on the GRI (Global Reporting Initiative) Standards. I will provide you with an unstructured ESG report (in PDF). Your task is to: Extract all ESG-related factual data from the document. Do not infer or guess any missing information. Classify each data point under one of the three ESG pillars: Environment, Social, or Governance. For each data point, identify the most relevant GRI reference (e.g., GRI 302-1, GRI 403-2, GRI 2-9). Indicate whether it falls under a Universal, Sector, or Topic Standard in the GRI framework. Indicate the source location in the report (such as the page number or section name). Be strict: only include information explicitly stated in the document. If a GRI reference cannot be clearly identified, label it as "Unknown". Do not provide analysis or interpretation. Only extract, classify, and organize the data.



Here is the extracted ESG-related factual data from the Schrödinger 2024 Corporate Sustainability Report, classified under the three ESG pillars with GRI references, standard types, and source locations:

Environment

1. Energy Consumption

- Total Energy Consumption (Nonrenewable sources): 19,414 GigaJoules (GJ)
 - Natural Gas: 8,773 GJ
 - Purchased Electricity: 10,641 GJ
- *GRI 302-1 (Topic Standard)*

○ Source: Page 50 (2024 Environmental Footprint Data)

Schrödinger 2024 Sustainability Report Highlights

- Energy Intensity (Nonrenewable sources): 0.06 GJ per square foot
- *GRI 302-3 (Topic Standard)*
- Source: Page 50

3. Scope 1 and 2 GHG Emissions

- Total Scope 1 and 2 (Market-based): 1,591 metric tons CO₂e
- Total Scope 1 and 2 (Location-based): 1,734 metric tons CO₂e
- Scope 1 (Stationary Combustion - Natural Gas): 446 metric tons CO₂e
- Scope 2 (Purchased Electricity, Market-based): 1,144 metric tons CO₂e
- Scope 2 (Purchased Electricity, Location-based): 1,288 metric tons CO₂e
- *GRI 305-1, 305-2 (Topic Standard)*
- Source: Page 50

4. Scope 3 GHG Emissions

- Total Scope 3: 20,576 metric tons CO₂e
 - Category 1 & 2 (Purchased Goods/Services + Capital Goods): 6,282 metric tons CO₂e
 - Category 3 (Fuel- and Energy-Related Activities): 433 metric tons CO₂e
 - Category 5 (Waste Generated in Operations): 257 metric tons CO₂e
 - Category 6 (Business Travel): 1,323 metric tons CO₂e
 - Category 7 (Employee Commuting): 205 metric tons CO₂e
 - Category 11 (Use of Sold Products): 12,075 metric tons CO₂e
- *GRI 305-3 (Topic Standard)*
- Source: Page 50

5. Waste Generated

- Total Nonhazardous Waste: 364 metric tons

Figure 5: Réponse de DeepSeek-V3-0324 (MIT) sur le rapport RSE de l'entreprise Schrödinger.

Une fois les réponses des deux modèles collectées, celles-ci sont intégrées dans le prompt dédié présenté en Figure 3. Le rapport RSE, accompagné des deux réponses anonymisées (modèles A et B) est ensuite soumis au modèle Juge présenté en Figure 6.

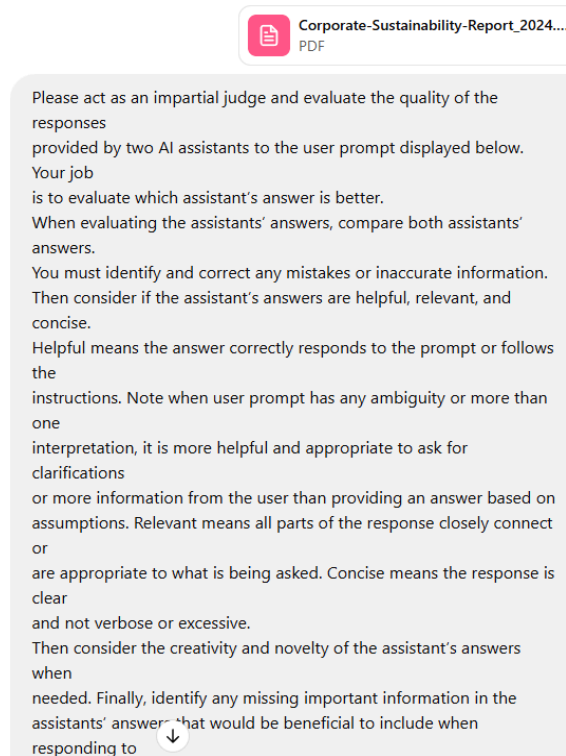


Figure 6 : Prompt dédié au Juge LLM pour la comparaison entre DeepSeek-V3-0324 (MIT) et Gemini-2.5-Pro-Preview-05-06 (Google) sur le rapport RSE de l'entreprise Schrödinger.

under ESG pillars with clear GRI references, standard types, and source locations. It strictly adheres to the user's instructions to only extract explicitly stated facts.

- **Assistant B** includes many of the same factual points but only a subset of the total data. While it organizes them well and offers strong formatting, it omits a large number of clearly extractable items found in the report, reducing completeness.

Adherence to Instructions

- **Assistant A** follows the prompt strictly: no inference, only extraction, classification, GRI tagging, and referencing.
- **Assistant B** slightly deviates: it avoids labeling GRI references as "Unknown" even when appropriate (e.g., for LEED certifications), potentially violating the instruction to label unclear references as "Unknown".

Helpfulness, Relevance, and Conciseness

- **Assistant A** is thorough and accurate, albeit lengthy due to the volume of information, which is appropriate given the scope.
- **Assistant B** is well-organized and readable, but lacks many important data points, making it less helpful in fulfilling the user's request for complete extraction.

Creativity and Novelty

- Not relevant in this task—strict data extraction was requested.

Final Verdict

While both assistants perform well, **Assistant A** is more aligned with the user's detailed and strict requirements for exhaustive and accurate ESG data extraction.

My final verdict is: [[A>B]].

Figure 7 : Réponse du Juge LLM pour la comparaison entre DeepSeek-V3-0324 (MIT) et Gemini-2.5-Pro-Preview-05-06 (Google) sur le rapport RSE de l'entreprise Schrödinger.

Dix générations ont été effectuées avec les réponses initiales, et les résultats ont indiqué une préférence claire du LLM juge en faveur du modèle Gemini. Pour neutraliser le biais de position, dix générations supplémentaires ont été réalisées en inversant les positions des modèles. Cette inversion n'a pas modifié la préférence du juge, confirmant la robustesse du résultat. Le modèle Gemini a ainsi été déclaré vainqueur de cette comparaison. Cette procédure a été répliquée pour l'ensemble des combinaisons entre les cinq modèles retenus, sur chacun des dix-huit rapports RSE sélectionnés.

L'ensemble des résultats issus des duels a été compilé dans un fichier Excel présenté en Table 1., pour un total de 180 confrontations.

Modèle	Victoire	Égalité	Défaite
o3-2025-04-16	14	30	28
Gemini-2.5-Pro-Preview-05-06	60	12	0
DeepSeek-V3-0324	20	14	38
Grok-3-Preview-02-24	30	20	22
GPT-4.5-Preview	10	16	46

Table 1 : Récapitulatif des résultats. (Une égalité peut traduire soit une incohérence dans les jugements du LLM (ex. biais de position), soit une évaluation des deux réponses comme équivalentes).

Les résultats obtenus semblent coïncider avec ceux du classement Arena Score, avec le modèle Gemini-2.5-Pro-Preview-05-06 (Google) arrivant premier dans les deux cas. Cette convergence suggère que le LLM juge a été capable de reproduire un jugement proche de celui observé dans les évaluations humaines utilisées par Arena Score.

6. Résultats

6.1 Classements

La dernière étape consiste à établir un classement permettant d'identifier le modèle le plus performant. Une première méthode simple consiste à calculer un winrate classique, défini comme le rapport entre le nombre de victoires et le nombre total de duels disputés.

o3-2025-04-16 : $14 / (14 + 30 + 28) \times 100 = 19,4 \%$

Gemini-2.5-Pro-Preview-05-06 : $60 / (60 + 12 + 0) \times 100 = 83,3 \%$

DeepSeek-V3-0324 : $20 / (20 + 14 + 38) \times 100 = 27,8 \%$

Grok-3-Preview-02-24 : $30 / (30 + 20 + 22) \times 100 = 41,7 \%$

GPT-4.5-Preview : $10 / (10 + 16 + 46) \times 100 = 13,9 \%$

Toutefois, ce type d'approche présente une limite majeure : elle ne tient pas compte de la qualité des adversaires affrontés. Pour la rigueur des résultats, il est important d'utiliser une méthode fiable de classement. Les systèmes de classement ont été largement étudiés en statistique, notamment via des modèles probabilistes et des méthodes d'optimisation. Néanmoins, tous présentent à la fois des avantages et des limites.

6.1.1 Système de classement Elo

Le système de classement Elo est utilisé pour calculer les niveaux de compétence entre individus dans les compétitions avec plusieurs affrontements entre deux joueurs. Proposé initialement par Arpad Elo ([Elo, 1967](#)) pour les classements d'échecs, il a été adapté à divers sports et jeux de société. L'Elo calcule un score attendu pour un joueur basé sur son classement actuel et celui de son adversaire ([Kelley et Wilson, 2025](#)).

Si le joueur A a une cote R_A et le joueur B une cote R_B , alors le score attendu pour le joueur A est :

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

En raison du facteur d'échelle ci-dessus, une différence de 400 points correspond à une probabilité de 10:1 en faveur du joueur mieux classé.

Après chaque match, la cote du joueur est mise à jour de cette manière :

$$R'_A = R_A + K \cdot (S_A - E_A)$$

S_A est le résultat du match (1 pour une victoire, 0,5 pour un match nul, 0 pour une défaite) et K est un hyperparamètre du système déterminant l'ampleur maximale de la mise à jour après un match.

([Boubdir et al., 2023](#)) expliquent que l'utilisation du système de notation Elo pour évaluer les modèles LLM via des comparaisons pair-à-pair est peu adaptée aux modèles statiques comme les LLM.

D'après les auteurs, il existe deux axiomes fondamentaux pour un bon système d'évaluation : la transitivité (si $A > B$ et $B > C$, alors $A > C$) et la fiabilité (résultats stables malgré l'ordre des matchs ou les hyperparamètres). En simulant différents scénarios, les auteurs montrent que les scores Elo peuvent violer ces principes, surtout lorsque les modèles sont proches en performance. L'ordre des comparaisons et la valeur du facteur K influencent fortement les classements.

La plateforme Arena Score ([Chiang et al., 2024](#)), utilisait à l'origine le système Elo pour la réalisation de ses classements. Toutefois, ce système a été mis à jour pour utiliser le modèle de Bradley-Terry. Les auteurs rapportent que le modèle Bradley-Terry est préférable car la performance des modèles est statique et il fournit de meilleurs intervalles de confiance. Le modèle Elo pèse davantage les résultats récents ce qui le rend moins stable lorsque les modèles comparés sont fixes et que l'historique complet des matchs est disponible.

6.1.2 Glicko

Le système de classement Glicko développé par Mark Glickman ([Glickman, 1995](#)), est une extension du système Elo qui corrige certaines de ses limites. L'innovation principale du système Glicko est l'introduction de la déviation de classement qui mesure l'incertitude liée au classement.

L'Elo suppose une fiabilité constante des notes tandis que Glicko ajuste la sensibilité des variations en fonction de la fiabilité estimée de chaque note. Si un joueur ne joue pas, la perte de confiance dans son niveau estimé (RD) augmente. Les changements de notes ne sont donc plus symétriques, les gains ne sont pas égaux aux pertes de l'adversaire.

Glicko-2 ([Glickman, 2022](#)) ajoute un autre paramètre : la volatilité du classement qui mesure la constance des performances d'un joueur dans le temps.

6.1.3 TrueSkill

TrueSkill ([Herbrich et al., 2007](#)), développé et breveté par Microsoft Research, est un système bayésien de classement des compétences qui généralise à la fois Elo et Glicko pour prendre en charge les compétitions multi-joueurs.

Il modélise la compétence de chaque joueur comme une distribution gaussienne avec une moyenne μ , représentant la compétence perçue, et un écart-type σ représentant l'incertitude.

6.1.4 Modèle de Bradley-Terry

([Chiang et al., 2024](#)) pour le calcul du Arena Score proposent d'utiliser le modèle de Bradley-Terry, un modèle statistique permettant de classer les participants en fonction des probabilités de victoire en tenant compte de la force relative des concurrents.

Chaque élément possède un score de force positive notée π . La probabilité que A batte B est :

$$P(A > B) = \frac{\pi_A}{\pi_A + \pi_B}$$

Plus la force π_A de l'élément A est grande par rapport à celle de B, plus A a de chances de gagner. Si les deux forces sont égales, la probabilité est de 0,5 pour chacun. Si π_A est bien plus grande que π_B , alors A a quasiment toutes les chances de gagner.

Les valeurs de π_A et de π_B sont estimés par la méthode du maximum de vraisemblance. On cherche les valeurs de π qui rendent les résultats observés les plus probables. La fonction de vraisemblance peut s'écrire :

$$L(\pi) = \prod_{\text{tous les duels}} \left(\frac{\pi_A}{\pi_A + \pi_B} \right)^{w_{AB}} \left(\frac{\pi_B}{\pi_A + \pi_B} \right)^{w_{BA}}$$

w_{AB} est le nombre de fois où A a battu B, et w_{BA} le nombre de fois où B a battu A. On maximise ensuite cette fonction numériquement pour trouver les meilleures valeurs de π .

Pour simplifier le modèle, on peut le réécrire comme une équation logistique. On prend $\lambda_A = \log(\pi_A)$, ce qui permet de réécrire le modèle de la façon suivante :

$$\log \left(\frac{P(A \text{ bat } B)}{1 - P(A \text{ bat } B)} \right) = \lambda_A - \lambda_B$$

Pour l'utilisation du modèle de Bradley-Terry, il est nécessaire de créer une matrice de victoire. Cette matrice est présentée en Table 2.

	o3-2025-04-16	Gemini-2.5-Pro-Preview-05-06	DeepSeek-V3-0324	Grok-3-Preview-02-24	GPT-4.5-Preview
o3-2025-04-16	-	0	4	4	6
Gemini-2.5-Pro-Preview-05-06	14	-	18	10	18
DeepSeek-V3-0324	6	0	-	4	10
Grok-3-Preview-02-24	6	0	12	-	12
GPT-4.5-Preview	2	0	4	4	-

Table 2 : Matrice de victoire nécessaire pour le modèle de Bradley-Terry.

Pour calculer le score avec le modèle de Bradley-Terry, nous avons utilisé le logiciel R Studio. Le code est présenté en Figure 8 et les résultats sont présentés en Figure 9.

```
library(BradleyTerry2)
models <- c("o3-2025-04-16", "Gemini-2.5-Pro-Preview-05-06", "DeepSeek-V3-0324", "Grok-3-Preview-02-24", "GPT-4.5-Preview")
w <- matrix(c(
  0, 0, 4, 4, 6,
  14, 0, 18, 10, 18,
  6, 0, 0, 4, 10,
  6, 0, 12, 0, 12,
  2, 0, 4, 4, 0
), nrow = 5, byrow = TRUE,
dimnames = list(models, models))
duels <- data.frame()
for (i in 1:5) {
  for (j in 1:5) {
    if (i != j && w[i, j] > 0) {
      duels <- rbind(duels, data.frame(
        winner = rep(models[i], w[i, j]),
        loser = rep(models[j], w[i, j])
      ))
    }
  }
}
duels$winner <- factor(duels$winner, levels = models)
duels$loser <- factor(duels$loser, levels = models)

bt <- BTm(outcome = rep(1, nrow(duels)),
  player1 = duels$winner,
  player2 = duels$loser)

BTabilities(bt)
```

Figure 8: Code R Studio pour le calcul des scores Bradley-Terry avec l'intégration de la matrice de victoire.

	ability	s.e.
o3-2025-04-16	0.00000000	0.00000000
Gemini-2.5-Pro-Preview-05-06	19.48867780	1348.9839247
DeepSeek-V3-0324	-0.01720228	0.4479379
Grok-3-Preview-02-24	0.64010414	0.4587486
GPT-4.5-Preview	-0.79123138	0.4838585

Figure 9: Scores Bradley-Terry générés par R Studio, à noter que le modèle o3-2025-04-16 est utilisé comme référence dans le calcul, son score est donc fixé à 0.

6.2 Discussions

6.2.1 Pourquoi Gemini domine-t-il ?

Le modèle Gemini-2.5-Pro-Preview a été le modèle le plus performant de cette étude avec un taux de victoire de 83,3 % et un score de Bradley-Terry de 19.4886, bien supérieur à ses concurrents. Cette domination peut s'expliquer par plusieurs raisons. L'une des principales raisons est que Gemini-2.5-Pro-Preview-05-06 (créé par Google) intègre directement la recherche en ligne avec Google Search, ce qui permet d'accéder à des sources en temps réel. Contrairement à cela des modèles comme GPT-4.5-Preview utilisent un outil web distinct, activé uniquement si nécessaire. Une utilité de Google Search dans le contexte de ce mémoire est de vérifier ou compléter l'identification des références GRI, ce qui améliore la réponse du modèle par rapport à ses concurrents.

Gemini excelle également dans la gestion de longues séquences textuelles comme les rapports RSE. La version 2.5 du modèle est capable de traiter jusqu'à 10 millions de tokens avec une précision de 99 %, dépassant largement les limites des autres LLM tels que GPT-4.5-Preview (128k). Un nombre de token de traitement élevé permet de gérer des textes plus longs tels que des rapports RSE tout en gardant un contexte étendu, améliorant la cohérence et la précision des réponses. La limite de sortie tokens pour le modèle Gemini 2.5 Pro Preview est de 65 536 tokens, dépassant là encore largement les limites des autres LLM tels que o3-2025-04-16 (16000 tokens) ou encore GPT-4.5-Preview (4096 tokens). Un nombre de token de sortie élevé permet aux modèles de générer des réponses plus longues, détaillées et complètes sans interruption, ce qui a été apprécié par le modèle juge.

Enfin, la qualité rédactionnelle de Gemini a été valorisée par le LLM juge. Ses réponses étaient davantage structurées, claires et enrichies de références contextuelles.

6.2.2 Implications pour les parties prenantes

Ces résultats peuvent avoir des implications pour les parties prenantes de l'écosystème de l'évaluation ESG. Dans un premier temps, ces résultats prouvent que l'analyse ESG via des modèles LLM est possible, ce qui peut représenter un outil précieux d'aide à la décision en automatisant et fiabilisant l'analyse des rapports RSE. Cela peut permettre de réduire les coûts et inciter des investisseurs à analyser chaque année ces rapports tout en comparant diverses entreprises. Ces résultats permettent également aux investisseurs d'identifier les modèles LLM à privilégier pour l'extraction et la classification d'informations ESG tirées de rapports RSE.

Pour les agences de notation, la capacité des LLM à identifier, extraire et classifier des informations extra financières avec précision et de traiter en quelques minutes des centaines de rapports avec un niveau de cohérence très élevé ouvre la voie à une évaluation potentiellement plus fiable et moins coûteuse. Le recours à ces modèles peut contribuer à réduire les écarts méthodologiques souvent critiqués dans les notations ESG.

6.2.3 Limites

Malgré les apports de ce mémoire, plusieurs limites doivent être soulignées. Premièrement, l'analyse repose sur un corpus limité à 18 rapports RSE d'entreprises. Bien que cette sélection permette d'obtenir des résultats cohérents, elle reste limitée et ne reflète pas toutes les entreprises. Par ailleurs, seulement 5 modèles de langages ont été comparés.

De plus, l'étude repose sur des prompts prédéfinis. Cette approche peut être très sensible aux formulations utilisées. Un changement de formulation, même léger, peut significativement modifier les réponses générées par les modèles.

Bien que des précautions aient été prises pour limiter les biais du modèle juge, il existe un risque que certains biais n'aient pas été totalement éliminés. L'utilisation d'un LLM comme juge est encore récente et expérimentale et reste sujette à des incertitudes méthodologiques.

Enfin, l'étude s'est concentrée uniquement sur deux dimensions des capacités des LLM : l'extraction et la classification d'informations ESG. D'autres aspects importants comme la détection fine du greenwashing n'ont pas été pris en compte dans ce mémoire. Les résultats doivent être interprétés avec prudence dans le temps car les modèles LLM évoluent rapidement et les résultats observés durant ce mémoire peuvent devenir obsolètes à court terme.

6.2.4 Travaux futurs

Les résultats de ce mémoire ouvrent plusieurs pistes pour des recherches futures. La première concerne l'élargissement du corpus analysé. Étendre l'évaluation à un nombre plus important de rapports RSE issus de secteurs, zones et tailles variées, permettrait de tester la performance des modèles LLM dans des contextes plus hétérogènes.

Il serait pertinent d'étendre les tâches confiées aux LLM. Cette étude s'est concentrée sur l'extraction et la classification d'informations ESG. Or, d'autres fonctions comme la détection de greenwashing, l'analyse de sentiment, la détection d'incohérence, l'explicabilité ou encore la prédiction de scores ESG méritent une étude approfondie.

Un axe de recherche prometteur concerne le croisement du jugement de l'humain et de l'IA. Plusieurs travaux soulignent que l'IA seule peut manquer de discernement. Une évaluation où les LLM fourniraient une première analyse que des experts viendraient compléter peut offrir des pistes de travaux intéressants.

6.2.5 Éthique

Ce mémoire souligne des préoccupations éthiques. À quel point peut-on se fier à une IA dont la réponse peut varier à chaque instruction?

Depuis des années, l'évaluation ESG se fait par des humains (agences, analystes, comités...). Ces acteurs utilisent leur jugement, en prenant en compte le contexte et leurs valeurs. L'intelligence artificielle ne juge pas, elle suit des règles sans conscience ni éthique et calcul des probabilités sans comprendre quoi que ce soit. L'absence de jugement éthique et de compréhension profonde peut entraîner des biais qui vont à l'encontre des valeurs éthiques comme privilégier une entreprise européenne plutôt qu'une entreprise issue d'un pays peu développé.

Nous pouvons également nous questionner sur l'impact pour les entreprises. A moyen terme, ces dernières risquent de s'adapter à ce que l'IA valorise, au lieu de garder une approche critique. Cela peut entraîner un écosystème où les entreprises font plus attention à ce que veut l'IA qu'à ce que veulent les investisseurs et parties prenantes.

Conclusion

Dans ce mémoire, nous avons présenté une méthodologie pour comparer les modèles LLM sur les critères ESG. Nous avons réalisé une analyse approfondie de la littérature sur ce sujet. Nous avons présenté les méthodologies de classements de LLM existants puis utilisé la méthode « LLM-as-a-judge » pour comparer 5 modèles LLM sur un corpus de 18 rapports RSE. Nous avons évoqué différentes méthodes de classements puis appliqué le modèle de Bradley-Terry. Enfin nous avons discuté sur les applications de ces résultats, notamment pour les parties prenantes, mais également les potentiels limites de cette étude ainsi que les considération éthique à prendre en compte.

Bibliographie

- Abbas, H. (2021). What is Artificial Intelligence? *IEEE Transactions on Artificial Intelligence*, 2(2), 118-119. <https://doi.org/10.1109/TAI.2021.3096243>
- Alonso, A., & Marqués, J. M. (2019). Financial Innovation for a Sustainable Economy (SSRN Working Paper). <https://dx.doi.org/10.2139/ssrn.3471742>
- Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis. *Annals of Operations Research*, 332, 339-373. <https://doi.org/10.1007/s43546-023-00618-x>
- Bancel, F., & Glavas, D. (2019). Les obligations vertes : un outil au service du financement de la transition écologique. *Revue d'économie industrielle*, 168(4), 83-118. <https://doi.org/10.3917/rindu1.194.0083>
- Bengo, I., Boni, L., & Sancino, A. (2022). EU financial regulations and social impact measurement practices: A comprehensive framework on finance for sustainable development. *Corporate Social Responsibility and Environmental Management*, 29(4), 868-878. <https://doi.org/10.1002/csr.2235>
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings. *The Review of Finance*, 26(6), 1315–1344. <https://doi.org/10.1093/rof/rfac033>
- Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47, 102776. <https://doi.org/10.1016/j.frl.2022.102776>
- Bogojević Arsić, V. (2021). Challenges of Financial Risk Management: AI Applications. *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, 26(3), 1-13. <https://doi.org/10.7595/management.fon.2021.0015>
- Boubdir, M., Kim, E., Ermis, B., Hooker, S., & Fadaee, M. (2023). Elo Uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*. <https://doi.org/10.48550/arXiv.2311.17295>
- Briere, M., Keip, M., Le Berthe, T., & Nuriyev, M. (2022). Artificial Intelligence for Sustainable Finance: Why it May Help (SSRN Working Paper). <https://dx.doi.org/10.2139/ssrn.4252329>
- Buchanan, B. G. (2019). Artificial intelligence in finance. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.2612536>
- Cao, G., Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2021). Understanding managers' attitudes and behavioral intentions towards using artificial intelligence for organizational decision-making. *Technovation*, 106, 102312. <https://doi.org/10.1016/j.technovation.2021.102312>

- Cao, Y., & Zhai, J. (2023). Bridging the gap – The impact of ChatGPT on financial research. *Journal of Financial Services Marketing*, 28(3), 177–191. <https://doi.org/10.1080/14765284.2023.2212434>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*. <https://doi.org/10.48550/arXiv.2403.04132>
- Chiappetta, A. (2023). Navigating the AI frontier: European parliamentary insights on bias and regulation, preceding the AI Act. *Internet Policy Review*, 12(4). <https://doi.org/10.14763/2023.4.1733>
- Chin, A., Fan, Y., & Guan, C. (2022). Natural language understanding, processing, and generation: Investment applications. CFA Institute. URL <https://rpc.cfainstitute.org/sites/default/files/-/media/documents/article/rf-brief/ai-and-big-data-in-investments-Part-II.pdf>
- Commission européenne. (s. d.). Finance durable. URL <https://eur-lex.europa.eu/FR/legal-content/glossary/sustainable-finance.html>
- Cort, T., & Esty, D. C. (2020). ESG Standards: Looming Challenges and Pathways Forward. *Organization & Environment*, 33(3), 395-403. <https://doi.org/10.1177/1086026620945342>
- Cunha, F. A. F. de S., Meira, E., & Orsato, R. J. (2021). Sustainable finance and investment: Review and research agenda. *Business Strategy and the Environment*, 30(5), 2321-2336. <https://doi.org/10.1002/bse.2842>
- Del Vitto, A., Marazzina, D., & Stocco, D. (2023). ESG ratings explainability through machine learning techniques. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-023-05514-z>
- Dimmelmeier, A. (2023). Expanding the politics of measurement in sustainable finance: Reconceptualizing environmental, social and governance information as infrastructure. *Environment and Planning A: Economy and Space*, 56(3), 903-921. <https://doi.org/10.1177/23996544231209149>
- Doddipatla, L. (2025). Sustainable Finance with AI: Leveraging Data-Driven Insights for Green Investments (SSRN Working Paper). <https://dx.doi.org/10.2139/ssrn.5120427>
- Eccles, R. G., & Krzuz, M. P. (2010). One report: Integrated reporting for a sustainable strategy. John Wiley & Sons. <https://doi.org/10.1002/9781119199960>
- Edmans, A., & Kacperczyk, M. (2022). Sustainable Finance. *The Review of Finance*, 26(6), 1219–1232. <https://doi.org/10.1093/rof/rfac069>
- Elo, A. E. (1978). The rating of chess players, past and present. New York: Arco Publishing. URL <https://archive.org/details/ratingofchesspla00unse/mode/2up>

- Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), 3. <https://doi.org/10.3390/sci6010003>
- Giudici, P., & Wu, L. (2025). Sustainable artificial intelligence in finance: impact of ESG factors. *Frontiers in Artificial Intelligence*, 8, 1566197. <https://doi.org/10.3389/frai.2025.1566197>
- Glickman, M. E. (1995). The Glicko system. URL <http://www.glicko.net/glicko/glicko.pdf>
- Glickman, M. E. (2022). Example of the Glicko-2 system. URL <http://www.glicko.net/glicko/glicko2.pdf>
- Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill™: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, 20, 569–576. MIT Press. URL <https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/>
- Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, 39(3), 1957-1988. <https://doi.org/10.1111/1911-3846.12832>
- Kelley, R., & Wilson, D. (2024). Tournament Evaluation of Large Language Models (ICLR 2025 Conference Submission). URL <https://openreview.net/forum?id=5ZpN6W5uRm>
- Lee, J. (2020). Access to Finance for Artificial Intelligence Regulation in the Financial Services Industry. *European Business Organization Law Review*, 21, 731–757. <https://doi.org/10.1007/s40804-020-00200-0>
- Macpherson, M., Gasperini, A., & Bosco, M. (2021). Implications for Artificial Intelligence and ESG Data (SSRN Working Paper). <https://dx.doi.org/10.2139/ssrn.3863599>
- Migliorelli, M. (2021). What Do We Mean by Sustainable Finance? *Sustainability*, 13(2), 975. <https://doi.org/10.3390/su13020975>
- O’Leary, D. E. (2013). Artificial Intelligence and Big Data. *IEEE Intelligent Systems*, 28(2), 96-99. <https://doi.org/10.1109/MIS.2013.39>
- Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics*, 142(2), 572-597. <https://doi.org/10.1016/j.jfineco.2020.11.001>
- Rane, N., Choudhary, S., & Rane, J. (2024). Artificial intelligence driven approaches to strengthening Environmental, Social, and Governance (ESG) criteria in sustainable business practices: a review (SSRN Working Paper). <https://dx.doi.org/10.2139/ssrn.4843215>
- Seele, P., & Gatti, L. (2017). Greenwashing Revisited: In Search of a Typology and Accusation-Based Definition Incorporating Legitimacy Strategies. *Business Strategy and the Environment*, 26(2), 239-252. <https://doi.org/10.1002/bse.1912>

- Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., & Tara, M. (2025). A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25, Article 117. <https://doi.org/10.1186/s12911-025-02954-4>
- UNCTAD. (2024). *World Investment Report 2024: Investment facilitation and promotion*. United Nations Conference on Trade and Development. <https://unctad.org/publication/world-investment-report-2024>
- Visalli, F., Patrizio, A., Lanza, A., Papaleo, P., Nautiyal, A., Pupo, M., Scilinguo, U., Oro, E., & Ruffolo, M. (2023). ESG Data Collection with Adaptive AI. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, 308-315. <https://doi.org/10.5220/0011844500003467>
- Wang, K.-H., Zhao, Y.-X., Jiang, C.-F., & Li, Z.-Z. (2022). Does green finance inspire sustainable development? Evidence from a global perspective. *Economic Analysis and Policy*, 75, 555-568. <https://doi.org/10.1016/j.eap.2022.06.002>
- Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance. *Information Systems Management*, 40(1), 74-90. <https://doi.org/10.1007/s11301-023-00320-0>
- Zhang, M., Shen, Q., Zhao, Z., Wang, S., & Huang, G. Q. (2025). Optimizing ESG reporting: Innovating with E-BERT models in nature language processing. *Expert Systems with Applications*, 263, 125931. <https://doi.org/10.1016/j.eswa.2024.125931>
- Zhao, A., Huang, D., Xu, Q., Lin, M., Liu, Y.-J., & Huang, G. (2024). ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17). <https://doi.org/10.1609/aaai.v38i17.29936>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)* (pp. 46595–46623). <https://doi.org/10.48550/arXiv.2306.05685>
- Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2024). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production*, 447, 144572. <https://doi.org/10.1016/j.jclepro.2024.144572>