

Language-Aware Evaluation of AI-Driven Search: A Benchmark Study of Albanian–English Asymmetry in a Low-Resource Setting

Lazerta Rista

Assistant Lecturer, Faculty of Information Technology,

“Aleksandër Moisiu” University Durrës, Albania

Doctoral Student, University of New York Tirana (UNYT), Tirana, Albania

lazertarista@uamd.edu.al

Abstract. *AI-driven search increasingly mediates information access, yet query language remains underexamined in low-resource settings. AlbSearchBench is a pilot benchmark of 30 matched Albanian–English query pairs tested on Google Search and Gemini across market, institutional, and cultural modules. In Google’s top-five results, Albanian queries surfaced Albanian-owned domains more often than English queries (Local Visibility Rate: 0.73 vs 0.49; Holm-adjusted $p < 0.001$). Across both systems, Albanian-language outputs received higher cultural-grounding ratings (4.77 vs 4.13; $p < 0.001$). Intermediary Saturation (IS) and Institutional-module Official Access Rate (OAR) showed directionally consistent patterns, although neither reached statistical significance. These findings support larger language-aware evaluations.*

Keywords. AI-driven search, language asymmetry, low-resource languages, generative AI, visibility allocation, Albanian language

1 Introduction

Information retrieval is shifting from conventional document retrieval toward AI-driven synthesis (Metzler et al., 2021). By producing direct responses rather than only ranked links, generative systems may reduce users’ search burden while increasing the mediation between users and underlying information sources. This study examines how that mediation is associated with access to local direct providers, public institutions, and culturally grounded information across languages.

An underexplored question is whether these systems provide equivalent access when the same intent is expressed in different languages. English benefits from substantially greater representation in NLP datasets and technologies, whereas languages such as Albanian remain comparatively under-resourced (Joshi et al., 2020; Kastrati & Biba, 2022). These resource imbalances may affect both the visibility and cultural specificity of information returned across languages.

This study investigates whether query language is associated with differences in procedural access, defined here as the pathway through which a search system directs users to information. Specifically, it examines whether equivalent English and Albanian queries differ in the ownership of surfaced access points, local or international,

and in their actor type, direct provider, official source, or intermediary. It also examines whether the cultural fidelity of returned information differs by language. The term procedural asymmetry describes systematic cross-language differences in these routing patterns.

Research Question: To what extent is query language associated with differences in visibility allocation, access pathways, and cultural fidelity in AI-driven search within a low-resource language setting?

The core contributions of this paper are threefold. First, it introduces AlbSearchBench, a matched-query benchmark for evaluating language-conditioned access in AI-driven search across three modules: Market Visibility, Institutional Access, and Cultural/Contextual Grounding. The benchmark combines intent-matched Albanian–English query pairs, a standardized collection protocol, and evaluation across Google Search and Gemini. It further distinguishes between structural visibility analysis based on Google’s top-five results and human assessment of outputs from both systems. Second, the study operationalizes procedural access through a reproducible classification framework that codes each surfaced access point along two dimensions: domain ownership, local or international, and actor type, direct provider, official source, or intermediary. These classifications support three visibility metrics: Local Visibility Rate (LVR), Intermediary Saturation (IS), and Official Access Rate (OAR). LVR is treated as the primary visibility measure because it directly captures whether query language is associated with differences in the visibility allocated to Albanian-owned domains. IS describes intermediary-controlled results, while OAR describes access to official sources and is inferentially evaluated within the Institutional module. Third, the study provides pilot-scale empirical evidence that query language is associated with differences in both procedural access and cultural fidelity. Significant differences are observed for local visibility and cultural grounding, whereas intermediary saturation and Institutional-module official access show directionally consistent but statistically non-significant patterns. Together, these contributions establish a reproducible foundation for larger-scale, language-aware audits of visibility allocation, access pathways, and cultural representation in low-resource language settings.

2 Related Work

Research on fairness in natural language processing has increasingly examined how language technologies may reproduce societal biases and generate representational and downstream harms (Bender et al., 2021; Dev et al., 2022). In multilingual settings, these concerns extend beyond model accuracy. Users expressing equivalent information needs in different languages may encounter differences in information quality, source visibility, access to authoritative resources, and the representation of culturally embedded concepts. However, existing evaluation research has largely concentrated on semantic performance, cross-lingual generalization, and retrieval relevance rather than on whether query language changes the ownership and actor type of the sources surfaced to users.

Multilingual NLP benchmarks have substantially broadened the evaluation of cross-lingual model performance. XTREME evaluates multilingual representations across a range of language-understanding tasks, while XTREME-R extends this framework through more challenging tasks and broader cross-lingual evaluation (Hu et al., 2020; Ruder et al., 2021). These benchmarks provide important evidence about whether models can transfer linguistic knowledge across languages. Nevertheless, their principal unit of analysis is model performance on predefined datasets and tasks. They do not directly examine how operational search systems distribute visibility when the same real-world information need is expressed in different languages.

Information-retrieval benchmarks provide a related but distinct evaluation perspective. BEIR supports heterogeneous zero-shot retrieval evaluation across multiple domains, while MTEB evaluates text embeddings across retrieval, reranking, classification, clustering, and other tasks (Thakur et al., 2021; Muennighoff et al., 2023). Multilingual resources such as mMARCO and MIRACL further extend passage-ranking and ad hoc retrieval evaluation across languages (Bonifacio et al., 2021; Zhang et al., 2023). These resources measure retrieval effectiveness, ranking performance, text representation, and cross-lingual transfer. Their settings, however, differ from the present study because they generally evaluate systems against predetermined document collections and relevance judgments. They do not treat domain ownership, direct-provider access, intermediary dependence, official-source visibility, or cultural grounding as primary evaluation outcomes. Albanian is also not consistently represented across these widely used resources, leaving its behaviour in multilingual retrieval and AI-mediated search comparatively underexamined.

A parallel literature on exposure and ranking fairness provides a conceptual basis for studying visibility allocation. Singh and Joachims (2018) formalize fairness in terms of how exposure is distributed through ranked positions. Biega, Gummadi, and Weikum (2018) examine whether attention is allocated equitably among ranked subjects over time, while Diaz et al. (2020) develop exposure-based approaches for evaluating stochastic rankings. This literature demonstrates that ranking systems do more than organize information: they allocate attention among items, individuals, and providers. However, exposure-fairness research generally examines allocation within a single ranking or language condition. It does not directly address whether changing the

language of an otherwise equivalent query redistributes visibility between locally and internationally controlled access points. The present study extends this perspective to a matched cross-language setting by examining both domain ownership and actor type.

The growing use of generative systems introduces an additional dimension to information-access evaluation. Conventional search presents identifiable links and domains, whereas generative systems may synthesize information into direct responses and increase the distance between users and the underlying sources. As a result, relevance alone may not capture the quality of the resulting access pathway. Usefulness, source transparency, confidence, and cultural grounding also become important, particularly when information is produced in or about a comparatively under-resourced language. AlbSearchBench therefore distinguishes between structural visibility analysis, based on Google's ranked results, and human evaluation of outputs from both Google Search and Gemini.

Research specifically addressing Albanian has documented the language's structural characteristics and persistent scarcity of digital resources, datasets, and NLP tools (Kastrati & Biba, 2022). More recent work has examined how pre-trained language models perform across multilingual tasks and how explainable-AI methods can be used to interpret their behaviour (Kastrati et al., 2025). Within the Albanian digital ecosystem, preliminary studies have also identified growing use of AI in marketing alongside continuing linguistic challenges for local enterprises, particularly in audience-facing activities such as copywriting and local search optimization (Thanasi Biti & Rista, 2025; Rista & Thanasi Biti, 2026a; Rista & Thanasi Biti, 2026b). These studies establish the linguistic and applied context of the present research, but they do not investigate how live search and generative systems distribute access, visibility, and cultural representation between Albanian and English query conditions.

Taken together, previous research provides established methods for evaluating multilingual model performance, retrieval effectiveness, text representation, and ranking exposure. Limited empirical work, however, examines whether intent-matched queries expressed in different languages systematically alter the ownership and actor type of surfaced access points or the cultural fidelity of returned information. This study addresses that gap by shifting the evaluation focus from semantic performance alone to procedural access, connecting multilingual evaluation with visibility allocation, access pathways, and cultural representation in operational AI-driven search.

3 Methodology

AlbSearchBench evaluates whether expressing the same information need in Albanian or English leads AI-driven search systems to surface different access points and provide differently evaluated outputs. The benchmark examines two dimensions. First, it measures structural access in Google's top-five results by classifying each surfaced domain according to actor type, direct provider, official source, or intermediary, and domain ownership, local or international. Second, it evaluates outputs from Google Search and Gemini through multi-rater assessment of relevance, usefulness, source transparency, cultural grounding, and

confidence. AlbSearchBench is organized into four layers: a Query Bank of intent-matched Albanian–English pairs distributed across the Market Visibility, Institutional Access, and Cultural/Contextual Grounding modules; a Result Classification Layer assigning each Google result an actor type and ownership category; a Metric Layer computing Local Visibility Rate, Intermediary Saturation, and Official Access Rate; and an Evaluation Layer applying human assessment, with inter-rater reliability assessed using the intraclass correlation coefficient.

3.1 Dataset Construction

The pilot dataset comprises 30 matched Albanian–English query pairs, with 10 pairs in each of three modules: Market Visibility, covering searches for local commercial services; Institutional Access, covering civic, governmental, and public-service information; and Cultural/Contextual Grounding, covering culturally embedded concepts and meanings. These modules are referred to hereafter as Market, Institutional, and Cultural, respectively. Each query pair was executed on Google Search and Gemini, producing 120 system–language runs. The 60 Google runs were used to compute the structural visibility metrics, while outputs from both systems were included in the human evaluation.

Queries were formulated as short, natural search expressions representing realistic user information needs rather than artificial keyword strings. Market queries reflected local transactional intentions, such as “Hotel në Berat me parkim” / “Hotel in Berat with parking.” Institutional queries addressed practical procedures and access to authoritative information, such as “Si të regjistroj një biznes në Shqipëri” / “How to register a business in Albania.” Cultural queries examined culturally embedded concepts, such as “Çfarë përfaqëson Kanuni në kulturën shqiptare” / “What does the Kanun represent in Albanian culture.” Within each pair, the Albanian and English formulations were matched as closely as practicable in information need, specificity, phrasing style, and length. The paired design was intended to limit differences arising from query construction, allowing observed variation to be interpreted as associated with the query-language condition.

Data collection was conducted in April 2026 from Durrës, Albania, using controlled incognito browser sessions to reduce the influence of search history, stored cookies, and account-based personalization. The same matched query formulations were used across systems, and collection conditions were kept consistent within each pair. The complete dataset and supporting materials are available in the [OSF repository](#).

3.2 Classification Rule and Metrics

Each of the top-five Google results was assigned exactly one actor type according to a strict access-point ownership rule. Classification was based on the organization controlling the surfaced domain rather than on the relevance or origin of the content displayed on it. A verified website controlled by a public institution was classified as Official, including government portals and verified domains of ministries, municipalities, agencies, and other public authorities. A non-governmental entity’s own website was classified as a Direct Provider. Domains controlled by third-party platforms, directories, aggregators, or hosting services, including booking.com, tripadvisor.com, facebook.com,

instagram.com, and youtube.com, were classified as Intermediaries, regardless of whose content appeared on them. When ownership or institutional status was not evident from the domain name, it was verified through the website’s About, contact, institutional, or legal-information pages.

Each result was also assigned one ownership category: local or international. Local denotes an access point controlled by an Albanian organization or institution at the national level; it does not necessarily indicate geographic proximity to the city named in a query. International denotes an access point controlled by an organization based outside Albania. Within each classification axis, the categories were mutually exclusive and collectively exhaustive: every result received one actor-type label and one ownership label.

The rule is deliberately conservative. An Albanian business’s Facebook page or booking.com listing is classified as an international intermediary even when the content concerns a local provider, because the surfaced access point is controlled by a foreign platform. This treatment affects Albanian queries more frequently than English queries: 21 of the 150 Albanian-query results appeared on foreign-controlled platforms, compared with 13 of the 150 English-query results. Accordingly, the ownership-based measure provides a conservative estimate of the Albanian advantage in locally relevant exposure.

Three visibility metrics were calculated for each Google run using the five classified results. Each metric is expressed as a proportion ranging from 0 to 1. Local Visibility Rate (LVR), defined in Eq. (1), is the proportion of the top-five results controlled by Albanian organizations or institutions. It is the primary visibility metric because it directly measures whether Albanian- and English-language queries differ in the visibility allocated to Albanian-controlled access points. Intermediary Saturation (IS), defined in Eq. (2), is the proportion of results controlled by third-party platforms, directories, or aggregators rather than by a provider or official source. Official Access Rate (OAR), defined in Eq. (3), is the proportion of results controlled by a government body, public institution, or official service portal.

$$LVR = (\text{Number of local results of a query}) / 5 \quad (1)$$

$$IS = (\text{Number of intermediary results of a query}) / 5 \quad (2)$$

$$OAR = (\text{Number of official results of a query}) / 5 \quad (3)$$

An LVR of 0.80 indicates that four of the five results are controlled by Albanian actors, whereas an LVR of 0.40 indicates that two are locally controlled. An IS of 0.60 indicates that three of the five results are intermediaries, while an OAR of 0.20 indicates that one result is an official source. A value of 0 means that none of the five results belongs to the relevant category, and a value of 1 means that all five do. Because LVR is derived from the ownership axis, whereas IS and OAR are derived from the actor-type axis, the three metrics are not expected to sum to 1. OAR was calculated descriptively for all Google runs, but its inferential comparison was restricted to the 10 Institutional pairs because official sources are an expected access outcome for civic and public-service queries, not for most Market or Cultural queries.

The visibility metrics were computed only for Google Search because Gemini did not consistently return a fixed ranked list of five source links during the April 2026 collection period. Gemini runs without a complete ranked

source list were excluded from the structural metrics rather than coded as zero. Gemini was instead examined through the human-evaluation results and descriptive source analysis reported in Sections 4.5 and 4.6.

This restriction reflects the observed behaviour of the Gemini 3.1 Flash consumer interface during the collection window. Source links were inconsistent and particularly sparse for general and explanatory queries. A later rerun displayed materially greater citation coverage, indicating temporal variation in source presentation. The findings are therefore interpreted as a snapshot of the system during the recorded period.

3.3 Human Evaluation and Reliability

Three independent raters, a researcher, an SEO specialist, and a business owner, represented analytical, professional, and end-user perspectives. Each independently assessed all 120 system-language runs on five-point scales for relevance, usefulness, source transparency, cultural grounding, and rating confidence. None was an author, and raters could not view one another's scores. The procedure produced 360 ratings per dimension and 1,800 ratings overall.

Predefined criteria guided scoring. Relevance measured whether the output addressed the query; usefulness, whether a user could act on it; source transparency, whether supporting sources were visible and credible; and rating confidence, the evaluator's certainty in the assigned score. Cultural grounding was a composite measure of whether an output reflected the Albanian context through appropriate language, relevant local providers or institutions, and, for cultural queries, accurate cultural meaning. Endpoint anchors ranged from 1 (very poor, absent, or low) to 5 (excellent, clear, or high).

For analysis, the three ratings for each run and dimension were averaged into one run-level score. Inter-rater reliability was assessed with a two-way random-effects, absolute-agreement intraclass correlation model.

Although the three raters were purposively selected to represent distinct evaluative perspectives, analytical, professional, and end-user, rather than randomly sampled from an undifferentiated rater pool, the random-effects specification was retained because the intended target of inference was the reliability of an equivalently constituted evaluation panel rather than agreement among these three specific individuals. ICC(2,3) therefore estimates the reliability of the mean judgment produced by the three-rater panel under this generalization assumption. Because only one rater represented each perspective, however, individual-rater effects cannot be separated from perspective-specific effects, and generalization beyond similarly constituted panels should be made cautiously.

Both single-measure reliability, ICC(2,1), and average-measure reliability, ICC(2,3), were calculated. ICC(2,1) estimates the reliability of one rater, whereas ICC(2,3) estimates the reliability of the mean across all three raters. The coefficients are reported in Section 4.1.

No formal calibration round preceded independent scoring. The strong reliability estimates reported in Section 4.1, particularly for relevance, cultural grounding, and usefulness, indicate substantial convergence under the predefined criteria. Nevertheless, a calibration session is planned for the full-scale study so that shared interpretation can be established prospectively rather than inferred solely from post hoc reliability estimates.

3.4 Statistical Analysis and Interpretation

Matched conditions were compared using the Wilcoxon signed-rank test (Woolson, 2007), which does not require normally distributed measurements. Language comparisons paired the Albanian and English versions of the same query within a system; system comparisons paired Google Search and Gemini for the same query and language. The test ranks non-zero paired differences and evaluates whether they are systematically centered above or below zero. Statistical significance was assessed at $p < 0.05$.

For descriptive language comparisons, differences were calculated as SQ minus EN. Positive differences therefore favour Albanian, while zero differences are ties and are omitted from the signed-rank calculation. Rank-biserial correlation, r , was reported as the effect size. Because the statistical output used the English-minus-Albanian contrast, negative r values favour Albanian. Absolute values below 0.10 were treated as negligible, 0.10-0.29 as small, 0.30-0.49 as medium, and 0.50 or above as large.

The Hodges-Lehmann estimate expresses the typical paired shift in the original units of a metric. For LVR, a shift of 0.20 corresponds to one additional Albanian-controlled result among five. Holm correction was applied to three prespecified structural comparisons: overall LVR, overall IS, and Institutional-module OAR, controlling familywise error (Holm, 1979). Descriptive patterns with $p \geq 0.05$ are reported as directionally consistent but statistically non-significant, rather than as evidence of no difference.

Human-evaluation means were calculated from the run-level averages. ICC values were interpreted following Koo and Li (2016): below 0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and above 0.90, excellent reliability.

4 Results

The results are presented in six subsections. Inter-rater reliability is reported first, followed by the structural visibility and human-evaluation findings. Two cross-language effects reached statistical significance at pilot scale: Local Visibility Rate and cultural grounding. Intermediary Saturation and Institutional-module Official Access Rate showed directionally consistent but statistically non-significant differences. Cross-system comparisons revealed task-dependent usefulness and lower source transparency for Gemini.

4.1 Inter-rater Reliability

Inter-rater reliability ranged from moderate to excellent for individual raters and from good to excellent for the mean of three raters (Table 1). Average-measure reliability was excellent for relevance, cultural grounding, and usefulness, and good for rating confidence and source transparency.

Table 1. Inter-rater reliability across 120 runs and three raters.

Dimension	ICC(2,1)	ICC(2,3)
Relevance	0.92	0.97
Cultural grounding	0.996	0.999
Usefulness	0.78	0.91
Rating confidence	0.71	0.88
Source transparency	0.58	0.80

The near-perfect cultural-grounding coefficient indicates highly consistent ratings under the applied criteria. It should nevertheless be interpreted cautiously because Albanian-language outputs were concentrated at the upper boundary of the scale, creating a ceiling effect and limiting discrimination among those outputs.

4.2 Local Visibility Rate

Across all 30 matched pairs, Albanian queries produced a substantially higher mean LVR than English queries (0.73 vs 0.49; $W=5.0$, $p<0.001$; rank-biserial $r=-0.95$; Hodges–Lehmann shift=0.20, 95% CI [0.10,0.30]). The difference remained statistically significant after Holm correction across the three prespecified structural comparisons (adjusted $p<0.001$). Eighteen pairs favoured Albanian, one favoured English, and 11 were tied.

Mean LVR was descriptively higher for Albanian in every module (Table 2). Module-level values are reported to show the consistency of direction across the 10 pairs in each module; inferential testing is reserved for the prespecified overall LVR comparison.

Table 2. Local Visibility Rate by module for Google Search (10 pairs per module; module values descriptive).

Module	SQ	EN
Market	0.78	0.54
Institutional	0.78	0.52
Cultural	0.64	0.40
All 30 pairs	0.73	0.49

A worked example illustrates how query language affected both ownership and actor type. For "Si të aplikoj për pasaportë shqiptare" / "How to apply for an Albanian passport," all five Albanian results were locally controlled, including four official sources. The English results contained two local official sources and three international intermediaries (Table 3).

Table 3. Worked example: top-five Google results for the passport query pair.

Rank	English SERP	Albanian SERP
1	e-albania.al (Official/L)	e-albania.al (Official/L)
2	futurepassports.com (Intermediary/Int)	identitek.al (Direct/L)
3	global-relocate.com (Intermediary/Int)	e-albania.al (Official/L)
4	ambasadat.gov.al (Official/L)	ambasadat.gov.al (Official/L)
5	wikipedia.org (Intermediary/Int)	punetegashtme.gov.al (Official/L)

For this pair, LVR was 1.00 for Albanian and 0.40 for English; IS was 0.00 and 0.60; and OAR was 0.80 and 0.40, respectively. In Table 3, L denotes local ownership and Int denotes international ownership.

4.3 Intermediary Saturation and Official Access Rate

Across all 30 pairs, mean IS was higher for English than Albanian queries (0.57 vs 0.51; $W = 36.0$, $p = 0.10$, $r = 0.47$; Holm-adjusted $p = 0.21$). The difference was directionally consistent but statistically non-significant.

OAR was evaluated inferentially only within the Institutional module, where official sources are a substantively expected access outcome. Albanian queries produced a higher mean OAR than English queries (0.44 vs 0.30; $r = -0.54$, $p = 0.30$; Holm-adjusted $p = 0.30$), but the difference was not statistically significant. Market and Cultural OAR values were near zero in both languages and are retained descriptively in Table 4 because those query types rarely call for official sources.

Table 4. IS and OAR by module for Google Search; only Institutional OAR is tested inferentially.

Module	IS SQ	IS EN	OAR SQ	OAR EN
Market	0.28	0.40	0.00	0.00
Institutional	0.28	0.44	0.44	0.30
Cultural	0.98	0.86	0.00	0.04
All 30 pairs	0.51	0.57	0.15	0.11

Metric behaviour also reflected module structure. In the Cultural module, almost all results in both language conditions were intermediaries, producing near-ceiling IS values of 0.98 for Albanian and 0.86 for English. LVR was less affected because it derives from the separate ownership axis.

4.4 Cultural Grounding and Conceptual Flattening

The second significant cross-language effect concerned cultural fidelity. Across all 60 matched query-system units, Albanian-language outputs received higher mean cultural-grounding ratings than English-language outputs (4.77 vs 4.13; $p < 0.001$; $r = -0.81$).

Within the Cultural module, Albanian outputs averaged 5.00 compared with 3.45 for English, and every matched unit favoured Albanian ($r = -1.00$). The gap was larger for Google Search than for Gemini (Table 5).

Table 5. Cultural-grounding ratings within the Cultural module (10 pairs per system).

System	SQ	EN
Google Search	5.00	2.90
Gemini	5.00	4.00

Qualitative inspection of the structured issue log illustrates a pattern termed conceptual flattening, understood here as the reduction of culturally specific ambiguity, variation, or context in the English-language condition. For the Zana query, the principal difference concerned characterization rather than translation: the English output more strongly emphasized benevolent and protective fairy-like roles, whereas the Albanian output more explicitly retained the figures' ambivalence, presenting them as beautiful and protective but also potentially dangerous or punitive. A comparable difference appeared in the black-cat query. The English output stated more generally that black cats are commonly considered bad luck in Albania, while the Albanian output qualified the belief as mixed, partly borrowed or urban, and culturally variable. In the traditional-food query, the English output provided a general list of Albanian dishes, desserts, and drinks, whereas the Albanian output connected particular foods with specific holidays and celebrations. These observations show how an

output may remain broadly relevant while losing local variation, symbolic ambiguity, or occasion-specific meaning. They are used to interpret, rather than independently establish, the statistical cultural-grounding result.

4.5 Task-Dependent Usefulness

The comparison between Google Search and Gemini revealed a task-dependent usefulness pattern. Usefulness scores were pooled across the Albanian and English conditions within each module, producing 20 matched query-language units per module. Relevance was consistently high for both systems (Gemini 5.00; Google 4.86), while usefulness differed by module (Table 6). For Market queries, Google was rated more useful (5.00 vs 3.52; $p < 0.001$; $|r| = 1.00$, favouring Google in all 20 matched units), frequently providing actionable elements such as map listings, telephone details, opening hours, and directions.

For Cultural queries, Gemini was rated more useful (4.70 vs 3.60; $p < 0.001$; $|r| = 1.00$, favouring Gemini in all 20 matched units), reflecting the value of synthesized responses for cultural and explanatory information needs. Institutional queries also favoured Google (4.87 vs 4.00; $p < 0.001$; $|r| = 1.00$, favouring Google in all 20 matched units), as these tasks often required direct access to an official portal or service.

Table 6. Mean usefulness ratings by module and system (20 query-language units per module; 60 overall).

Module	Google	Gemini
Market	5.00	3.52
Institutional	4.87	4.00
Cultural	3.60	4.70
All runs	4.49	4.07

Although the overall mean favoured Google, the module-level results reveal a crossover pattern. Google was significantly more useful for actionable Market and Institutional queries, whereas Gemini was significantly more useful for Cultural queries requiring synthesized explanation.

4.6 Source Transparency and Verification Risk

Unlike usefulness, source transparency consistently favoured Google. Gemini received a substantially lower mean rating than Google (1.96 vs 4.07; $p < 0.001$; $|r| = 1.00$, favouring Google in all 60 matched comparisons). Raters also reported lower rating confidence for Gemini outputs than for Google outputs (3.06 vs 4.28; $p < 0.001$; $|r| = 0.89$). Because source transparency showed moderate single-rater reliability ($ICC(2,1) = 0.58$) but good reliability for the averaged three-rater score ($ICC(2,3) = 0.80$), this comparison is interpreted using the run-level mean ratings.

The structured issue log records reduced or distorted treatments of culturally specific concepts and English result sets relying on low-specificity sources. These cases do not estimate a general hallucination rate; they show that sparse or weak sourcing increases the verification burden and makes unsupported reduction or distortion harder to detect.

Within Gemini, no visible sources were provided in 23 of 30 Albanian runs, compared with eight of 30 English runs. Google source-transparency ratings were nearly

identical across languages (4.08 vs 4.06). A later rerun displayed materially greater citation coverage, so these findings describe Gemini 3.1 Flash during the April 2026 collection window rather than a stable property of later versions.

5 Discussion

The findings show that query language is associated with both the structure of information access and the cultural fidelity of returned information. Albanian queries allocated more visibility to Albanian-controlled domains, whereas English queries more often routed users through internationally controlled access points. Albanian-language outputs also preserved culturally embedded meaning more consistently. These patterns do not imply intentional bias or establish a causal mechanism; they describe systematic associations within a matched pilot dataset.

The practical implications differ by module. In local service search, reduced visibility for Albanian-controlled providers may shift attention and economic value toward international platforms. In institutional search, the Institutional-module OAR comparison suggests that query language may affect how directly users reach authoritative sources, although the difference was not statistically established. For cultural information, conceptual flattening can obscure distinctions that are central to local knowledge. Generative synthesis offered greater usefulness for explanation but imposed a higher verification burden, with the source-omission pattern appearing more pronounced in Albanian during the collection window.

Methodologically, AlbSearchBench separates two analytical layers. The structural layer identifies who controls the surfaced access point through reproducible ownership and actor-type classification. Its strength is consistency; its limitation is that a local provider hosted on a foreign platform is coded as international intermediation. The human-evaluation layer provides an initial assessment of what outputs deliver in terms of relevance, usefulness, transparency, and cultural grounding. Together, the layers extend multilingual evaluation beyond semantic accuracy toward visibility allocation and procedural access.

6 Limitations

Several constraints limit generalization. First, the pilot contains 30 matched pairs, 10 per module, and is designed to establish the framework and detect large effects rather than provide a census of search behaviour. The secondary overall IS and Institutional-module OAR comparisons are limited by the pilot sample and are therefore reported cautiously as directional. Second, the data form a cross-sectional snapshot from April 2026. Gemini 3.1 Flash returned sources inconsistently, and a later rerun showed greater citation coverage; system-level transparency findings are therefore bounded to the collection window. Third, all queries were executed from Durrës, Albania, so language is not fully separated from a localized IP and search environment. Fourth, all queries targeted Albanian providers, institutions, or cultural content. The design therefore cannot fully distinguish an effect of query language itself from an effect of alignment between the query language and the target

content. Future work should include reverse-case queries, such as Albanian-language queries for non-Albanian targets, to separate these mechanisms.

The ownership rule classifies access points rather than page content. This improves reproducibility but treats an Albanian provider's page on a foreign platform as international intermediation. The conceptual-flattening examples are illustrative and were not subjected to an independent qualitative coding scheme. Cultural-grounding scores also showed a ceiling effect for Albanian outputs, while relevance was maximal for one system, limiting discrimination. Because each evaluator perspective was represented by one rater, generalization of the random-effects ICC to comparable evaluators remains provisional. Finally, the five-point criteria used endpoint anchors rather than fully specified descriptors for every scale point; a more granular anchored protocol is planned for subsequent work.

7 Conclusion

AlbSearchBench identifies two significant cross-language asymmetries in this low-resource setting. Albanian queries produced greater visibility for Albanian-controlled domains in Google's top-five results, and Albanian-language outputs received higher cultural-grounding ratings across Google Search and Gemini. Overall Intermediary Saturation and Institutional-module Official Access Rate followed directionally consistent patterns but did not reach statistical significance at pilot scale.

The cross-system findings further show that usefulness depends on the task: Google was significantly more useful for actionable Market and Institutional queries, whereas Gemini was significantly more useful for Cultural explanation. This explanatory value was accompanied by lower source transparency and greater verification difficulty during the recorded Gemini 3.1 Flash collection window. Evaluating AI-driven search through semantic relevance alone would therefore miss how language affects the ownership of visible access points, the route to authoritative information, and the preservation of culturally specific meaning.

The benchmark provides a reproducible basis for larger language-aware audits. Future work will expand the query bank; add reverse-case tests, including Albanian-language queries for non-Albanian targets, to separate query-language effects from language-content alignment; introduce a coded content-level analysis; test additional systems and collection locations; and monitor temporal changes in ranking and source presentation.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623.
- Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. *Proceedings of SIGIR 2018*, 405-414.
- Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., & Nogueira, R. (2021). mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*.
- Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On measures of biases and harms in NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*.
- Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. *Proceedings of CIKM 2020*, 275-284.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Proceedings of ICML 2020*.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the ACL*, 6282-6293.
- Kastrati, M., & Biba, M. (2022). Natural language processing for Albanian: A state-of-the-art survey. *International Journal of Electrical and Computer Engineering*, 12(6), 6432-6439.
- Kastrati, M., Imran, A. S., Hashmi, E., Kastrati, Z., Daudpota, S. M., & Biba, M. (2025). Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 149, 110136.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163.
- Metzler, D., Tay, Y., Bahri, D., & Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *ACM SIGIR Forum*, 55(1), 1-27.
- Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014-2037.
- Rista, L., & Thanasi Biti, T. (2026a). The strategic integration of artificial intelligence in digital marketing: Global case studies and implications for the Albanian emerging market. *Interdisciplinary Journal of Research and Development*, 13(1), 177. <https://doi.org/10.56345/ijrdv13n120>
- Rista, L., & Thanasi Biti, T. (2026b). A real-world mixed-methods investigation across service and e-commerce sectors in Albania. *Proceedings of the 16th International Scientific Conference of Business Faculty (ISCBF 2026)*, 419-429.

- Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10215-10245.
- Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. *Proceedings of KDD 2018*, 2219-2228.
- Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.
- Thanasi Biti, T., & Rista, L. (2025). Leveraging AI for personalized digital marketing campaigns in Albania. *Interdisciplinary Journal of Research and Development*, 12(3 S1), 73. <https://doi.org/10.56345/ijrdv12n3s110>
- Woolson, R. F. (2007). Wilcoxon signed-rank test. *Wiley Encyclopedia of Clinical Trials*.
- Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11, 1114-1131.