FAST GROWTH Advisors

WHITE PAPER

AUGUST 2026

# THE RECOVERABLE MESSAGE

## What answer engines really read of your site

> *Twenty edge cases, three extractors, 94 audits. A measurement, not an opinion.*

### Measured

Observed in a published large-scale study, or tested by us and reproducible.

### Deduced

A necessary logical consequence of a measured fact, flagged as such.

### Not verified

Circulates without a primary source. We say so rather than repeat it.

### What is not here

No tactics, no visibility promises. The limits are written in the body text.

*This document sells nothing. It measures what an engine keeps from a site, and shows how to check it yourself.*

SECTION 01

# IN BRIEF

A B2B company takes care of two readings of its website: the buyer's and Google's. A third has settled in unannounced, the reading done by answer engines, and it follows the rules of neither.

This document establishes three things.

First, that this reading can be measured. We submitted a test page of twenty edge cases to the three extractors used to prepare the corpora of large models, and several results contradict reflexes inherited from classic search optimisation.

Second, that French tech fails at it massively. Across 94 full audits of post-funding startups, readability by AI engines is the weakest of the fifteen sub-criteria measured, at 0.91 out of 5.

Third, that the loss is not uniform, and this is the result that matters. The narrative survives. The proof dies. An engine can then say what the company does, not what sets it apart, and it cites as the category reference the competitor who put its proof in HTML.

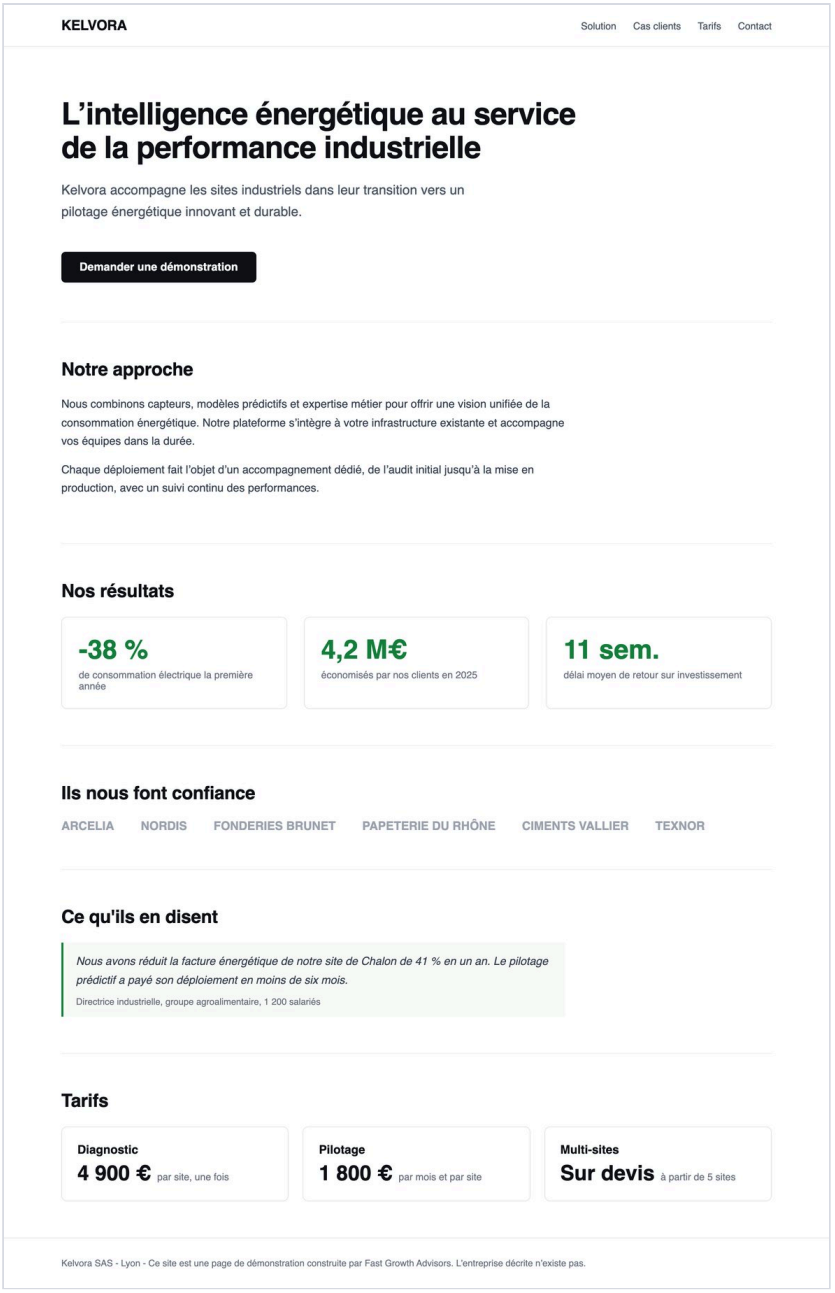
From this we derive a new audit measure, the recoverable message: the share of positioning that survives a read without JavaScript, extraction included. And an action plan that calls for a targeted intervention, not a rebuild.

SECTION 02

# WHAT DOES THE LOSS LOOK LIKE, ON A CONCRETE CASE?

Before explaining the mechanism, let us show it.

We built a fictional B2B startup homepage carrying the defects we meet most often in audits: the result figures, the client logos, the testimonial and the pricing grid are injected by JavaScript, and an older version of the positioning was left in the code, hidden by CSS. The company described does not exist. The page is in French, and both page and protocol are reproducible.




What a visitor sees. The page displays a promise, three result figures, six client logos, a quantified testimonial and three prices.

Here is what three extractors keep from that same page, without executing JavaScript, exactly as the bots that build corpora do. These outputs are not reconstructed: they come from actually running trafilatura, jusText and Resiliparse on the served bytes.

### What trafilatura keeps

```
Kelvora accompagne les sites industriels dans leur transition vers un pilotage énergétique innovant et durable.
Demander une démonstrationNous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s'intègre à votre infrastructure existante et accompagne vos équipes dans la durée.
Chaque déploiement fait l'objet d'un accompagnement dédié, de l'audit initial jusqu'à la mise en production, avec un suivi continu des performances.
Kelvora divise par deux le coût énergétique d'une ligne de production en quatre-vingt-dix jours, sans arrêt d'exploitation et sans investissement matériel.
```

### What jusText keeps

```
L'intelligence énergétique au service de la performance industrielle
Notre approche
Nous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s'intègre à votre infrastructure existante et accompagne vos équipes dans la durée.
Chaque déploiement fait l'objet d'un accompagnement dédié, de l'audit initial jusqu'à la mise en production, avec un suivi continu des performances.
Nos résultats
Ils nous font confiance
Ce qu'ils en disent
Tarifs
Ancienne version du positionnement, laissée dans le code lors de la refonte
Kelvora divise par deux le coût énergétique d'une ligne de production en quatre-vingt-dix jours, sans arrêt d'exploitation et sans investissement matériel.
```

### What Resiliparse keeps

```
L'intelligence énergétique au service de la performance industrielle

Kelvora accompagne les sites industriels dans leur transition vers un pilotage énergétique innovant et durable.

Demander une démonstration

Notre approche

Nous combinons capteurs, modèles prédictifs et expertise métier pour offrir une vision unifiée de la consommation énergétique. Notre plateforme s'intègre à votre infrastructure existante et accompagne vos équipes dans la durée.

Chaque déploiement fait l'objet d'un accompagnement dédié, de l'audit initial jusqu'à la mise en production, avec un suivi continu des performances.

Nos résultats

Ils nous font confiance

Ce qu'ils en disent

Tarifs

Ancienne version du positionnement, laissée dans le code lors de la refonte

Kelvora divise par deux le coût énergétique d'une ligne de production en quatre-vingt-dix jours, sans arrêt d'exploitation et sans investissement matériel.
```

| READING | BYTES OF TEXT | SHARE | TRACKED ELEMENTS KEPT |
|---|---|---|---|
| Seen by a human, JavaScript executed | 1450 | 100 % | 8 / 8 |
| trafilatura | 706 | 49 % | 1 / 8 |
| jusText | 796 | 55 % | 2 / 8 |
| Resiliparse | 949 | 65 % | 2 / 8 |

Volume of text kept, and survival of the proof elements, on the demonstration page.

Volume alone would be reassuring: between 49% and 65% of the text survives. It hides the point. What disappears is not spread at random, it is the part that proves. Of the eight elements tracked, the extractors keep one or two, and never the ones that matter: the promise gets through in two of them, the hidden old positioning gets through in all three.

The six proof elements proper, the three result figures, the client logos, the testimonial and the prices, survive nowhere.

Two details deserve a pause.

The first: jusText and Resiliparse keep the section *headings*, "Nos résultats", "Ils nous font confiance", "Tarifs", but empty. So the engine knows results, clients and prices exist. It knows none of them. This is the situation described later in this document: it can say what the company does, not what sets it apart.

The second is more awkward. The old promise left in `display:none`, the one the team believed it had removed, comes back in all three extractors. It is the only commercial figure this page actually transmits, and it is the one it no longer wanted to say.

That leaves the metadata channel, which follows other rules. Trafilatura does not keep the displayed heading but the `og:title`, here "Kelvora, pilotage énergétique industriel". A tag many treat as a social detail therefore decides the title under which the page enters the corpus.

SECTION 03

# WHY DOES THIS MATTER NOW?

Three dated facts, none of them ours.

Visits referred by ChatGPT to B2B websites multiplied by four in a year: roughly 645,000 monthly visits in June 2025, 2.6 million in June 2026, a 303% rise, across more than 11 billion visits analysed by Demandbase and published on 12 August 2026. IDC (International Data Corporation) projects that by 2028, 70% of B2B buyers in the United States will rely on generative AI to discover, evaluate and select their vendors. And since 13 August 2026, Microsoft Clarity has displayed a ratio between pages scraped and visitors referred that puts the subject on every executive agenda: about 6,000 pages scraped per referred visitor, in Microsoft's published example.

That last figure feeds a debate about compensation which is not our subject. The operational question lies elsewhere. Since these engines read your pages in volume, what do they keep?

One convention, before we go further. Every claim in this document is labelled Measured (observed in a published large-scale study, or tested by us and reproducible), Deduced (a necessary logical consequence of a measured fact) or Not verified (circulating without a primary source). A document that skips this distinction ends up circulating inferences as facts. We have experienced that internally. We would rather avoid it in public.

SECTION 04

# WHY THINK IN THREE STAGES RATHER THAN ONE?

The most common mistake is to reason in a single step: "the bot sees the page."

There are three, and a piece of information can survive the first two then die at the third.

The first stage is the fetch. The agent requests the URL. It can fail on a firewall that blocks, a missing page, a rate limit, a consent banner that masks the body. The fix belongs to infrastructure.

The second is the response. The server returns bytes, and no JavaScript is executed at this point. If the text is not in those bytes, it does not exist for what follows. The fix belongs to rendering architecture.

That leaves the third, extraction. A tool turns the HTML into text, and discards part of the document along the way. It is the stage almost always forgotten, and the one where we measured the least intuitive results.

Three stages, three causes of invisibility, three fixes with nothing in common. An audit that conflates these three levels produces a false diagnosis and sends the budget to the wrong place, typically towards a technical rebuild when the defect sat in the markup, or the other way round.

SECTION 05

# WHICH BOTS RUN JAVASCRIPT, AND WHICH DO NOT?

The answer is documented, and it is clear-cut.

Measured, by Vercel and MERJ (a web analytics firm) in December 2024, across the Vercel network and two third-party stacks: the bots of OpenAI (GPTBot, OAI-SearchBot which is its search bot, ChatGPT-User), Anthropic (ClaudeBot), Perplexity, Meta, ByteDance and Common Crawl do not execute JavaScript. Googlebot and Applebot do render pages, in a headless browser.

The detail that settles it: these bots download JavaScript files without executing them. 11.5% of ChatGPT's requests and 23.8% of Claude's target scripts, treated as text, never as programs.

Two caveats about freshness come with that picture. The measurement dates from December 2024 and remains the last one published at scale. No model publisher documents its rendering capability, so everything rests on third-party observation, to be reconfirmed periodically.

And rendering now exists elsewhere. Agentic browsers do execute JavaScript, but on the user's own machine, one page at a time, without writing anything durable.

Hence the paradox that structures this document. Rendering exists, and it happens exactly where it feeds no index. The layer that decides future citations, the one that builds corpora and answer indexes, renders nothing.

SECTION 06

# WHAT ACTUALLY SURVIVES EXTRACTION?

On 29 July 2026 we submitted a test page to the three extractors documented in corpus preparation pipelines: jusText, Resiliparse and trafilatura. Twenty unique markers, one per edge case, inside a body of 5,389 characters. Length is not a detail: on a document that is too short, the jusText classifier files everything as boilerplate and the test becomes false. Exact versions are recorded in our test files, and the measurement is to be replayed at every version upgrade.

## The four findings

**Content hidden by CSS is read.** A block set to `display:none`, served by the server, comes back in the text extracted by jusText and Resiliparse. The reason is mechanical: not executing JavaScript means not applying CSS, so the extractor has no way of knowing a block is hidden. This is the complete inversion of classic search optimisation. The happy consequence is that an FAQ folded into a server-rendered `details` element is read in full. The awkward one is that rebuild leftovers, A/B test variants and old content left in `display:none` pollute the reading an engine makes of your positioning. Nobody audits what their own HTML hides.

**The level-one heading is not the title that wins.** Trafilatura does not include the `h1` in the extracted body. It takes its title from the metadata, and in our test the `og:title` prevails. That tag everyone treats as decorative, meant for social sharing, is in at least one common extraction chain the title that reaches the corpus. A neglected `og:title`, or one whose vocabulary disagrees with the `h1`, creates two competing positionings: one for the visitor, one for the corpus.

**The test recommended everywhere produces a false positive in exactly the case it should detect.** A `curl` followed by a `grep` on the raw response finds the phrase you are looking for even when it is injected by JavaScript: it sits as a literal string inside the `script` block, a portion every extractor discards and no engine reads as content. Measured here: naive search, found; after removing scripts, absent; in the extracted text, absent. This false positive is silent and systematic. Any serious measurement strips `script` and `style` blocks and comments before searching, or works on the output of a real extractor.

**The empty shell, quantified.** A pure client-rendered application serves 128 bytes of HTML. Useful text extracted: zero bytes. The case is binary, and it is becoming rarer as several platforms now pre-render their sites by default. It still exists, and it forgives nothing.

To which the silent losses must be added. `data-*` attributes, iframe titles and JSON embedded in the page are lost by all three extractors, even when present in the response. The `alt` attribute of images is unreliable. Any content inside an iframe, or carried by an image, should be treated as not acquired.

## What we do not know

By symmetry, here is what remains not verified and must not be sold as fact. The causal link between server rendering and citation rate: no controlled study exists, and the spectacular visibility gains in circulation come from vendor case studies. The share of the web in pure client rendering: no public source measures it. The extractor actually used by each model publisher: none publishes it, hence our rule of trusting only what passes all three.

SECTION 07

# WHY DOES THE PROOF GO FIRST?

Here is the result that connects extraction mechanics to revenue.

On a typical startup, text coverage is high. The page body gets through, the narrative gets through, the promise gets through. What disappears is concentrated on one precise family of elements: client logos in JavaScript carousels, reviews in third-party widgets, the monthly-annual pricing toggle, dynamic comparison tables, key figures animated on scroll, the integrations list loaded from an interface, case studies in modal windows.

Everything that proves.

What an answer engine then produces is predictable. It can say what the company does and describes the category correctly. It cannot say what sets the company apart, so it cites as the category reference the competitor who put its proof in HTML.

This mechanism combines with a pattern the Fast Growth Advisors Observatory finds in 61% of detailed audits: the figures that prove the value, gains, savings, impact metrics, do exist, but in the press and in funding announcements, not on the website. The proof is therefore doubly absent, editorially first, technically second. One Series B deeptech in our corpus, 30 million euros raised, a process that cuts production costs by 45%, offers on its homepage “innovative advanced recycling solutions.” Audit score: 24 out of 75.

SECTION 08

# WHERE DOES FRENCH TECH STAND, MEASURED?

The Fast Growth Advisors Observatory audits the messaging of French post-funding startups. Our measurements cover 375 simple audits, 94 full audits across fifteen sub-criteria, and 346 startups cross-referenced with the amount they raised.

The weakest of the fifteen sub-criteria is readability by AI engines: 0.91 out of 5, less than 20% of the potential. None of the fifteen exceeds 60% of the potential, and the best rated, differentiation and clarity, peak at 52%. It is the weakest point of an already weak set.

Two further figures set the context. 75.9% of the 369 French post-funding startups audited by the Observatory sit below the critical clarity threshold, set at 37.5 out of 75 (Q2 2026 edition). And the correlation between the amount a startup raised and its messaging score is close to nil, $R^2 = 0.036$, which means the two quantities are effectively independent of one another.

In other words: raising 50 million rather than 5 makes neither the message clearer nor the site more readable. Funding does not solve this problem.

SECTION 09

# HOW DO YOU GET FOUND, AND IN WHICH LANGUAGE?

Extraction decides what is kept from a page already found. You still have to be found. Our measurements on the engines themselves add two counter-intuitive results.

The first concerns language.

When a user asks a question, the engine does not search for the question. It breaks it down into queries it issues itself, and those queries can be read in the official interfaces. Measured across two independent corpora, 20 B2B technical questions then 30 questions spread across 15 sectors: on B2B technical subjects, ChatGPT issues 46% of its queries in English for questions asked in French. The reference literature of software is English, and the engine searches in the language where the authority on the subject lives.

A consequence observed on our own site, with three converging signals but modest volumes, therefore to be confirmed: the page most fetched by ChatGPT, apart from the homepage, is the English version of our AI visibility page, ahead of its French twin. The English version of a French B2B site does not only serve international clients. It serves French clients, through engines that search in English and answer in French.

A final result: not every subject triggers a search. On our panel, advice questions (“which one should I choose”, “how do I avoid”) are answered from memory, with no page able to weigh in, whereas questions naming entities do trigger queries. Before investing in a subject, the first measurement is therefore: will this engine even search?

SECTION 10

# WHAT IS THE RECOVERABLE MESSAGE?

A messaging audit classically measures two things: the claimed message, what the company says, and the understood message, what the buyer retains. This document justifies a third, independent of the first two.

The recoverable message is the share of positioning that survives a read without JavaScript, extraction included.

The three cannot be deduced from one another.

A message can be clear, differentiating, and still not recoverable. This is in fact the most common case among companies that invested at the same time in their narrative and in a modern site, since it is precisely the most carefully built components, carousels and animated counters, that disappear at extraction.

The useful output is not a score.

It is a location: which precise blocks of the page do not survive, and among those, which carry proof. On the sites we audit, the answer almost always fits in a short list, fixable in one intervention. That is what makes the subject tractable: the diagnosis is surgical, not architectural.

SECTION 11

# WHAT SHOULD YOU DO ON MONDAY MORNING?

The thirty-second test first. Disable JavaScript, reload the homepage, look at what remains. It is the rough approximation of the recoverable message, and it is enough to detect the serious cases. It says nothing, however, about what an extractor discards once the page has been fetched, which is the second half of the diagnosis.

Then, in order of return: take the proof out of JavaScript, with logos, figures, testimonials and prices in server-rendered HTML, keeping the animation on top if you wish; replace hydrated accordions with server-rendered `details` elements; align the vocabulary of the `og:title` with that of the `h1`; audit what your HTML hides, since rebuild leftovers in `display:none` are read; and check your redirect plan, since a substantial share of generative engines' fetch budget is lost on dead pages, 34.8% of ChatGPT's fetches according to the same Vercel study.

On the `llms.txt` file, which you will hear about: no major engine has confirmed reading it. Putting one in place costs little. Presenting it as a lever is an unmeasured promise, and this document makes none.

SECTION 12

# WHAT ARE OUR LIMITS?

The extraction measurements date from 29 July 2026, on trafilatura, jusText and Resiliparse, with versions recorded in our test files. The query decomposition and language measurements date from 5 August 2026, through the engines' official interfaces with the search tool enabled, across two independent corpora, for a total cost under 20 dollars. The audit figures come from the Fast Growth Advisors Observatory 2026, whose corpus and thresholds are documented in the report.

Our limits, written down before anyone finds them for us. The Observatory sample is French and post-funding, it does not describe the web. The measurements on our own site rest on modest volumes. And the reference measurement on JavaScript execution dates from December 2024, which is old for this subject.

Every figure in this document carries its base and its date. If one of them does not survive your own test, write to us. That is the whole point.

SECTION 13

# FREQUENTLY ASKED QUESTIONS

### How does the recoverable message differ from the understood message?

It is the share of a company's positioning that survives a read without JavaScript, extraction included. It differs from the claimed message, what the company says, and from the understood message, what the buyer retains. A message can be clear, differentiating, and still not recoverable by an answer engine.

### Do answer engines execute the JavaScript on my site?

No, not the bots that build corpora and indexes. GPTBot, ClaudeBot, PerplexityBot and Common Crawl download JavaScript files without executing them, according to the Vercel and MERJ measurement of December 2024. Googlebot and Applebot do render pages. Agentic browsers do too, but on the user's machine, without feeding any index.

### Is the "disable JavaScript" test enough to know what an AI reads of my site?

It gives a faithful, free first approximation, enough to detect the serious cases, which is a lot for a test that takes thirty seconds. It says nothing, however, about the next stage, extraction, where a tool turns HTML into text and discards part of the document. A complete measurement therefore goes through a real extractor, not through a reading of the raw HTML.

### Why do my client logos and figures disappear when my text gets through?

Because those elements are almost always injected by JavaScript: carousels, review widgets, pricing toggles, animated counters. The body text, by contrast, is server-rendered. The loss is therefore concentrated on what proves, not on what describes, which explains why an engine can say what a company does without being able to say what sets it apart.

### Do I need to rebuild my site to be readable by AI engines?

Rarely. In most of the cases we audit, the narrative already gets through and it is the proof that is missing, which calls for a targeted intervention on a few blocks. A rebuild is only justified in the binary case of pure client rendering, where the server returns no usable text at all.

SECTION 14

# GLOSSARY

### Recoverable message

Share of positioning that survives a read without JavaScript, extraction included. An audit measure distinct from the claimed message and the understood message.

### Extraction

Third stage of machine reading, where a tool turns HTML into text and sets aside what it considers boilerplate. It is the stage most often forgotten in diagnoses.

### Server rendering

Mode of building pages where the HTML is assembled by the server before being sent. It contrasts with client rendering, where content is built by the browser executing JavaScript.

### jusText

An extractor that separates content from boilerplate by measuring function word density. It files lists of proper nouns without sentences as boilerplate, which makes client reference lists disappear.

### Trafilatura

A main content extractor. It does not include the level-one heading in the body and takes its title from page metadata, where the og:title prevails.

### Resiliparse

A fast extractor from the ChatNoir pipeline, used on very large corpora. It keeps text hidden by CSS, since CSS is not applied.

### og:title

A metadata tag originally intended for social sharing. It reaches the corpus in at least one common extraction chain, which makes it structural rather than decorative.

SECTION 15

# SOURCES


1. Vercel and MERJ. *The rise of the AI crawler*, 17 December 2024. Server log study: none of the major generative bots executes JavaScript, and 34.8% of ChatGPT's fetches target non-existent pages. vercel.com
2. Microsoft Clarity. *Scrape-to-Referral insights*, 13 August 2026. Introduction of the ratio between pages scraped and visitors referred, with a public example of about 6,000 to 1. clarity.microsoft.com
3. Labs by Demandbase. *ChatGPT referrals to B2B websites*, 12 August 2026. Visits referred by ChatGPT to B2B sites: about 645,000 per month in June 2025, 2.6 million in June 2026, a 303% rise, across more than 11 billion measured visits.
4. IDC (International Data Corporation), 2024. Forecast: by 2028, 70% of B2B buyers in the United States will rely on generative AI to discover, evaluate and select vendors. idc.com
5. trafilatura, official documentation. trafilatura.readthedocs.io
6. jusText, project repository. github.com/miso-belica/jusText
7. Resiliparse, official documentation. resiliparse.chatnoir.eu
8. Nvidia NeMo Curator, a corpus preparation pipeline documenting the use of these extractors. github.com
9. Message-Market Fit Observatory, Fast Growth Advisors, 2026. 375 simple audits, 94 full audits across fifteen sub-criteria, 346 startups cross-referenced with the amount raised.