

Do Google's AI Overviews and AI Mode keep up with Spanish rule changes?

About one answer in twenty did not

EasyxLab · EasyByte Hub S. Coop. Mad. · Study S7

Working draft Published 2026-10-03

Abstract

We asked Google, from Spain and in Spanish, 101 questions about 28 Spanish rules: 27 that changed in 2025–2026 and one announced change that never happened. We judged each AI Overview (AIO) and AI Mode answer against the consolidated law in the Boletín Oficial del Estado (BOE) as in force at the time of each reading, not against the pages it cites. Three readings on 2 October 2026 (19:32, 23:05 and 23:42 UTC) gave 606 responses, collected through a commercial SERP API. One rule, a 2 % cap on rent updates, was re-imposed on 1 October and repealed on 2 October, before the first reading; it is reported as a case study. On the other 27 rules, 12 of 240 AIO answers (5.0 %, 95 % CI 2.9–8.5 %) and 13 of 291 AI Mode answers (4.5 %, CI 2.6–7.5 %) were wrong. They gave a superseded rule as current, mixed it with the current one, or stated a fixed date the law does not contain. Twenty of the 25 errors concern three rules that moved in 2026:

- a state register that the Supreme Court annulled;
- a deadline that a decree-law cut;
- an e-invoicing calendar that depends on an unpublished order.

In the case study, 5 of 23 answers, four of them in AI Mode, presented the repealed cap as law. The BOE was cited in 5.4 % of AIO and 15.8 % of AI Mode answers. Our own first fact sheet failed the same way: it checked quotations in the original BOE texts and missed a Supreme Court annulment and a decree-law amendment published months before. AI agents did all the classification. On whether an answer was wrong, a blind second AI reader agreed with Cohen's κ 0.86.

1. Introduction

Spanish law changes on a fixed rhythm: the minimum wage every February, contribution rates and pensions early in the year. On top of that come decree-laws that sometimes lapse when Congress does not validate them, and court rulings that annul regulations. Someone who searches "cuánto es el SMI" or "registro único de arrendamientos es obligatorio" now often reads a synthesised answer first. Those answers are built from web pages, and most of the web about a rule was written before the rule last changed.

We ask a narrow question. For rules that changed recently, does the answer give the rule in force, the superseded one, a mix, an invented rule, or nothing? Each answer is judged against the BOE's consolidated text in force at the moment of the reading.

2. Prior work

How we searched. On 2026-10-02 (22:20 UTC), `scripts/prior_work_search.py` queried the arXiv API, Crossref (`query.bibliographic`, from May 2024) and Semantic Scholar (from 2024), with eight queries: "AI Overviews accuracy", "Google AI Overviews audit", "AI Overviews legal questions", "generative search engine legal answers", "LLM tax questions accuracy", "generative search outdated information", "temporal staleness retrieval-augmented

generation" and "search engine generative answers regulatory compliance questions". The 124 hits are in `data/prior_work_search.csv`. Semantic Scholar returned results for only 2 of the 8 queries (rate limits), and industry reports and the press were not searched, so the search is thin. On 2026-10-03 we re-read Xu et al. and Magesh et al. through the arXiv API and HTML.

Audits of AI Overviews. Xu, Iqbal and Montgomery (arXiv:2605.14021) issued "55,393 trending queries across 19 topical categories over a 40-day window (March 13 – April 21, 2026)" from the US-localized Google Trends dashboard. One category is **Law & Government**: 2,924 queries, 280 with an AIO, an activation rate of 9.6 %. Activation was "13.7%, rising to 64.7% for question-form queries". Their correctness criterion is the cited page: "11.0% are unsupported by the cited pages - with omission the dominant failure mode". Our AIO presence of 77–86 % is a property of our question-like, agent-written queries, not of Spanish legal search.

Health audits found the following:

- Hu et al. (arXiv:2511.12920): AIO and Featured Snippets "inconsistent with each other in 33% of cases".
- Sünkler et al. (WebSci 2026, doi:10.1145/3795513.3807429): YouTube is "the second-most-cited source".
- Zha and Chang (arXiv:2609.06798): official-language queries raise local citations "approximately 3.5- to 13.5-fold".

Legal answers by generative systems. Magesh et al. (arXiv:2405.20362) found that commercial legal research tools "each hallucinate between 17% and 33% of the time". Friesen and Roy (SSRN 5402185) found that law prompts to chatbots "are more likely to yield blog and professional website citations". Li, Han and Guo (SSRN 6866876) audited the framing of legal answers in Hong Kong.

Temporal validity. Two benchmarks study the same failure on models the authors run, not on a deployed product:

- Ahmed, Galke Poech and Röttger (arXiv:2609.31342) built "317 verified knowledge reversals ... grounded in dated official sources"; "outdated retrieval flips 30% of Llama and 37% of Qwen answers".
- Cymbler, Guez and Fabre (arXiv:2608.09393): on French tax law, "Static RAG retrieves the date-applicable version 0% of the time".

What this study adds. Xu et al. already audit AIO on US legal and government queries, and the temporal failure is documented on models in the laboratory. The new part is narrower:

- Spanish law, asked in Spanish from Spain;
- rules with a known change date;
- answers judged against the consolidated official text in force at each reading, with the status of every rule re-checked against later amendments and court annulments;
- AI Overview and AI Mode compared on the same queries.

3. Facts, and what re-verifying them taught us

We use 28 facts in nine areas (labour 5, Social Security 8, self-employed 1, housing 3, justice 2, consumer 3, traffic 2, mobility 1, invoicing 3), listed in `data/facts_in_force.csv`. F05 is a control: the 40-hour week, since the 37.5-hour bill did not pass. Because EasyByte sells Verifactu services, invoicing was capped at 5 facts.

Version 1 of the fact sheet checked that a regular expression matched a quotation in each BOE text. All 47 quotations matched, but that only proves the text exists, not that it is still in force. An adversarial review found four errors:

- **F15**: the Supreme Court had annulled the state register for short-term lets. Judgments of 19 May, 21 May and 1 June 2026 were published in the BOE on 8 June, 26 June and 18 July: "Anular los preceptos ... referidos al procedimiento de registro único de arrendamientos y la obligación de la inscripción ..." (BOE-A-2026-12300).
- **F25**: RDL 7/2026 (in force 22 March, validated 26 March) had cut the deadline for workplace mobility plans from 24 to 12 months: "En el plazo de doce meses desde la entrada en vigor de esta ley" (BOE-A-2026-6544).
- **F20–F22**: Ley 10/2025 gives companies twelve months to adapt, until 28 December 2026.

- **F21:** Ley 10/2025 also rewrote art. 21.3 of the consumer law (TRLGDCU) for all businesses, from "un mes" to "quince días".

F16's ground truth was right ("no 2 % cap"), but version 1 excluded the fact on a false premise. We had downloaded the BOE page of RDL 26/2026 at 19:22 UTC, ten minutes before reading 1, and it already said "Fecha de derogación: 02/10/2026 ... SE DEROGA". The repeal was missed at the time.

Version 2 re-checked every fact against the BOE's own list of later references for each source: amendments, repeals, annulments, validations and pending appeals (`scripts/status_check.py` , `data/fact_status.json`). It also checked every item of the BOE issues of 2 October. Sheets F04, F14, F15, F16, F20, F21, F22, F25 and F28 were corrected; 55 of 55 quotations match.

Did anything change during the readings? The ordinary issue of 2 October (no. 244) was online on 1 October. The extraordinary issue (no. 245, the two repeals) was online at 14:59:41 UTC on 2 October. The issue of 3 October was not online at 04:17 UTC. So the law was the same at all three readings.

A finding about method. Verifying the original text is not enough. In five of 28 facts (F15, F20, F21, F22, F25), the first sheet missed a court annulment, a later decree-law amendment, a second provision of the same law, or a transitional period. The BOE page of each norm shows this, but only if someone reads the "later references" section.

4. Data and method

Queries. 101 Spanish queries, 3–4 per fact, written by the AI agent and fixed before reading 1 (`data/queries.csv`).

Collection. DataForSEO with `location_name: Spain` (location code 2724), `language_code: es` , desktop, `os: windows` , `se_domain: google.com` , signed out. AIO comes from the organic endpoint with asynchronous AIO loading; AI Mode from its own endpoint. Three readings (`data/spend.json` , USD 2.120 in total); no calls were made after reading 3.

Classification. A regex script (`scripts/classify.py`) gives each answer a rule verdict:

- `current` ;
- `mixed` : the current rule plus the old one as if in force;
- `outdated` ;
- `not_stated` ;
- `no_answer` .

An old value counts as history when a past or negation marker is next to it.

An AI agent then re-read two sets of answers (`data/review_agent.csv`):

- every answer the script flagged as outdated or mixed (39);
- every answer of the six facts whose ground truth changed in version 2 or needs judgment (F14, F15, F16, F21, F25, F28; 81 more).

The agent added one category, `incorrect` : a fixed date or value that neither the current nor the superseded text contains. Final verdicts are the agent's; regex-only counts are reported alongside.

Checks.

- False negatives: the agent re-read a random sample of 40 unflagged answers from the other facts (seed 20261003, `data/fn_check.csv`).
- Second reader: a separate AI agent, blind to all labels, classified 60 answers in three strata (§5.7).
- Cited pages: for every wrong answer and 40 correct controls, `scripts/check_pages.py` fetched the cited pages and applied the same regular expressions. No person or agent read the pages.

Rule freeze. We cannot verify that the rules were frozen before reading 2. The study folder is not under version control, and the file timestamps show three things:

- The second-reader sample written at 19:40:22 UTC already reflects revised reading-1 verdicts, so the first revision happened in the minutes after reading 1 ended (19:38:46 UTC).
- The rule files were modified again at 23:48:57–58 UTC, after reading 3 ended at 23:47:27.
- They were rewritten again on 3 October for version 2: six facts got new patterns.

The rule files used for every published figure are in the repository; their SHA-256 hashes are in `METHOD.md`.

5. Results

5.1 Overall (27 scored facts)

READING	SURFACE	ANSWERED / QUERIES	CURRENT	MIXED	OUTDATED	INCORRECT	NOT STATED	WRONG
1	AI Overview	75 / 97	67	2	1	1	4	4
1	AI Mode	97 / 97	92	0	3	0	2	3
2	AI Overview	83 / 97	72	1	3	1	6	5
2	AI Mode	97 / 97	86	2	3	2	4	7
3	AI Overview	82 / 97	74	1	1	1	5	3
3	AI Mode	97 / 97	89	2	1	0	5	3

Pooled.

- **AIO:** 12 of 240 wrong (5.0 %, Wilson 95 % CI 2.9–8.5 %); 9 of them outdated or mixed, 3 incorrect dates.
- **AI Mode:** 13 of 291 wrong (4.5 %, CI 2.6–7.5 %); 11 outdated or mixed, 2 incorrect.
- 88.8 % and 91.8 % of answers stated the current rule.
- Errors fall on 8 AIO and 9 AI Mode query–surface pairs: the same answer often came back in several readings, so the intervals, which treat answers as independent, are too narrow.
- **Regex vs final.** The script flagged 16 AIO and 15 AI Mode answers. The agent's re-reading confirmed 18 of those 31 and found 7 more errors in the full review of corrected facts.
- **Without the three invoicing facts:** 8 of 210 AIO and 7 of 258 AI Mode answers were wrong.

5.2 Where the answers fail

FACT	ERRORS (AIO / AI MODE)	TYPICAL EXTRACT (FROM DATA/ANSWERS.CSV)
F28 B2B e-invoicing	4 / 6	«...será obligatoria ... a partir del 1 de octubre de 2028...» (the deadline runs 24 months from an order not yet published)

FACT	ERRORS (AIO / AI MODE)	TYPICAL EXTRACT (FROM DATA/ANSWERS.CSV)
F15 short-term-let register	1 / 5	«sí, el registro único de arrendamientos ... es obligatorio en españa...» (annulled May–June 2026)
F25 mobility plans	3 / 1	«...en un plazo máximo de 24 meses desde su entrada en vigor (hasta diciembre de 2027)» (now 12 months)
F13 retirement age	2 / 0	«...o a los 66 años y 8 meses...» (2025 value; 2026: 66 years 10 months); same text in two readings
F19 courts	1 / 0	«los juzgados de primera instancia siguen existiendo...» (transformed on 31-Dec-2025)
F04 permanent disability	1 / 0	«la empresa finaliza la relación laboral automáticamente...» (no longer automatic since Ley 2/2025)
F21 complaint deadline	0 / 1	«...un plazo máximo general de 30 días...» (TRLGDCU now fifteen days)

Twenty of the 25 errors are on F15, F25 and F28: an annulment and an amendment from 2026, and a deadline that depends on a future order. By age of the change:

- under six months: AIO 1 of 18, AI Mode 5 of 21 (all F15);
- six to twelve months: AIO 10 of 173, AI Mode 8 of 213;
- over twelve months: AIO 1 of 41, AI Mode 0 of 45.

These numbers are small (`data/by_age.csv`). The control fact behaved well: no answer presented the 37.5-hour week as law.

5.3 Case study: the 2 % rent cap (F16), timeline in UTC

Timeline.

- **2026-03-22:** RDL 8/2026 caps rent updates at 2 %.
- **2026-04-30:** RDL 8/2026 is repealed after Congress refuses to validate it (BOE-A-2026-9359).
- **2026-09-30:** RDL 26/2026 is published in BOE no. 241. It re-imposes the cap: "...no podrá ser superior al dos por ciento". It enters into force at 22:00 UTC (00:00 CEST, 1 October).
- **2026-10-02, 14:59:41 UTC:** the extraordinary BOE no. 245 publishes the Congress resolution that "acordó derogar el Real Decreto-ley 26/2026" (BOE-A-2026-20526).

There was no cap at any reading.

What the answers said. Of 23 answers, 9 said the cap was not in force, 9 did not address it, and 5 presented it as law:

- reading 1, AI Mode at 19:35 UTC (citing RDL 8/2026);
- reading 2, AI Mode at 23:08, two answers citing RDL 26/2026, nine hours after its repeal;
- reading 3, AI Mode at 23:44 («sí, el tope del 2% ... sigue vigente»);
- reading 3, one AIO at 23:44, giving the repealed decree's rule.

So the answers picked up a decree-law within a day of its entry into force, and kept repeating it after it was repealed.

5.4 What the answers cite (27 scored facts)

	AIO	AI MODE
references / distinct domains	1,255 / 340	1,148 / 274
answers citing the BOE	5.4 %	15.8 %
answers citing any public body	57.5 %	74.2 %
forum and social-network references	13.2 %	1.8 %
YouTube, Facebook and Instagram references	11.6 %	1.8 %

The most cited AIO domains were youtube.com (70), seg-social.es (44), facebook.com (39) and grupo2000.es (38). For AI Mode they were seg-social.es (63), lamoncloa.gob.es (49) and boe.es (47). Types are in `data/citation_types.csv`. Version 1 matched domain types by substring (taxfix.com counted as `x.com`, cuatrecasas.com as `as.com`); domains are now matched by label.

Citing a public body did not protect against error. 11 of the 13 AI Mode errors cited one, against 3 of the 12 AIO errors.

5.5 Source or synthesis?

For the 25 wrong answers, the script fetched the cited pages and applied the same regular expressions:

- in 14 answers, at least one cited page states only the superseded rule;
- in 8, a page states both rules (often history tables);
- in 3, nothing readable addresses the rule;
- in none did the pages state only the current rule.

In 40 correct control answers, the cited pages stated only the superseded rule in 2 cases. The errors look inherited from the pages, not introduced by the synthesis, but pages were fetched on 3 October, not when Google read them.

5.6 Stability

The same query–surface pair got the same final verdict in all three readings in 66.0 % of cases for AIO and 89.7 % for AI Mode (`data/stability.csv`). Most AIO changes are an AIO appearing or disappearing.

5.7 Consistency checks

Second reader. A separate AI agent (a Claude Sonnet model; prompt in `data/second_reader_prompt.md`) classified 60 answers blind. The sample had three strata (seeds 31–34):

- A: 20 of the 30 confirmed errors;
- B: all 17 answers the regex script flagged but the re-reading rejected;
- C: 23 of 502 answers flagged neither way.

	OVERALL	A ERRORS	B FLAGGED, REJECTED	C UNFLAGGED
exact agreement	83.3 %	75.0 %	88.2 %	87.0 %
agreement on "wrong or not"	93.3 %	100 %	88.2 %	91.3 %

False negatives. In 40 random unflagged answers from the facts that were not fully re-read, the agent found 0 missed errors (0 of 40; 95 % CI 0–8.8 %).

6. Discussion

On these 101 agent-written queries, one evening, about one answer in twenty was wrong on both surfaces.

The errors follow the law's own changes. Of 25 errors, 20 concern three facts. A Supreme Court annulment, a decree-law amendment and a deadline tied to an unpublished order each leave a large body of pages describing the earlier state. The page checks are consistent with answers repeating those pages.

The two surfaces failed differently. AIO sometimes did not answer at all (14–23 % of queries per reading). AI Mode always answered, cited the BOE three times as often, and still repeated a repealed decree-law hours after its repeal. Official sources were no protection: most AI Mode errors cited a public body.

The study's own first draft failed in the same way, and was caught only by an adversarial review. Checking that a quotation exists in the BOE is not checking that it is in force. Annulments, amendments by later decree-laws and transitional periods sit in the "later references" section of each norm's BOE page.

For anyone auditing current law. Re-check the status of every rule on the day of each reading, from the consolidated text and its later references, not from the original text.

7. Limitations

- **Population.** 101 queries written by an AI agent for 28 facts it chose for having a crisp answer, issued from one country-level location, signed out, on desktop, on one evening. The numbers describe this set, not Spanish legal search.
 - **Variability.** Answers change between requests, and identical answers recur across readings. The Wilson intervals treat 531 answers as independent, but only 17 distinct query–surface pairs carry errors.
 - **Ground truth.** The rule in force was read from the BOE by AI agents, and needs a lawyer's confirmation. This applies especially to F14 (the self-employed brackets are unchanged, but fees rise slightly with the MEI), F21 (general rule vs sectoral rules), F28 (no order published, checked in the BOE but not elsewhere) and F15 (regional registers remain).
 - **Judgment-heavy verdicts.** `incorrect` (fixed dates) and the line between `mixed` and `current` are judgment calls. The second reader agreed that these answers were wrong, but not always on the category.
 - **Rule freeze.** Not verifiable (§4). Final verdicts rest on the agent's re-reading, which happened after all readings.
 - **Page checks.** Pages were fetched by script on 3 October and judged by regular expressions; some were unreadable. Six pages whose `robots.txt` reserves them (five on facebook.com, one on legatiqu.ai) were removed from the analysis afterwards; no figure changed (`METHOD.md` §8).
 - **Prior-work search.** Thin: three APIs, eight queries, no grey literature.
-

8. Data, code and licences

All published figures are regenerated by `scripts/aggregate.py` from the published files.

`scripts/check_headline.py` asserts each headline number against `data/` and against this text.

`scripts/classify.py` needs the full answers in `data/raw/responses/`, which are not redistributed: they are Google-

generated text obtained through DataForSEO. `data/answers.csv` carries extracts of at most 300 characters for flagged or wrong answers, as evidence.

Licences:

- code: Apache-2.0;
- our data and text: CC BY 4.0;
- third-party content (Google answers, DataForSEO responses): not redistributed.

9. Automation and review

- **Done by AI agents:**
 - choosing the facts and checking their status in the BOE;
 - the queries;
 - collection;
 - the regex rules, which a script applies;
 - re-reading 120 answers and the 40-answer false-negative sample;
 - typing the cited domains from the domain name;
 - the literature search;
 - this text.
- **Second reader:** a separate AI agent, given only the corrected fact sheet, the query and the answer.
- **Review:** an independent AI agent reviewed version 1 adversarially and found the errors corrected here (`private/REVIEW.md`, "Fixes applied").
- **Checks:** the facts and their legal reading, the verdicts, the domain types and the figures were checked by AI agents. Where this text says something was checked, an AI agent or a script checked it.

10. Competing interests

None related to Google or DataForSEO. EasyByte Hub S. Coop. Mad. sells Verifactu-related software and services, and publishes Spanish legal and tax content on its own website. That content competes for some of the same searches as the advisory and software sites cited in the answers. EasyByte's domains were never cited. The study received no funding.

11. Ethics

Readings were obtained through DataForSEO's commercial SERP API, which queries Google; we did not query Google directly. DataForSEO's terms make the customer responsible for the use of SERP data. We publish only aggregates and extracts of at most 282 characters. The queries contain no personal data. Cited pages were fetched once per check, with an identifying User-Agent; six pages whose `robots.txt` reserves them were removed from the analysis (§7). Their text was not stored, and social-network URLs are published as domains only.

How to cite

EasyxLab (2026). Do Google's AI Overviews and AI Mode keep up with Spanish rule changes? Study S7. EasyByte Hub S. Coop. Mad. <https://github.com/easybytehub/easyxlab>

References

- H. Xu, U. Iqbal, J. M. Montgomery, "Measuring Google AI Overviews: Activation, Source Quality, Claim Fidelity, and Publisher Impact", arXiv:2605.14021v1, 2026 (abstract via the arXiv API; Table 3 via arxiv.org/html, accessed 2026-10-03).
- V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools", arXiv:2405.20362, 2024.
- D. Hu et al., "Auditing Google's AI Overviews and Featured Snippets: A Case Study on Baby Care and Pregnancy", arXiv:2511.12920, 2025.
- S. Sünkler, D. Lewandowski, S. Schultheiß, O. Koop, "Information Diversity and Authority Shifts in Google's AI Overviews: An Audit of Health Queries", WebSci 2026, doi:10.1145/3795513.3807429.
- M. Zha, H.-C. H. Chang, "Who Anchors AI Overviews in Health? Baidu, Google, and the Geography of Authority", arXiv:2609.06798, 2026.
- E. Friesen, A. Roy, "Better than a Google Search? ...", SSRN 5402185, 2025, doi:10.2139/ssrn.5402185.
- Li, Han, Guo, "When Legal AI Speaks with Authority ...", SSRN 6866876, 2026, doi:10.2139/ssrn.6866876.
- M. S. Ahmed, L. Galke Poech, R. Röttger, "Stale-Document Poisoning: When Outdated Retrieval Overrides Correct Model Answers", arXiv:2609.31342, 2026.
- R. Cymbler, D. Guez, L. Fabre, "Temporal Misgrounding in Legal RAG: A Versioned-Corpus Benchmark for French Tax Law", arXiv:2608.09393, 2026.
- Agencia Estatal Boletín Oficial del Estado: documents, consolidated texts and open-data API, accessed 2026-10-02/03; each norm with its identifier and quotation in `data/facts.csv` and `data/facts_in_force.csv`.
- DataForSEO, SERP API (Google Organic and Google AI Mode, live advanced), <https://docs.dataforseo.com/v3/>.

Cite this study

CITATION

EasyxLab (2026). Do Google's AI Overviews and AI Mode keep up with Spanish rule changes? About one answer in twenty did not. Study S7. EasyByte Hub S. Coop. Mad.
<https://github.com/easybytehub/easyxlab/tree/main/studies/s7-aio-spanish-regulation>

BIBTEX

```
@techreport{easyxlab_s7,  
  title      = {Do Google's AI Overviews and AI Mode keep up with Spanish rule changes?  
About one answer in twenty did not},  
  author     = {{EasyxLab}},  
  institution = {EasyByte Hub S. Coop. Mad.},  
  number     = {S7},  
  year       = {2026},  
  url        = {https://github.com/easybytehub/easyxlab/tree/main/studies/s7-aio-spanish-regulation}  
}
```