

Synthèse de la thèse : GEO – Generative Engine Optimization : du référencement naturel aux moteurs génératifs (ChatGPT/Gemini/Perplexity/Claude etc...)

Résumé exécutif.

La thèse montre que la recherche générative déplace la performance de la page vers la réponse : être cité, bien résumé et correctement attribué.

Elle propose un cadre GEO articulé à une chaîne RAG et un protocole reproductible par variantes contrôlées.

Les résultats soulignent l'importance de l'extractibilité (structure), de la fiabilité (preuves, dates, périmètre) et de la clarté d'entité (attribution).

Les entreprises doivent compléter le SEO par des métriques « réponse » et une gouvernance de test/traçabilité.

Le GEO apparaît comme un prolongement stratégique du SEO, adapté au contexte « zéro clic » et aux moteurs génératifs.

1) Contexte & problématique

Constat. L'essor des assistants conversationnels et des moteurs intégrant des réponses générées (ChatGPT, Gemini, Perplexity, Claude, mais aussi les expériences de recherche enrichies par IA) modifie la nature de la visibilité en ligne. Là où le référencement naturel (SEO) vise classiquement la performance au niveau du document (être exploré, indexé, bien

classé, puis cliqué), les interfaces génératives déplacent une partie de la valeur vers la visibilité « dans la réponse » : être sélectionné comme source, être résumé fidèlement, et être correctement attribué (marque, auteur, organisation).

Problématique. La thèse pose une question centrale : comment optimiser un contenu pour qu'il soit récupéré et utilisé de manière fiable par des moteurs génératifs, sans se limiter aux métriques traditionnelles (positions, impressions, clics) qui deviennent partiellement insuffisantes dans un contexte de consommation « zéro clic ». Elle met en avant trois tensions structurantes : la compression de l'information (peu de sources retenues), la variabilité des systèmes (formats de citations, instabilité, mises à jour) et les risques d'ambiguïté d'entité (confusion de marque, attribution fluctuante, paraphrases inexactes).

Apport original. L'apport de la thèse est de traiter le GEO (Generative Engine Optimization) non comme un ensemble de « trucs » isolés, mais comme un prolongement méthodique du SEO, adapté aux chaînes de traitement de type Retrieval-Augmented Generation (RAG) : récupération de passages, ancrage sur des éléments vérifiables, puis génération.

2) Objectifs de la recherche

La recherche poursuit quatre objectifs complémentaires. **(1) Clarifier conceptuellement** le passage d'un référencement centré « page » (SEO) à une optimisation centrée « réponse » (GEO), en explicitant ce qui change dans la sélection et l'usage des contenus. **(2) Proposer un cadre opérationnel** reliant les étapes techniques et éditoriales (découvrabilité, extractibilité, fiabilité, attribution). **(3) Définir un protocole d'évaluation reproductible** pour mesurer l'effet d'optimisations GEO via des scénarios de requêtes, des variantes de contenus et des métriques simples. **(4) Produire des recommandations actionnables** (checklists, formats, bonnes pratiques) mobilisables par des entreprises et des équipes marketing/contenu.

3) Méthodologie

Approche comparative. La méthodologie s'appuie sur une comparaison entre deux chaînes : une chaîne SEO (exploration-indexation-classement-clic) et une chaîne « générative » de type RAG (récupération-ancrage/grounding-génération).

Variantes contrôlées. Un contenu « baseline » (A) est décliné en versions (A', A'', A''') afin d'isoler l'effet de familles de leviers, tout en conservant la même intention éditoriale. Trois familles de leviers sont explicitement distinguées : **(i) structure et extractibilité** (sections autonomes, définitions courtes, listes, FAQ), **(ii) fiabilité** (chiffres datés, périmètre,

sources, auteur, méthodologie) et (iii) **entités et attribution** (cohérence de nommage, métadonnées, éléments structurants de type données structurées).

Mesures. L'évaluation repose sur des scénarios de requêtes et des métriques pragmatiques : présence en citations/mentions, erreurs et omissions dans la reprise, stabilité d'attribution, et cohérence des éléments clés repris. Une attention particulière est portée à la traçabilité (version du contenu, date/heure, contexte d'outil), indispensable compte tenu de l'évolution rapide et de la stochasticité des systèmes génératifs.

4) Résultats clés

1. Visibilité « dans la réponse » : la visibilité se déplace du document vers la réponse, deux contenus proches en SEO peuvent diverger en GEO lorsque la sélection opère au niveau de passages. **2. Extractibilité :** les contenus structurés en blocs autonomes (définitions, étapes, tableaux simples, FAQ) sont plus facilement récupérés et intégrés, avec moins d'ambiguïtés. **3. Fiabilité :** l'ajout de périmètre, d'hypothèses, de dates, de sources et d'éléments vérifiables améliore la sélection et réduit les déformations. **4. Attribution :** l'attribution (marque/auteur) est un enjeu distinct ; une information peut être reprise sans association correcte à l'entité, d'où l'importance de cohérence de nommage et de métadonnées. **5. Unité de récupération :** le « chunk » (section) devient une unité d'optimisation implicite, ce qui favorise des sections auto-suffisantes et contextualisées. **6. Mesure :** les métriques documentaires (clics/positions) doivent être complétées par des indicateurs orientés « réponse » (citation/mention, fidélité, stabilité). **7. Continuité avec le SEO :** le GEO prolonge le SEO sans le remplacer ; les prérequis techniques de découvrabilité restent nécessaires, mais ne garantissent pas l'usage dans une réponse générée.

5) Implications pour les entreprises

Enjeu. Le GEO revient à piloter une nouvelle « part de voix » : être la source qui structure la réponse, y compris lorsque l'utilisateur ne clique pas.

Implications opérationnelles. Cela conduit à repenser la production éditoriale en formats compatibles avec la synthèse (définitions prêtes à reprendre, procédures en étapes, comparatifs, FAQ, encadrés « limites/périmètre »), à renforcer l'hygiène de confiance (auteur identifié, dates, versions, méthodologie, sources primaires et données contextualisées) et à travailler l'identité d'entité (marque/produit/expert) de façon cohérente sur l'ensemble des contenus pour limiter dilution et confusion. Cela implique également une gouvernance de

test (jeux de requêtes, suivi régulier, archivage des sorties et des versions) adaptée à un environnement instable, et un alignement transverse entre SEO, contenu, data et produit.

6) Limites & perspectives

Limites. Variabilité des modèles et interfaces, opacité partielle des mécanismes de sélection, difficulté d'isoler parfaitement les causes (mises à jour, contextes d'outil), et formats de citations hétérogènes selon les plateformes. Les résultats s'interprètent comme des tendances robustes plutôt que comme des garanties universelles.

Perspectives. La thèse ouvre notamment vers la standardisation de protocoles de test GEO, l'industrialisation de métriques de fidélité/attribution, le rôle croissant des données structurées et de la gestion d'entités, ainsi qu'une stratégie « multi-surfaces » intégrant SEO, réponses génératives et cohérence de marque.

Thesis summary (English version) : GEO – Generative Engine Optimization : du référencement naturel aux moteurs génératifs (Chat GPT/Gemini/Perplexity/Claude etc...)

Executive summary

This thesis shows that generative search shifts performance from the page to the answer : being cited, accurately summarized, and correctly attributed.

It proposes a GEO framework aligned with a RAG pipeline and a reproducible evaluation protocol based on controlled content variants.

Results emphasize extractability (structure), reliability (evidence, dates, scope), and entity clarity (attribution).

Companies should complement SEO with answer-level metrics and a testing/traceability governance model.

GEO emerges as a strategic extension of SEO, adapted to a zero-click context and to generative engines.

1) Context & research problem

Observation. The rise of conversational assistants and engines that provide generated answers (ChatGPT, Gemini, Perplexity, Claude, as well as AI-enhanced search experiences) is changing the nature of online visibility. While traditional SEO targets document-level

performance (being crawled, indexed, ranked, and clicked), generative interfaces move part of the value to visibility *within the answer* : being selected as a source, being summarized faithfully, and being correctly attributed (brand, author, organization).

Research problem. The thesis asks how to optimize content so that it can be retrieved and used reliably by generative engines, without relying solely on traditional metrics (rankings, impressions, clicks) that become partly insufficient in a zero-click consumption context. It highlights three tensions : information compression (few sources retained), system variability (citation formats, instability, updates), and entity ambiguity (brand confusion, unstable attribution, inaccurate paraphrases).

Original contribution. The thesis positions GEO (Generative Engine Optimization) not as a set of isolated tricks, but as a methodical extension of SEO adapted to Retrieval-Augmented Generation (RAG) pipelines : passage retrieval, grounding on verifiable elements, then generation.

2) Research objectives

The research pursues four complementary objectives. (1) **Conceptual clarification** : explain the shift from page-centric SEO to answer-centric GEO, and what changes in source selection and content usage. (2) **Operational framework** : connect technical and editorial stages (discoverability, extractability, reliability, attribution). (3) **Reproducible evaluation** : define a protocol to measure the effect of GEO optimizations through query scenarios, controlled content variants, and simple metrics. (4) **Actionable recommendations** : produce checklists, formats, and best practices usable by companies and marketing/content teams.

3) Methodology

Comparative approach. The methodology compares two pipelines : an SEO pipeline (crawling-indexing-ranking-click) and a generative pipeline of the RAG type (retrieval-grounding-generation).

Controlled variants. A baseline content (A) is transformed into controlled variants (A', A'', A''') to isolate families of levers while keeping the same editorial intent. Three lever families are distinguished : (i) **structure and extractability** (self-contained sections, short definitions, lists, FAQ), (ii) **reliability** (dated figures, scope, sources, author, methodology) and (iii) **entities and attribution** (consistent naming, metadata, structured elements).

Measurements. Evaluation relies on query scenarios and practical metrics : presence as citations/mentions, errors and omissions in the reuse, attribution stability, and coherence of the key elements reproduced. Particular attention is paid to traceability (content version, date/time, tool context), given the fast evolution and stochasticity of generative systems.

4) Key findings

1. Answer-level visibility : visibility shifts from the document to the answer ; two contents can be close in SEO yet diverge in GEO when selection happens at the passage level. **2. Extractability :** structured content in autonomous blocks (definitions, steps, simple tables, FAQ) is retrieved and integrated more easily, with fewer ambiguities. **3. Reliability :** adding scope, assumptions, dates, sources and verifiable elements improves selection and reduces distortions. **4. Attribution :** attribution (brand/author) is a distinct challenge ; correct information can be reused without being associated with the right entity, hence the need for consistent naming and metadata. **5. Retrieval unit :** the section-level *chunk* becomes an implicit unit of optimization, favoring self-contained, contextualized sections. **6. Measurement :** document metrics (clicks/rankings) should be complemented with answer-level indicators (citation/mention, fidelity, stability). **7. Continuity with SEO :** GEO extends SEO rather than replacing it ; technical discoverability prerequisites remain necessary but do not guarantee inclusion in generated answers.

5) Implications for companies

Stake. GEO means managing a new share of voice : becoming the source that structures the answer, even when users do not click.

Operational implications. This requires adapting editorial production to synthesis-friendly formats (ready-to-reuse definitions, step-by-step procedures, comparisons, FAQs, scope/limits boxes), strengthening trust hygiene (identified author, dates, versions, methodology, primary sources and contextualized data), and ensuring coherent entity identity (brand/product/expert) across content to reduce dilution and confusion. It also implies a testing governance model (query sets, regular monitoring, archiving outputs and versions) suited to an unstable environment, and cross-functional alignment across SEO, content, data and product.

6) Limits & outlook

Limits. Model and interface variability, partial opacity of selection mechanisms, difficulty in fully isolating causal factors (updates, tool contexts), and heterogeneous citation formats across platforms. Results should be read as robust tendencies rather than universal guarantees.

Outlook. The thesis points to standardizing GEO test protocols, industrializing fidelity/attribution metrics, increasing the role of structured data and entity management, and adopting a multi-surface strategy that integrates SEO, generative answers, and brand consistency.