

Unseen AI: what Europe's public sector buys, and what its procurement registers cannot show

Pablo Ramos-Sánchez, independent researcher · arquestra.dev · White paper ·

Oct 25, 2026 · v1.2 · <https://doi.org/10.5281/zenodo.23561547>

Executive summary

Public bodies across Europe now buy AI: chatbots that answer citizens, software that reads scans, systems that sort job applications. The EU AI Act places duties on the public bodies that use high-risk AI, which the Act calls *deployers*. These duties apply from 2 December 2027 for the uses listed in its Annex III, such as access to public services, employment or emergency triage. They apply from 2 August 2028 for AI inside regulated products, such as medical devices. Systems already in use by public authorities must comply by 2 August 2030.

Those who supervise these duties will need to know where the AI is, but the Act does not make a purchase say whether it contains AI, and the proposed EU Public Procurement Act leaves what notices must record to later delegated and implementing acts (European Commission, 2026, recitals 60-61). Public procurement is the most complete public record of what the state buys: every contract above a small amount is published in a procurement register, such as the EU's Tenders Electronic Daily (TED) or Spain's public procurement platform. This report asks what those registers can show about AI, and what they cannot.

The report examines TED and the national registers of Spain, Italy, France, Germany and the Netherlands in four steps. It **finds** the candidates: it scans the records for AI keywords and searches them by meaning, with an open model that ranks notices by what they say, in any language. It **classifies** them under a codebook: written rules, anchored in the AI Act, that decide what counts as AI. It **counts** them, correcting the keyword count with the error measured on a random sample of 2,317 purchases, each one reviewed (the gold standard, table 4.2). And it **compares** sources only at a common cut and in a stated unit. Chapter 3 says which step each figure comes from. The published counts of public AI reviewed for this report state no such error (chapter 2, table 2.1).

I. Much of the AI the public sector buys cannot be seen in the registers (chapter 4).

Registers record what is bought, not whether it contains AI. Some AI is *described in other words*: "software that reads mammograms", "speech recognition for the hospital", or only a brand name, such as licences of Claude, ChatGPT and Copilot bought by city councils, a

public broadcaster and a national museum. Depending on the source, keywords miss 4% (France) to 45% (TED) of the AI bought among IT purchases, with wide intervals (table 4.2). A search by meaning finds part of what they miss. Other AI is *not described at all*: a notice for a new car says nothing about its driver assistance, and one for a CT scanner says nothing about its learnt image reconstruction. No method can find what the text does not contain. Such AI can only be bounded, by sampling product categories (annex A). Without a label, no buyer, supervisor or researcher can find these systems.

II. Keyword counts give the wrong number (chapter 4). In this report, an AI-related purchase is one that buys an AI system, an AI course or AI governance. Of the records that name AI, only 35% (Netherlands) to 65% (TED) are AI-related purchases, and only 18% to 59% buy an AI system itself. The rest are records such as those of a buyer called "Directorate for Digitalisation and Artificial Intelligence", file numbers that start with "AI", or catering for AI conferences. These rates count records, as keyword counts of notices do. On the French records reviewed, the rate is 48% counted by record but 81% counted once per procedure, because one buyer's purchasing system, which publishes a record for each supplier it admits, fills half of France's records that name AI. In TED the error may run the other way: the estimate is above the keyword count, though its interval still includes it.

III. The money spent on AI cannot be read from the registers (chapter 4, annex A). Most TED and German notices give no amount. Some records hold the ceiling of a framework agreement (the most that may be spent over several years) instead of what an order costs. In Italy, three records hold EUR 2.3 billion of the EUR 2.6 billion recorded for contracts that name AI. This report therefore counts AI by number of purchases, where its error is measured. It gives money only as recorded, with and without the largest records.

IV. Most AI is bought in small contracts that TED does not see (chapter 4). In Spain, minor contracts (small contracts awarded directly) are 77% of AI purchases but about 4% of their value. In Italy, 93% of the contracts that name AI are direct awards. Studies built on TED alone see most of the money but miss most of the purchases. They also miss most of the small governance purchases: 22 of the 25 Spanish governance purchases are minor contracts. Comparisons between countries are therefore made at a common cut of EUR 140,000. Above it, the share of IT purchases that name AI is 1.1% to 2.4% in every source but Germany, where few IT records carry a usable amount, and with these samples the estimates of most sources cannot be told apart. The figures do not rank countries.

V. Whether that AI is governed cannot be seen either (chapter 5). A register shows governance only when it is bought as a contract of its own: an impact assessment, an audit, compliance support, an adoption strategy or an AI Act course. A search of TED and Spain finds 40 such purchases among 3,836 purchases that a classifier marks as AI (1%). That 1% is a floor. In the random sample, weighted by how it was drawn, governance is

about 4% of AI purchases (95% interval 1.4% to 7.8%), about 1 in 25, though on only seven purchases. A governance purchase costs about as much as a typical AI purchase, EUR 13,000 at the median. Governance done in house, or written into the contract for the system, leaves no trace.

VI. AI can help watch procurement, if a person oversees it the right way (chapter 3, chapter 7, annex B). This report is itself a use of AI to read procurement registers, a kind of monitoring that, in Sanchez-Graells's reading, the proposed EU Public Procurement Act does not address ([Sanchez-Graells, 2026](#)). Its two checks show what overseeing such a tool requires. Read blind, and weighted to the same design, 22% of purchases ended in doubt; reviewed with the agent's proposal in view, about 1% did: shown a firm answer, a reviewer settles what the text cannot. That is the automation bias that the AI Act asks oversight of high-risk AI to guard against (Articles 14 and 26). Neither side of such a review is free of bias: read blind, the person was pulled by words; reviewing the agent, the person deferred to it. The report offers no cure, but practices that expose the bias: an agent proposes and a person confirms, a sample is read blind, and the two are crossed and the cases of conflict studied. Ranking by meaning first, an agent reading next and a person auditing a blind random sample last reduces both the errors of keywords and the cost of reading everything.

VII. In health, where much of the high-risk AI is bought, the assurance visible is mostly certification written into the product, rarely the hospital's own governance (chapter 6). Health is 8.4% of the purchases marked as AI in TED and Spain: clinical assistants, AI contouring for radiotherapy (software that outlines organs on scans), AI mammography. Of 253 health candidates found by a search by meaning and then reviewed, four require the AI to meet the medical-device rules. Several AI imaging purchases carry no such condition; among them is AI mammography tendered for EUR 1.32 million. Certification governs the product through its manufacturer. It does not show that the hospital oversees how the product is used, which the AI Act also requires. Whether the December 2027 date is realistic, procurement data cannot tell: the registers cannot say which public bodies use such systems.

Recommendations (chapter 7). R1: a mandatory AI label on procurement notices, set by the buyer and declared by the supplier, so that a missing label can be checked, written into the EU Public Procurement Act now under negotiation and in force before 2 December 2027. R2: registers that record what was bought, for how much, and where it is registered: a description in enough words, the real cost of each order rather than the ceiling it draws on, and, for high-risk systems, a link to their registration in the EU database. R3: governance that is easy to buy. It should be budgeted in every contract for high-risk AI, offered as a lot in AI framework agreements, and available to small buyers

from central purchasing bodies as call-off orders (orders placed under an existing framework, without a tender of their own). R4: every published count of public AI states how it was validated: its definition, a sample checked by hand, and its precision and recall.

What this report does and does not show. It measures, on a random sample and with its sampling error, how wrong keyword counts are and how much AI described in other words they miss. It bounds, without counting, the AI built into products whose notices do not describe it. It describes governance bought as a contract of its own, as a floor. It measures what public bodies buy and how registers record it, not what they use. Whether a product contains AI, only its supplier can tell; that is why the report asks for a label. And it shows that the rules meant to govern public AI, the AI Act among them, assume a visibility the registers do not give. They can still be made to work, if the label and the oversight of the tools that read the registers are written into law while the EU Public Procurement Act is open.

1. Introduction: procurement is where much of the state's AI is decided

Much of the AI a public administration uses is bought, not built in house. A hospital buys a system that reads mammograms; a city, an assistant that answers residents; a ministry, a licence for a language model. Each purchase decides what is automated, with which data and which safeguards, and who answers when it fails. That is why procurement is where AI governance stops being a principle and becomes a clause, a budget line or a commissioned audit: it is the gatekeeper of the public sector's digital transformation ([Sanchez-Graells, 2024a](#)). Europe's rules assume that this gate is watched. The AI Act places duties on the public bodies that use high-risk AI, and the proposed EU Public Procurement Act will set how they buy it (European Commission, 2026). Both take for granted that someone can tell which AI the state buys. This report shows, with data, that no one can: registers do not say when a purchase contains AI, keyword counts err in both directions, and governance is visible only when it is bought as a contract of its own. It also shows that the remedies on paper, human oversight of AI and model clauses for buying it, stay what Sanchez-Graells (2026) calls an *empty shell* until they are made operational, and it tests one way to make them so.

The timing matters. The EU AI Act ([Regulation \(EU\) 2024/1689](#)) places obligations on the deployers of high-risk systems, among them the fundamental rights impact assessment of Article 27. After the Digital Omnibus on AI ([Regulation \(EU\) 2026/1744](#)), which amends Article 113 of the Act, these obligations apply from 2 December 2027 for the uses listed in Annex III and from 2 August 2028 for AI in regulated products. Systems already on the

market are caught after a significant change in their design, and those used by public authorities must comply by 2 August 2030 in any case (Article 111(2)). A high-risk system bought today will still be in use when the rules bite.

The question, and what purchase data can answer

This report asks what the procurement registers can show about the AI the public sector buys, and what they cannot. The question echoes two reports: the OECD's *Governing with Artificial Intelligence: Are governments ready?* ([OECD, 2024](#)) and the Open Contracting Partnership's *Is your government ready to buy AI?* ([Sklar, 2026](#)). Both answer from policies and practice. This report answers from purchases, and its answer concerns a step before readiness: the registers do not show much of the AI the public sector buys, so they cannot show what it has to govern, or whether it does. The problem is not that the public sector buys high-risk AI, which the Act allows under conditions. The problem is that no one can tell when it does.

What is new is the measurement. Published counts of public AI rely on keywords, language models or both, where they describe their method at all, and none reports its error (table 2.1). This report scans the registers for keywords and also searches them by meaning. The search by meaning uses an open embedding model: a model that places notices that mean the same thing close together, whatever words and language they use. The report then classifies what it finds under a codebook: written rules, anchored in the AI Act, that decide what counts as AI. Finally, it measures the error of every step on a gold standard, a random sample of 2,317 purchases, each one reviewed, so that each count comes with its sampling margin.

Chapter 2 sets out what is known and why the published figures disagree. Chapter 3 explains the four steps of the method and the sources. Chapter 4 shows what the registers do not show: AI that is named but miscounted, AI described in other words, AI built into products, and the small contracts TED does not reach. Chapter 5 turns to governance, chapter 6 to health, chapter 7 to recommendations and chapter 8 to limits. Annex A gives the method in detail, and annex B the checks on the classification.

What a purchase can and cannot tell

The report measures the public demand side: what is bought, not what is used or what works. A purchase does not prove deployment, and an awarded amount is not money spent.

2. Context: what is known, and why the published figures disagree

This report sits where two fields meet, and they barely overlap. The first, on AI in the public sector, has many principles and little measurement. Qualitative work asks whether public buyers are equipped to procure AI in the public interest ([Ada Lovelace Institute, 2024](#)) and how procurement can serve AI governance ([Sanchez-Graells, 2024a](#); [Shamir, 2026](#)); it does not measure what they buy. The second, public procurement analytics, is large and established. The Government Transparency Institute publishes contract-level data for 42 countries ([Fazekas et al., 2024](#)), on which machine-learning screens for cartels are now tested ([Fazekas et al., 2026](#)). The World Bank's [ProACT](#) covers 110 countries; [DIGIWHIST and Opentender](#) gathered 17.5 million tenders; the Open Contracting Partnership keeps a [data registry](#); and the European Commission runs the [Single Market Scoreboard](#) and the [Public Procurement Data Space](#). This field measures corruption risk and competition with red flags, such as single bidding, where a tender receives only one bid ([Fazekas and Kocsis, 2020](#)). It sorts purchases almost entirely by their CPV code, from the EU's Common Procurement Vocabulary, which classifies purchases by product. Work that uses procurement data to measure how AI is bought and governed sits in the very small intersection of the two fields.

Almost no government keeps a register of its AI

Only 6 of 36 countries (17%) keep an open register of public algorithms or AI: Canada, Estonia, France, Korea, the Netherlands and the United Kingdom. Only 3 make registration mandatory ([OECD, 2026](#), 4.5.3 and 4.5.4). Civil society asks for such a register: AlgorithmWatch told the German Bundestag in March 2026 that public-sector AI should be registered ([Sombetzki, 2026](#)).

Without such a register, procurement is the most complete public trace. Every purchase above a small amount must be published, in every member state, and in a common format above the EU thresholds. Procurement misses the AI a body builds in house, but no other source covers so many buyers on the same terms. Other routes answer other questions. A World Bank survey of 60 economies relies on what governments report, and finds that "governance frameworks are lagging" ([Lundy et al., 2026](#)). A pilot that detects AI-written text in public documents set procurement data aside as "noisy and difficult to compare across countries" ([Atkinson and O'Bryan, 2026](#)). This report takes that difficulty as its subject.

The market, from what is published

Eight measurements published in the last year cover the United Kingdom, the United States, Spain, Sweden and a worldwide pool (table 2.1).

Measurement	Scope	AI purchases found (each study's definition and unit)	How AI is found	Error published
World Bank, WDR 2026	6.13 million contracts (World Bank, OCP, TED)	17,865 (0.29%)	A language model reads every record; model versions not given, reproduction "not guaranteed"	No
Tussell, market analysis (Feb 2026)	UK, 2025	435 contracts, £699m	Not described	No
Tussell, tracker (Aug 2026)	UK, 2025	521 contracts, £1.18bn	Keywords, then ChatGPT for themes	No
BidSkim (Sep 2026)	UK, 2025	555 procurements, £505m	Five routes, one of them semantic search, then a language model reads each candidate	No
Brookings (May 2026)	US federal, Sep 2023 to Mar 2026	1,743 contracts, \$7.2bn obligated	"Artificial intelligence" or "AI" in the description	No
CSIS (Sep 2026)	US federal, FY2019 to FY2025	2,255 awards, \$4.1bn obligated	AI terms, plus review of named programmes	No
licitaciones.io (Aug 2026)	Spain, Jan 2025 to Jul 2026	541 tenders, over €208m	Not described	No
Alice Labs (2026)	Sweden, 2016 to 2025	250+ contracts, about SEK 1.3bn	Keywords, then a manual check against tender documents	No

Table 2.1. Published measurements of public AI purchases, with their method as each describes it. "Error published": precision or recall of the classification.

The measurements agree on two things. AI buying grows fast: BidSkim counts 271 UK AI procurements in 2022 and 555 in 2025. And it is a sliver of public buying: the World Bank screen flags 0.29% of the contracts it reads.

On size they disagree, even within one country and one year. For UK purchases in 2025, Tussell's February analysis found 435 contracts worth £699m; six months later, its tracker found 521 contracts worth £1.18bn; BidSkim found 555 worth £505m. Four choices explain most of the gap, and each is about the measurement, not the market. The first is what counts as AI: Tussell itself acknowledges that "measuring what is and isn't AI is highly subjective", and BidSkim finds that 37% of the AI purchases it includes would be missed by a search for AI terms. The second is what is counted: notices, contracts, procedures or awards. The third is which value: obligated, awarded, estimated, or the ceiling of a framework agreement. The fourth is revision: a tracker that improves its method changes its past figures.

The two kinds of tool fail in opposite ways. **Keywords are cheap and repeatable, but blind to meaning**: they miss AI described in other words and count records that merely name it (chapter 4). **Language models read meaning, but their result is hard to check**. A correction means paying for a full reading again, and the same contract can get a different answer on another run. Model errors are also not random, so they bias estimates even at 90% accuracy ([Egami et al., 2023](#)). None of these measurements publishes how often its classification is right. The methods to do so are known: define the concept, check a sample by hand, measure agreement between readers, and report the whole procedure ([Desai, Card and Jacobs, 2026](#)). The gap is one of practice, not of technology.

What this report adds

Classifying procurement text by meaning is not new. Sentence embeddings have been used to find circular-economy purchases in TED, with 800 cases on both sides of a similarity cut-off checked by hand ([Lepers et al., 2026](#)), and to find sustainability criteria in Swiss tenders ([Rolshoven et al., 2025](#)). The Open Contracting Partnership has piloted the approach for green procurement ([McKinney and Tarnay, 2024](#)). An open model tuned on Spanish notices retrieves the right CPV division for 68% of them ([openplacsp-e5-small](#)). [BidSkim](#) is the closest work on AI and reaches the same diagnosis on keywords. This report adds four things, applied to AI across TED and five national registers, small contracts included:

- a gold standard (a random sample reviewed in full) drawn across the whole ranking by meaning, so that recall (the share of all AI that a method finds, and so how much it misses) is estimated with intervals, and not only precision;

- a written codebook anchored in the AI Act, which keeps doubtful cases as a range rather than forcing them either way;
- a fixed, open embedding model, so that the search by meaning can be rerun exactly;
- comparisons between sources made at a common cut (one contract value applied to every source; chapter 3) and in a stated unit (record or procedure).

The classification step also uses a language model, the review agent, so the criticism above applies to it: it is not deterministic, and the same input can give a different answer. That is why every code is recorded, the author reviews it, and the agent's stability is tested (annex B).

3. Data and method: four steps, each with its check

Key terms

- **Record:** one row as a register publishes it (a notice, an award, a lot, an order). **Procedure:** one purchase, which a register may publish as several records. **Purchase** is used for either where the difference does not matter; where it does, the figure says which it counts.
- **Keyword hit:** a record whose text contains an AI term in some EU language. Records are *scanned* for keywords, *searched by meaning* (ranked by an embedding model) and *reviewed* (read by a person or the review agent).
- **AI-related purchase:** one that buys an AI system, an AI course (literacy) or AI governance. **AI-system purchase:** the subset that buys a system meeting the AI Act's definition, as the object or as a component. Unless stated otherwise, "AI purchase" means AI-related.
- **Codebook:** the written rules that decide what counts as AI (annex A). **Review agent:** a language model (Claude) instructed with the codebook; the author then reviews its codes. **Code:** a reviewer's classification of a purchase. A **firm** code is AI or not AI. **Doubtful:** the text cannot settle it; doubtful cases are kept with the way they lean, and counts are given as a range.
- **Closeness band:** a stratum (layer) of the ranking by meaning: the top 1% of records without AI words, 1-5%, 5-25%, and the rest.
- **Gold standard:** 2,317 purchases drawn at random in every closeness band and source, reviewed in full, against which every method is checked. Drawing within every band means the sample also looks far down the ranking, where AI is rare. **Precision:** of what a method marks as AI, the share that is AI; low precision means a count is inflated. **Recall:** of all the AI there is, the share a method finds; low recall means a count misses

AI. 95% interval: the range within which the true figure is likely to lie, given that the gold standard is a sample. It is not the same as the range from doubtful cases, and it is usually much wider.

The four steps

Every figure in this report comes from one of four steps (table 3.1), and each step has its own check, reported next to it. The gold standard is not a step. It is drawn once, at random across the whole ranking, and the error of every step is measured directly on it, not on the output of the step before, so no check depends on an earlier step being right.

Step	What it does	Check	Where
1. Find	Keywords in every EU language mark the records that name AI; a ranking by meaning orders the rest by closeness to descriptions of AI	Recall: of all the AI in the gold standard, the share found at each depth of review, and so how much is missed	Table 4.2, table A.2
2. Classify	The codebook decides which candidates are AI, applied by the review agent and then by the author	Precision: how many of the purchases coded AI are AI; agreement between two readers who cannot see each other's codes, on purchases not reviewed before; stability: whether the agent gives the same codes when rerun or reworded	Table 4.2, annex B (table B.4)
3. Count	Keyword counts are corrected with the precision and recall of steps 1 and 2; the unit, record or procedure, is stated	95% intervals: how far the count could move because the gold standard is a sample; the grouping of records into procedures, checked on samples	Tables 4.1 and 4.3, table A.4
4. Compare	Sources are compared only at a common cut, EUR 140,000	The number of sampled purchases behind each figure: the fewer, the less a difference between sources can be trusted	Table 4.3, table A.5

Table 3.1. The four steps of the method and the check that tests each.

TED shows the four steps at work. Of 209,139 IT notices, 3,627 name AI (step 1). Of those, 65% are AI-related and 59% buy an AI system (step 2). Counting what keywords miss, about 4,260 to 4,780 records are AI-related, possibly more than the keywords find (step 3). From EUR 140,000, about 1.1% of TED's IT notices are AI, against about 1% to 2.5% in Spain, the Netherlands and France (step 4; Italy and Germany in table 4.3).

The sources

The six sources differ in the smallest contract they record, and that decides which part of the market each can show (table 3.2). TED covers every member state in one format, but only above the EU thresholds. It is also the source behind most cross-country studies of procurement, so its blind spots become theirs. The national registers of Spain and Italy reach contracts of a few hundred euros. France's starts at EUR 40,000. Those of Germany and the Netherlands publish mostly above the EU thresholds.

Source	Country	Smallest contract recorded	Period	Records scanned by keywords	That name AI
TED, Tenders Electronic Daily	EU and EEA	EU thresholds (EUR 140,000 to 221,000 for supplies and services)	2024-Oct 2026	209,139 IT notices; 341,954 health notices	3,627 in IT; 890 in health
PLACSP	Spain	EUR 0, including minor contracts	2024-Oct 2026	3,444,260	3,870 in IT; 169 in health
ANAC	Italy	EUR 0, including direct awards	2024-2025	9,047,179	3,908
DECP	France	EUR 40,000	2024-Oct 2026	802,356	888
Datenservice Öffentlicher Einkauf	Germany	As published nationally	2024-Oct 2026	611,312	1,680
TenderNed	Netherlands	As published nationally	2024-Sep 2026	49,258	138

Table 3.2. The six sources. PLACSP is Spain's Public Sector Procurement Platform; ANAC, Italy's anti-corruption authority; DECP, France's essential public-procurement data. Thresholds by year and licences are in annex A, and the sets also searched by meaning in table A.1.

What a register leaves out matters, because by number most AI purchases are small. In Spain, minor contracts (awarded directly below EUR 15,000 for services and supplies) are 77% of the AI purchases of the Spanish dataset of chapter 5 and about 4% of their value

(chapter 4); figure 1 shows the same pattern for the records that name AI.

Most records that name AI are below EUR 140,000, but they carry little of the money

Share of records that name AI (a keyword count, not corrected) below EUR 140,000, roughly the EU threshold for TED, 2024-2026

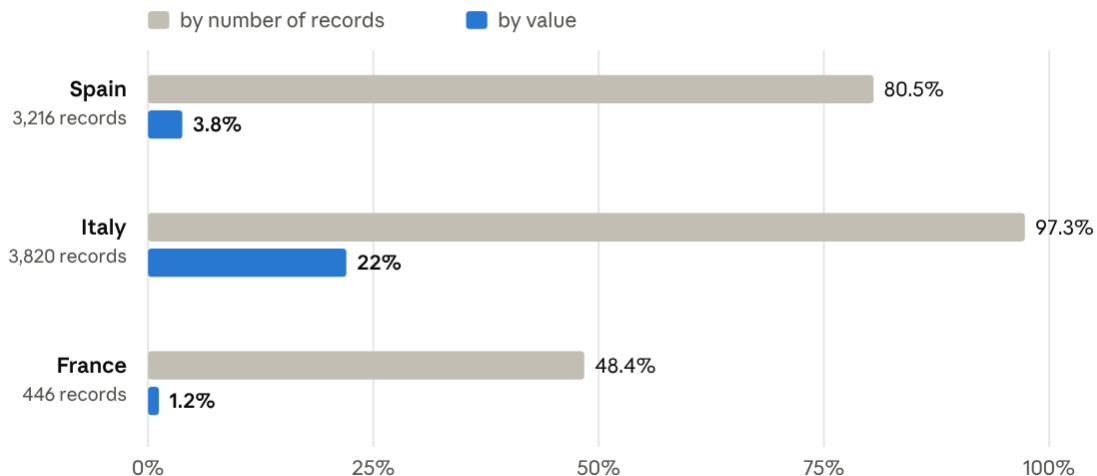


Figure 1 · PLACSP and TED (Spain), ANAC (Italy), DECP (France) · records that name AI, 2024-2026 · framework agreements excluded in Spain and Italy

Only public procurement is covered: banks, funds and private companies do not appear. A record may be a tender, an award or a modification, which is why the unit of each count is stated (step 3).

Step 1. Find

Keywords in every EU language mark the records that name AI. A search by meaning ranks the rest for review. An open-weights embedding model (BAAI/bge-m3, at a fixed version) places each record in a space of meaning, and each record is ranked by its closeness to published descriptions of AI and of ordinary IT. The ranking decides what is reviewed first, not what is counted. Because the model is multilingual, one set of descriptions finds the same kind of purchase in every language, whereas keyword lists fall into each language's traps (chapter 4).

In IT purchases, reviewing the top 1% by meaning adds one to three points of recall to keywords (table A.2), because keywords already find most of the AI there. Reviewing down to the top 5% by meaning raises recall to between 63% (TED) and 100% (France). The cost is reviewing 1.7 to 3.8 times as many records (table A.2). The value of the search by meaning lies in ordering the review, and in purchases that name no AI. Figure 2 shows the whole procedure. Records confirmed by hand return as seeds (example descriptions that steer the ranking) for the next round, while the gold standard is used only to measure.

Seeds learn from what is confirmed; the gold standard only measures

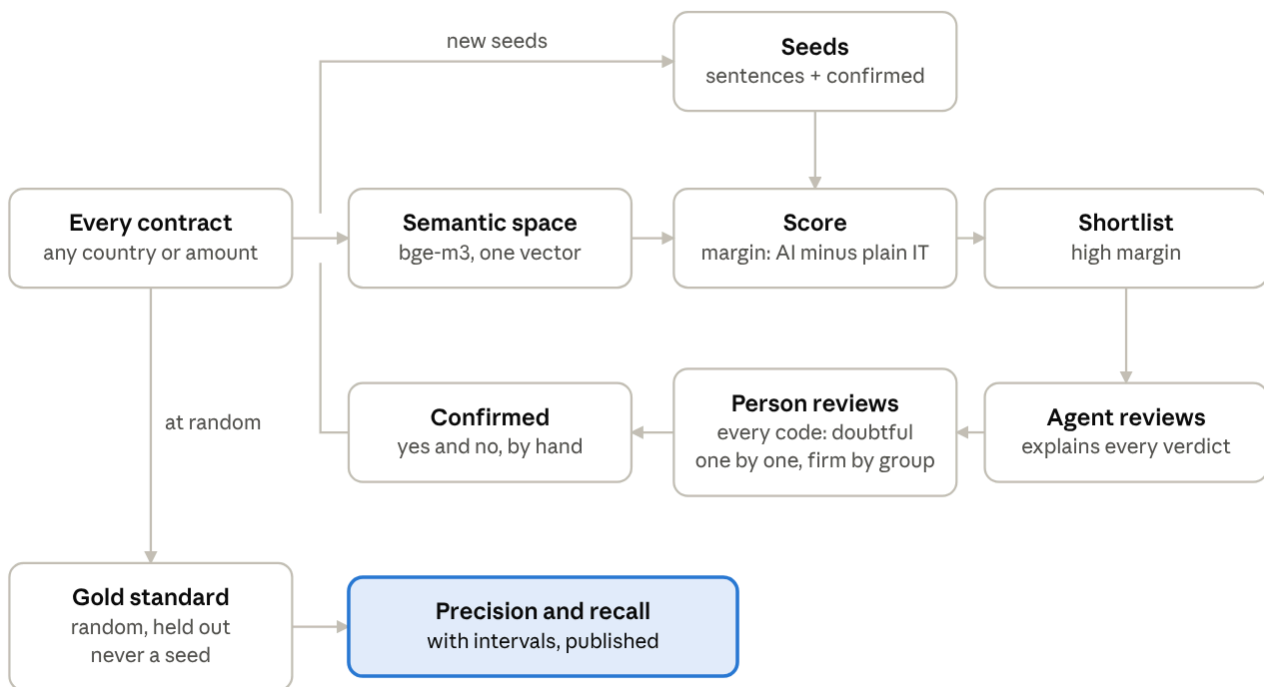


Figure 2 · The method at a glance · search by meaning, review, and a held-out gold standard

Step 2. Classify

What counts as AI is written down before it is applied. The first condition is the AI Act's definition of an AI system (Article 3(1)), with the Commission's guidelines on it. The Act regulates systems, not purchases, so a codebook settles the boundaries the Act does not. The codebook also gives every AI purchase a role: the object bought, a component of another product, infrastructure, literacy or governance (annex A). It narrows the operational definition of the Joint Research Centre ([Samoili et al., 2020](#)) to what a notice can show. Table 3.3 gives four boundary cases.

What the notice buys	Decision	Why
GPU servers for a research centre	Not AI; counted apart as compute (computing hardware)	Hardware infers nothing; it may run AI or anything else
An AI course for public staff	AI-related (literacy), not an AI system	It builds AI capacity (Article 4) but buys no system
Text recognition (OCR) for archives	AI system if the notice names learnt recognition; doubtful otherwise, with its lean	Template OCR is classical; modern recognition is learnt, and notices rarely say which

What the notice buys	Decision	Why
Licence-plate cameras for a low-emission zone	AI system (object)	Reading plates from images is learnt vision, whatever the notice calls it

Table 3.3. Four boundary cases under the codebook.

Which definition is used is one of the choices that change a count most. Of the records that name AI, 18% (Italy) to 59% (TED) buy an AI system, but 35% to 65% are AI-related (table 4.2). The gap comes above all from AI courses, which are 28% of the records that name AI in Spain and 32% in Italy.

Every code answers one question: what does the notice show? "Not AI" means not AI on the text, not that the product contains none, and no reader can do better with the same words. That is why the counts of this report are a lower bound on the AI that products contain (chapter 8), and why the first recommendation is a label set at the source.

The review agent codes each candidate under the codebook, gives a reason, and groups the cases by reason. The author then reviews every code: the doubtful ones one by one, the firm ones by group. Agreement between two readers is measured with Cohen's kappa, where 1 is full agreement and 0 is what chance alone would give (it can fall below 0). Reviewed with the agent's proposal in view, the author changed 0.8% of the agent's firm codes, a kappa of 0.97. That figure measures the review as it was run, not the reliability of the codes, because the second reader saw the first reader's codes.

Reliability is measured blind. On 186 purchases that no review had seen, an independent agent coded all of them and the author 136, without seeing each other's codes. Where both gave a firm code (110 purchases), they agreed on 78%, a kappa of 0.57: moderate agreement. Because hard groups were read in full and easy ones only in part, this figure describes the harder cases more than the population. The two tests were then crossed: for every disagreement, the agent set out both codes and the reason each reader gave, and the author reread the case under the codebook. Of the 24 firm disagreements, 7 are umbrella categories that the text cannot settle and are treated as doubts. In 14 of the other 17, the blind code was the author's misreading under the written rules, such as a hotel night for a speaker at an AI seminar coded as AI; in 3, the agent had read AI that the author passed over. On these cases rereading under the codebook favoured the agent's code, so the gold standard keeps its codes. The rereading itself was done with the agent's reasons in view, the very condition this check examines: in a review that combines a person and an agent, both sides carry bias, and no step removes it; it can be exposed and measured (annex B). The blind check also shows what reviewing with a proposal in view loses: doubt. Read blind, 24% of purchases ended in doubt (22% when weighted to the design of the gold standard); in the gold standard, 1.3% did. The ranges of chapter 4 are

therefore narrower than a blind reading would give, by up to 6 points of keyword precision (table 4.2). The agent itself is stable: two runs of the same prompt gave identical codes, and a reworded prompt or a smaller model changed 4% to 7% of codes. Annex B gives the design, the agreement measures (table B.4), a third reader, and the stability tests.

Step 3. Count

A keyword count is corrected with the error measured on the gold standard. The estimated number of AI purchases is the number of keyword hits times precision, divided by recall: multiplying by precision removes the hits that are not AI, and dividing by recall adds back the AI that keywords missed. Both rates are read the same way: doubtful cases count as not AI for the low end of the range, and as AI for the high end. Each count therefore carries two spreads. The range comes from doubtful cases. The wider 95% interval shows the uncertainty from sampling, and comes from resampling within closeness bands (redrawing the gold standard from itself many times and recomputing the count). Counts are of records unless stated otherwise, as keyword counts of notices are; where the unit changes the result, the count per procedure is given too. Registers carry no common procedure identifier, so records are grouped into one procedure when they share buyer and text. The grouping was checked on two samples: it rarely merges separate procedures but often leaves one procedure as several records, so counts of procedures are upper bounds (table A.4).

Step 4. Compare

A share needs its denominator, and a comparison between sources needs the same part of the market. Shares are therefore given within the IT set of each source: CPV divisions 48 (software) and 72 (IT services), plus any record that names AI elsewhere, which makes shares of the set slightly high. Sources are compared only at a common cut of EUR 140,000, the lower EU threshold for central government (table 4.3; detail in table A.5). Records without a usable amount are left out of both the count and the denominator, which matters most for Germany and TED (table A.3). Where an estimate rests on few sampled purchases, their number is given with it, because a share estimated from a handful of purchases can move a long way.

Open data

The report is published with the codebook and the error patterns (annex A); every review decision, with who took it, when and why; the gold-standard samples and the estimates drawn from them; and the AI purchases found. These, with the 3.7 million purchases searched by meaning, can be explored in the Arquestra AI Procurement Observatory

(arquestra.dev/assets/ai-observatory), where the part of the random sample that it holds can be filtered. The review sheets and the scripts that build every figure are deposited on Zenodo with the report, at <https://doi.org/10.5281/zenodo.23561547>.

The observatory also gives every purchase that was not reviewed an estimated classification from a separate model. Annex A explains how that model was tested; its classifications are not used in the figures of chapters 3 and 4. The method is not specific to AI: the same steps apply to any category of purchase that the registers do not identify.

4. What the registers do not show

A register can miss AI in four ways. Three lie in the text of the notice. When AI is named, keywords count it wrongly, because they also catch notices that only mention the word. When AI is described in other words, or bought by brand name, keywords miss it, but a search by meaning finds part of it. When AI is built into a product and the notice does not describe it, no method can find it from the text; it can only be bounded. The fourth is one of reach: most AI purchases are small, and TED does not see them. The chapter first gives the scale and the keyword error, then compares sources on the same terms, and then takes the other ways in turn.

The scale

Each source holds up to a few thousand records that name AI, and the estimated number of AI-related records is of the same order (table 4.1). Their value is another matter. Most TED and German notices give no amount. Some records hold the ceiling of a framework agreement rather than the cost of an order: in Italy, three such records hold EUR 2.3 billion of the EUR 2.6 billion recorded for contracts that name AI. From the registers as they are, the number of AI purchases can be stated within a margin, but their value cannot (table A.3).

Source	Period	Records that name AI	AI-related records, estimated [95% interval]
TED	2024 to Oct 2026	3,627	4,260 to 4,780 [2,670-7,210]
Spain, PLACSP	2024 to Oct 2026	3,870	2,870 to 3,370 [2,280-4,210]
Italy, ANAC	2024 to 2025	3,908	2,660 to 3,700 [2,030-5,130]
France, DECP	2024 to Oct 2026	888	380 to 540 [290-790]

Source	Period	Records that name AI	AI-related records, estimated [95% interval]
Germany	2024 to Oct 2026	1,680	970 to 1,260 [760-1,570]
Netherlands	2024 to Sep 2026	138	75 to 95 [45-135]

Table 4.1. AI-related records in absolute terms, in the IT set of each source. The estimate runs from firm codes only to counting doubtful cases as AI; the interval runs from the low end of the first to the high end of the second. Periods and smallest contracts differ (table 3.2), and national purchases above the EU thresholds also appear in TED, so rows are neither added up nor ranked; they follow the order of table 3.2, not their size.

Named AI, counted wrongly

Table 4.2 compares, on the same random sample, what a keyword count finds with what is there. Keyword counts are inflated. Of every 100 records that name AI, only 35 (Netherlands) to 65 (TED) are AI-related purchases, and 18 to 59 buy an AI system.

Keyword counts are also incomplete. They miss 4% (France) to 45% (TED) of the AI bought, and more where texts are short or multilingual (TED, Italy, the Netherlands); counting the doubtful cases as AI, they miss about half in TED. Much of what they miss sits where no one would look: in the lowest closeness band, the three quarters of each IT set farthest from AI by meaning, a review of 340 records per source found AI such as a speech-recognition model to assess spoken Estonian and a licence-plate recognition system for the Kraków police, described as a "solution" or a "system" (annex A).

In TED the two errors may not cancel: the estimate, about 4,260 to 4,780 AI-related records, is above the 3,627 records that name AI, though its interval (2,670 to 7,210) still includes them, and the difference rests on two records in the lowest closeness band, where each stands for about 450. In every other source the false hits weigh more. By value, the error in TED is smaller: in the sample, 94% of the value of keyword hits is AI, a figure that rests on a few large records, because the false hits (conference catering, books, staffing, compute) are small purchases.

IT set	Keyword hits that are AI-related (firm / counting doubtful as AI)	Of the hits, AI systems (firm)	Share of all AI that keywords find: firm [95% CI] / counting doubtful as AI	Top 1% by meaning, without AI words, that is AI
TED (EU)	65%	59%	55% [36-89] / 49%	5%

IT set	Keyword hits that are AI-related (firm / counting doubtful as AI)	Of the hits, AI systems (firm)	Share of all AI that keywords find: firm [95% CI] / counting doubtful as AI	Top 1% by meaning, without AI words, that is AI
Spain	61%	31%	82% [68-96] / 70%	8%
Italy	53%	18%	78% [61-94] / 56%	2%
France	41%	25%	96% [89-100] / 67%	1%
Germany	49%	44%	85% [74-94] / 66%	13%
Netherlands	35% / 40%	32%	66% [43-96] / 59%	3%

Table 4.2. Keyword error measured on the gold standard, per record. The gold standard is 2,317 purchases drawn at random with known probabilities and reviewed in full: per source, 100 keyword hits and 300 records without AI words in four closeness bands (217 without AI words in the Netherlands). A single value means there were no doubtful rows among the hits. Intervals come from resampling within closeness bands; where one AI purchase is found in a large band, it stands for hundreds or thousands, and the interval says so. Recall after also reviewing the top 1% by meaning, and the records that would then need reviewing, are in table A.2. Read blind, the author left as doubtful, leaning to AI, about one in ten keyword hits that the gold standard codes as not AI (annex B); counting those as AI would raise the upper end of precision by 4 to 6 points.

Records, not procedures. These rates count records, and a register can publish one procedure as many records. In France, France Télévisions fills 474 of the 888 records that name AI (53%). Of these, 443 are admissions of suppliers to its dynamic purchasing system for "Data and AI needs" (an open list of suppliers from which orders are later placed), almost all with a token amount of EUR 0.1 or 0.2. On the 114 French rows read in the gold standard and the blind check, a wider base than table 4.2, the share of keyword hits that are AI-related is 48% counted by record but 81% [69-92%] counted once per procedure. In TED, on the same kind of base, it hardly moves (68% by record, 70% by procedure). But TED's 3,627 records are at most about 1,930 procedures, because a notice and its award often repeat the same title. Any count of public AI should say which unit it counts (annex A, table A.4).

Comparing sources at a common cut

A keyword count is wrong by a different amount in each country. On TED's own phrase search (figure 3), Sweden, the Netherlands and Estonia appear to buy the least AI. That is not because they buy less AI. It is because their notices often write the bare acronym "AI", which a phrase search does not count. The report's own word list, which adds the bare acronym and other spellings, finds about as many IT notices naming AI in total (3,627 of the 209,139 IT notices in the report's download, against 3,579 of 209,778 counted by TED's search; figure 3 leaves out four countries with under 1,000 IT notices each), but places them differently: 129 in the Netherlands instead of 23, 661 in Germany instead of 495, and 148 in France instead of 200. Any ranking of countries built on keywords partly measures how each language names AI.

By keywords alone, AI is under 2% of TED's IT notices, and the ranking partly measures how each language names AI

IT notices (CPV 48, 72) found by TED's phrase search · 26 countries shown: 3,524 of 207,472 (1.7%) · 2024-Oct 2026

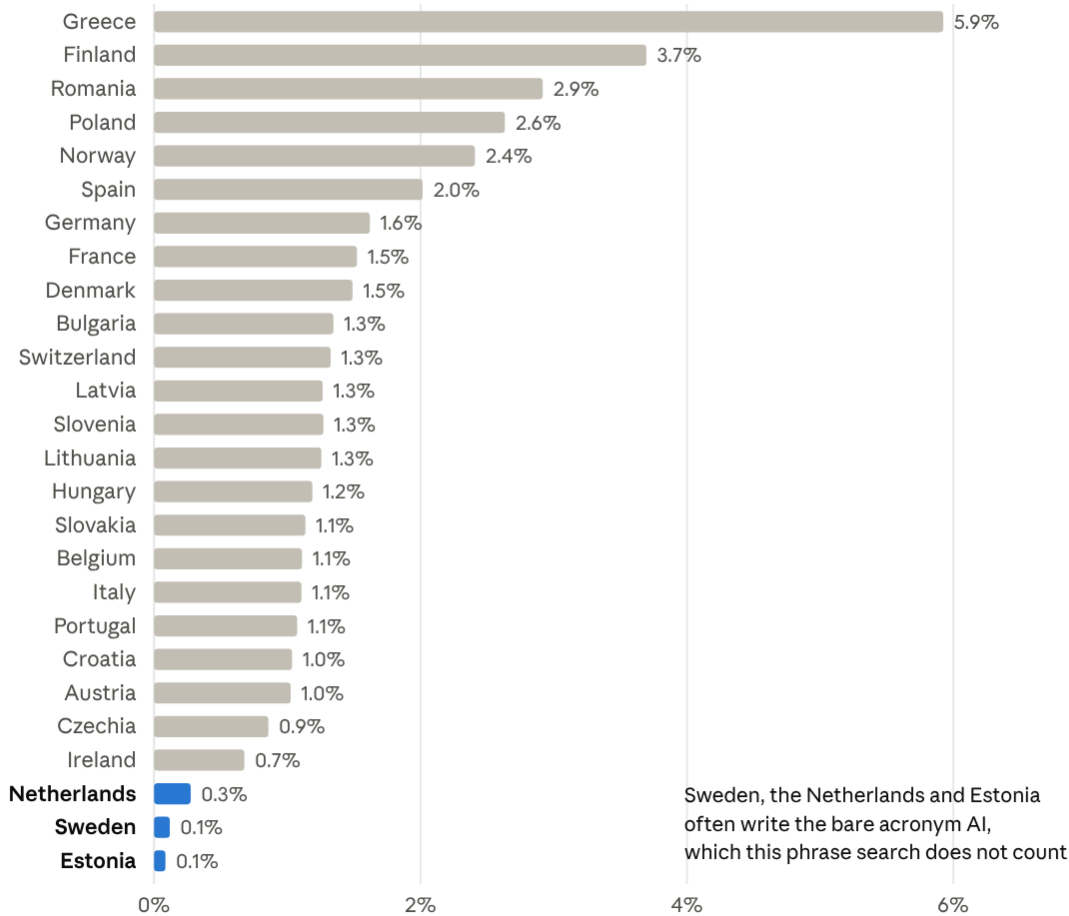


Figure 3 · TED, IT notices (CPV 48 and 72) and those found by TED's own phrase search for AI in 32 phrases, January 2024 to October 2026 · a keyword count, shown for its bias between countries · not a ranking of countries

Table 4.3 gives AI within IT purchases, corrected with the gold standard and then cut at the same amount, so that every source shows the same part of the market.

IT set	Records	Naming AI	AI, estimated (firm / counting doubtful as AI) [95% interval]	From EUR 140,000: naming AI	From EUR 140,000: AI, estimated [95% interval]
TED (EU)	209,139	1.7%	2.0% to 2.3% [1.3-3.4]	1.6%	1.1% [0.7-1.6]
Spain	60,223	6.4%	4.8% to 5.6% [3.8-7.0]	2.4%	2.0% to 2.5% [0.0-5.7]
Italy	136,959	2.9%	1.9% to 2.7% [1.5-3.7]	1.1%	0.9% to 4.8% [0.0-13.3]
France (from EUR 40,000)	25,845	3.4%	1.5% to 2.1% [1.1-3.1]	1.8%	1.0% to 1.2% [0.5-1.8]
Germany	25,871	6.5%	3.7% to 4.9% [2.9-6.1]	5.7%	4.5% to 6.1% [1.9-10.8]
Netherlands	3,840	3.6%	1.9% to 2.4% [1.2-3.5]	2.1%	1.0% to 1.1% [0.5-1.7]

Table 4.3. AI within the IT set of each source, for all records and for records from EUR 140,000 with a usable amount (detail, bases and sampled purchases in table A.5). Not a ranking of countries. The first columns mix registers that reach different parts of the market; the last two compare like with like, with wide intervals where few sampled purchases lie above the cut (Italy: 48). Below the EU thresholds, national data are generally not comparable, because publication thresholds vary by country ([Fazekas and Czibik, 2021](#)).

Much of the gap between rows comes from how far down each register reaches. At the common cut, Spain's share of IT purchases naming AI falls from 6.4% to 2.4%, and Italy's from 2.9% to 1.1%. In every source but Germany, 1.1% to 2.4% of IT records then name AI, and with these samples the estimates of Spain, Italy, the Netherlands, France and TED cannot be told apart. Germany's rests on a thin base: only 16% of its IT records carry a usable amount, and on those its estimate is higher than those of TED, the Netherlands and France. Even TED, one register with one set of EU thresholds, shows only a narrow spread among the five countries studied: on the report's own word list and before the common cut, the share of their IT notices naming AI lies between 1.1% (France) and 2.2% (Germany), against 0.3% to 2.0% on TED's phrase search in figure 3. Either way it is a keyword share, and the two word lists do not even agree on the order. Small contracts are where Spain and Italy buy most of their AI; above the common cut, the countries studied here differ far less.

Why keywords and CPV codes fail

Keywords catch records that are not AI in more than one way. Some hits are linguistic traps, all corrected before any figure was used. In Italy, a first pass over 14 months of contracts returned 80,866 "AI" contracts, because "ai" is an Italian preposition written in capitals ("AFFIDAMENTO AI SENSI DELL'ART. 50"). In France, "(AI)" marks a lot in building works; in the Netherlands, it is the prefix of Amsterdam's tender numbers ("AI 2023-

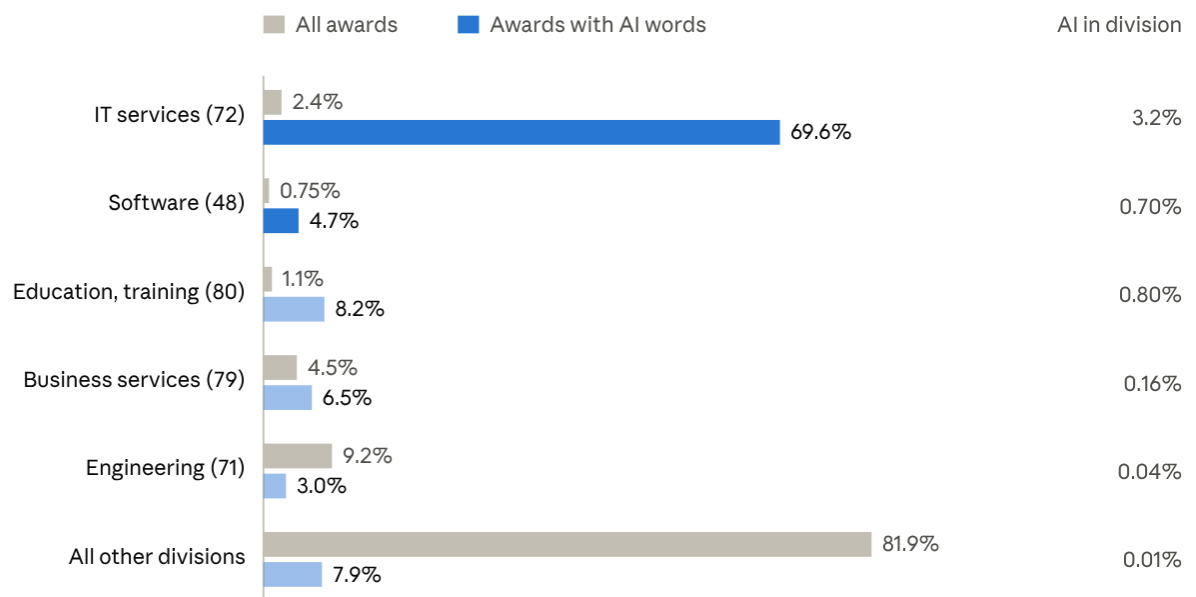
0147"); in Germany, "AG KI" is Kiel's local court. The stems "generativ" and "generativn" matched degenerative conditions in Italian and Croatian. Each language needs its own word list, audited by looking at what it found.

Other errors come from the role of AI in the purchase. Spanish research foundations pay to publish articles such as "A Machine Learning Approach" to mortality risk: AI is the topic, and nothing AI is bought.

Nor do the classifications that tenders carry solve the problem. The CPV was last revised in 2008 ([KPMG Law, 2021](#)) and has no code for AI; the closest are software (48) and IT services (72). In eForms, the EU's standard notice forms, the field for "strategic procurement" (BT-06) has four values: environmental impact, social objectives, innovative purchase and none ([OCDS for eForms](#)). In France, 74% of contracts that name AI sit in the IT divisions of the CPV, which hold only 3.2% of all contracts; the other quarter sits in training, business services and engineering (figure 4). In Italy, 71% of all contracts carry no CPV code. And 29% of Italian health contracts are described in six words or fewer ("DM OSTEOSINTESI"), which no reader, human or machine, can classify.

In France, the IT codes hold 3% of awards but three quarters of those that name AI

Share of all awards and of awards with AI words, by CPV division. Right: how much of each division names AI.



A filter on the IT divisions keeps three quarters of them; the rest is training, business services, engineering.

Figure 4 · DECP (France), awards from EUR 40,000, January 2024 to October 2026 · 802,356 awards, 888 with AI words · records that name AI

Described in other words, and built into products

What keywords miss is not a hidden kind of AI; it is the AI people mean. In the review, most of the AI that no keyword found was bought by brand name. Examples are licences of Claude (two Spanish city councils, the Spanish public broadcaster, a university), ChatGPT

Plus and Copilot Pro for a hospital, ChatGPT Business for a national museum, Copilot Studio, Perplexity Pro, Google Agent Builder, and Limbus Contour for radiotherapy. Next came speech recognition, licence-plate cameras, and network-security products that learn from traffic. Some AI is named only by its technique: a Murcian health service tendered a "clinical coding solution using natural language processing" without ever writing "artificial intelligence". A list of words looks for the technology; the state buys the product.

Outside IT, AI is rarely named at all, and where it is present it is built into something else. Samples were drawn, in four registers, from purchases outside the IT divisions of the CPV that name no AI. In them, purchases with a firm AI code were almost absent: none in France or Spain, and very few in the German and Dutch samples, though in Germany one sampled purchase stands for hundreds (table A.6). The picture changes only if every product that may carry AI is counted, and then the upper bound is large (table A.6).

That upper bound is almost entirely new vehicles. Since July 2024, the EU General Safety Regulation ([Regulation \(EU\) 2019/2144](#)) has required intelligent speed assistance on every new car, van, bus and truck, and emergency braking on new cars and vans. Both often rely on learnt vision. The same applies to MRI scanners with deep-learning reconstruction, milking robots that find teats with cameras, and counter-drone systems. None of these notices says so. Much of this equipment may meet the Act's definition, but the high-risk obligations reach only some of it. For vehicle safety systems, those obligations come through type-approval law (the EU rules for approving vehicle models). Providers carry their obligations whether or not a notice mentions AI. What is lost is the ability to check which public bodies deploy AI, and where.

None of this is fixed by a better search. A search by meaning finds what keywords miss, but it cannot find what a notice does not say, and a reviewer can judge a text but not a product. Only the supplier knows whether a product contains AI, and only the buyer can put that on the record. This is the case for the label of recommendation R1.

Most AI is bought small, below what TED sees

By number, most public AI purchases are small and direct (figure 1, chapter 3). In Italy, 93% of contracts that name AI are direct awards, and 65% are worth less than EUR 5,000. Below the EU threshold lie 97% of Italy's contracts that name AI and 22% of their value (without framework agreements). In France, 48% of the records that name AI and 1.2% of their value lie below it, on the 446 records with a usable amount. In the Spanish dataset of chapter 5, minor contracts are 77% of the Spanish purchases and 4% of their value. The median AI purchase there is EUR 526,362 in TED, EUR 57,327 in a national tender and EUR 7,000 in a minor contract.

Spain's own procurement supervisor states that there are still no aggregated data on minor contracts ([OIReScon, 2025](#)). Studies built on TED alone capture most of the money but miss most of the purchases. The small purchases are also the least scrutinised, and are made by the bodies with the least capacity to know whether a product contains AI.

Like those of this chapter, the figures in the rest of this report are not raw keyword counts. Chapter 5 starts from purchases that a classifier marks as AI, checked against the same gold standard, and asks how many of them buy governance. Chapter 6 asks what health buyers buy.

5. Governance bought as a purchase of its own is rarely visible

To govern AI, a public body first has to know that it has it. In this report, a governance purchase is one whose object is to govern AI that the buyer uses or will use. It may be an impact or risk assessment, an audit, support to comply with the AI Act, a strategy for adopting AI, or a course whose content is AI law, ethics or standards. General AI courses are counted apart, as literacy. Purchases by which a regulator governs AI in general, such as the Commission's AI Act service desk, are counted too, and shown apart where they matter.

This chapter does not compare countries. It uses TED for the large purchases of every country, and Spain, with all its data, as a case. Of the registers studied for this report, only Spain and Italy publish even their smallest contracts, and Spain is the one searched here for governance below TED.

The data for this chapter

The chapter uses a dataset built from TED and PLACSP (January 2024 to October 2026) and counted by procedure: the records of one purchase are merged into one procedure. A Spanish purchase published in both TED and PLACSP is merged when value, buyer and title match (from EUR 1 million); smaller duplicates may be counted twice. A search retrieved 6,463 candidate purchases. A logistic classifier, a statistical model that predicts from a purchase's position in the space of meaning whether it is AI, marks 3,831 of them as AI. Five confirmed governance purchases that it scores below its cut-off were added by hand, which gives a base of 3,836 (814 in TED, 3,022 in Spain). This chapter calls them the purchases marked as AI. Because the candidates come from a search, the dataset holds fewer TED purchases than the estimate of chapter 4 for all of TED's IT notices (about 4,260 to 4,780 AI-related records, table 4.1; the records that name AI are at most about 1,930 procedures).

The classifier was checked against the gold standard, on the 119 gold-standard purchases from TED and Spain that it also scored. It is right on 71% of what it marks as AI [95% CI 62-79%], and it finds 87% of the AI there [78-92%]. So, among the candidates of the search,

29% of the purchases marked as AI are not AI, and every rate below is a rate among them. The review agent and then the author reviewed 59 candidate governance purchases under the codebook, and confirmed 40 (15 of 19 in TED, 25 of 40 in Spain). Two of the 40 are preliminary market consultations rather than purchases.

How much governance

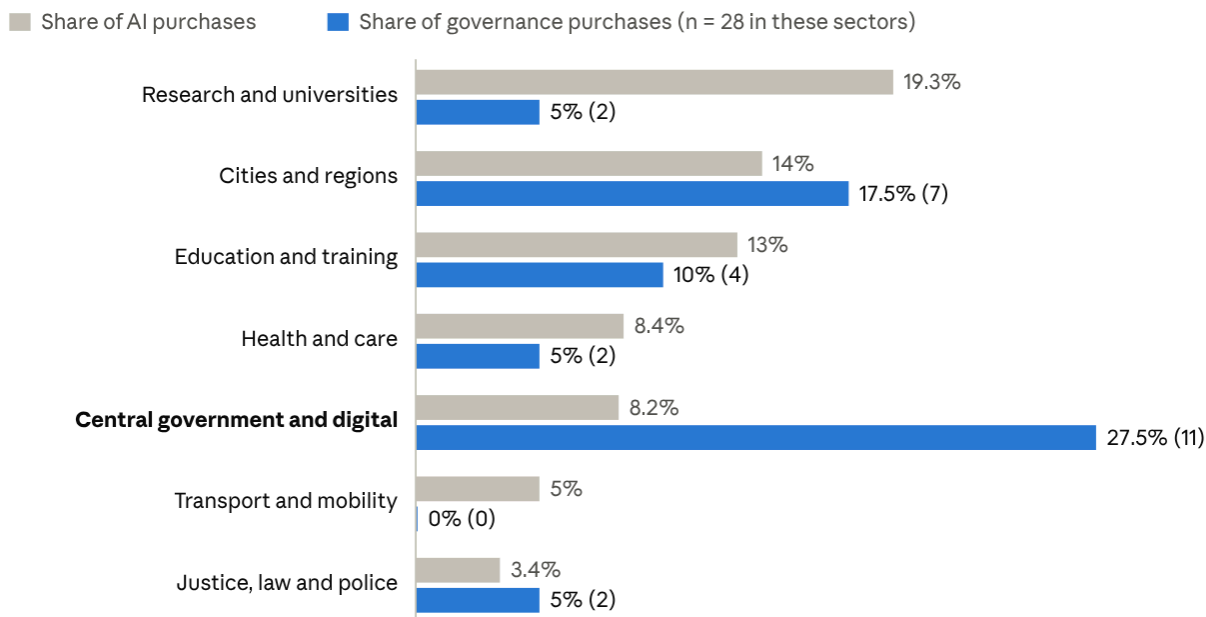
Of the 3,836 purchases marked as AI, 40 buy governance (1.0%; 35 of 3,831, or 0.9%, without the five added by hand). In TED it is 15 of 814 (1.8%). Five of these were bought by the European Commission: three for the AI Act and its own AI Observatory, and two as policy studies; without them, TED's rate is 10 of 814 (1.2%). In Spain it is 25 of 3,022 (0.8%), 22 of them minor contracts. That 1% is a floor. The gold standard holds seven governance purchases in TED and Spain, and the search had found two of them, a recall of 2 in 7 [95% CI 8-64%]. Two of the others lay outside the dataset the search covered, and three inside it were missed. In the random sample of TED's and Spain's IT purchases, weighted by the probabilities with which it was drawn, governance is about 4% of AI purchases (95% interval 1.4% to 7.8%), about 1 in 25; unweighted, it is 7 of the 150 records coded AI-related in those samples after compute and staffing were recoded (annex A). What this chapter can claim is the pattern, not the rate. Governance is a small and unevenly bought share of AI purchasing, and a register is a poor place to find it.

By value, governance is 2.9% of the EUR 3,138 million recorded for AI, but most of the value recorded for AI in this dataset (60%) is framework ceilings, not spending. In governance, 83% of the value is the ceiling of a single French framework of EUR 75 million (CANUT: training, audits, legal and security advice on AI); without it, governance is 0.5%. A governance purchase is as small as a typical purchase marked as AI: EUR 13,000 at the median, against EUR 13,550.

Who buys it

Central government buys a quarter of AI governance, research almost none

Share of AI purchases and of governance purchases by sector of the buyer



40 governance purchases in all: a pattern to test elsewhere, not a finding. Only central government differs after correction.

Figure 5 · TED and PLACSP, January 2024 to October 2026 · 3,836 purchases marked as AI and 40 governance purchases, by sector of the buyer (the seven largest named sectors) · data of table A.7 · small base

In these data (figure 5; table A.7), central government buys governance well above its share of AI purchases (11 of the 40, two of them the Commission's, against 8.2% of the purchases marked as AI; the other three Commission purchases fall in other sectors). Research buys it well below its share (2 of the 40, against 19.3%). The pattern was seen after looking at the data, on a small base, and with sector assignments that come partly from a classifier. It is therefore stated as a hypothesis to test on other countries (annex A).

What a governance purchase looks like

The clearest case cost EUR 1,500: the government of the Canary Islands paid for a legal assessment of its own virtual assistant, LEO, against the AI Act. It is the purchase this report looks for: a buyer that deploys AI and checks it. The City of Lyon tendered a conformity assessment for data protection and the AI Act. MAZ, a mutual insurer that works with the Spanish social security, paid EUR 12,000 to define its AI governance framework. ISDEFE, Spain's state defence engineering company, paid for advice on the strategy and the rules for its use of AI, and trained staff in ISO 42001, the management standard for AI.

About a third of the governance purchases are courses and talks for civil servants on the AI Act, AI ethics or ISO 42001, at EUR 1,490 to 5,400 each.

After the CANUT framework, the largest by value were bought by the regulator. The Commission's directorate for digital policy (DG CNECT) paid EUR 7.4 million for technical assistance on AI safety, EUR 2.5 million to support the AI Observatory and EUR 2 million for an AI Act service desk. And one purchase is a model for recommendation R3. RESAH, the French hospital purchasing network, offers a EUR 800,000 AI compliance platform through a framework that hospitals in its network can use without a tender of their own.

TED shows governance mostly when a central purchasing body or a regulator buys it in bulk, as CANUT, RESAH and the Commission do. The purchases by which a single body checks its own AI, such as LEO's assessment, are small: they sit below the EU threshold, as minor contracts, and only a register that publishes minor contracts, such as Spain's, can show them.

6. Health, and whether December 2027 is realistic

Health is where the public sector buys much of the AI that the Act treats as high-risk. It is 8.4% of the purchases marked as AI in TED and Spain (chapter 5, table A.7): auto-contouring for radiotherapy, AI reading of mammograms and chest scans, speech recognition for clinical notes, and clinical decision support. Much of it is bought under a product name or inside a device, so it is the AI least visible in the register (chapter 4). Its governance, bought as a purchase of its own, is as rare as elsewhere: 2 of the 40 governance purchases. That is 0.6% of health AI purchases against 1.0% overall, both counts from the search and so both floors; the difference is easily due to chance.

Several explanations could account for governance so rarely appearing as a purchase of its own:

- it is written into the contract for the system;
- health has a regime of its own for medical products;
- it is not yet due;
- any count misses it, because health AI is bought by brand or inside devices;
- it is done in house, judged low-risk, or left to clinicians.

Only the first two can be tested with purchase data. The others are left to interviews and surveys, and what purchase data show is stated as compatible with an explanation, not as proof of it. Testing the first fully means reading tender documents, not notices, which this report does not do. A study of 20 selected AI tenders in health had to read their specifications to establish what AI was bought, and with what guarantees on data, transparency and accuracy ([Gutiérrez David and Quintana Cortés, 2023](#)).

The second explanation was tested by searching the health purchases of Spain and TED by meaning for certification, CE marking (the EU mark of conformity), conformity assessment or clinical validation of software or AI. The review agent reviewed 253 candidates, and the author checked every decision. Purchases to certify or validate an AI system are almost absent. In four purchases by three buyers, the medical-device regime governs AI, always as a condition of the product and never as a purchase of its own. A German hospital group requires its radiology and early-detection decision support to meet the Medical Devices Regulation (MDR). Valencia's health service requires an AI virtual assistant certified as a class IIb device. CIBER, a Spanish research network, requires the same class for the AI software of 51 eye scanners. Several AI imaging purchases carry no such condition. Among them is AI mammography in Castilla y León, tendered for EUR 1.32 million and awarded for EUR 0.89 million. These cases are compatible with the first two explanations: where governance of health AI shows in the register at all, it shows inside the specification of the product, not as a purchase of its own.

Certification and governance are not the same thing, and they fall on different parties. The MDR ([Regulation \(EU\) 2017/745](#)) asks the manufacturer to certify the product. The AI Act also asks the hospital that uses high-risk AI to look after it: a person who oversees it, use as the manufacturer instructs, and records of how it works (Article 26). For some uses, such as triage, the hospital must also assess the impact on patients' rights and register the use (Articles 27 and 49). And much health AI, such as appointment chatbots or transcription, is not a medical device at all. The register shows the manufacturer's side, when the tender asks for it, but rarely the hospital's: 2 of the 40 governance purchases are in health and care.

The deadlines follow the same split. From 2 December 2027, the rules apply to uses such as emergency triage or deciding who is eligible for public healthcare, and the hospital must assess their impact before use. From 2 August 2028, they apply to AI inside medical devices from class IIa upwards (the classes that need assessment by an independent notified body), and they fall mainly on the manufacturer. Systems already in use by public bodies must comply by 2 August 2030, or earlier if their design changes significantly. Both the 2027 and the 2028 dates come from the 2026 amendment of the Act (chapter 1). That amendment also lets the Commission relax the Act where medical-device law already covers the same ground; whether it will do so is still open.

Is December 2027 realistic? Procurement data cannot tell. With thirteen months to go, the register cannot say which public bodies use such systems, governance is rarely bought, and much of the relevant AI is not described as AI. A label introduced now (recommendation R1) would show where the first impact assessments are due.

7. Discussion and recommendations

The problem this report documents is not that the public sector buys high-risk AI, but that no one can tell when it does. Europe's rules for AI are written as if someone could. The purchases that add most to the uncertainty of the report's estimates are the ones the register describes least. Examples are a pantograph monitoring system for trains (the pantograph is the arm that draws power from overhead lines), intrusion detection for traffic systems, and an assistant in an operating theatre. Each lies in a closeness band where one notice stands for thousands. A notice of a few words, with no CPV code and a framework ceiling instead of a price, cannot say what was bought. Still less can it say whether the purchase contains AI, or whether that AI is governed. The same blindness affects those who count public AI. The published counts reviewed in chapter 2 rest on keywords or language models, and none of them says how often its classification is right (table 2.1).

Measuring well need not be expensive. The search by meaning ran on one laptop graphics card (annex A). Review agents review only the shortlist, and a person audits their codes on random samples with known probabilities. The approach is not specific to AI. A search by meaning, with a published codebook and a measured error, serves any category that the CPV marks badly, from green purchases to cybersecurity.

The limits of this report (chapter 8) point to the next question. Some purchases are hard to classify because the product hides its AI, some because the notice says too little, some because the definition does not settle the case, and some because the reviewer errs, whether a person or a model. A follow-up study will test several language models and several review procedures against the same gold standard. It will ask when a public body can trust a model to classify what it buys, and with how much human review.

The method can also gain a second stage. The review found the kinds of purchase that are not AI systems, such as courses, compute, staffing and AI as a mere topic. A second pass could sort every candidate into those categories, in the same space of meaning, before anyone reviews it. Social science calls this keeping discovery apart from measurement: categories found on one part of the data are measured on another ([King, Lam and Roberts, 2017](#); [Grimmer, Roberts and Stewart, 2022](#); [Egami et al., 2022](#)). The random sample of this report already keeps the two apart, since none of its cases was used to search.

Watching procurement with AI

The order of the tools matters. Keywords find what is named, and fall for words: a Polish hospital's link to a "platform of intelligent services" counts as AI by keyword, though nothing in the notice is. A search by meaning avoids that trap and says where to look first. But it misses systems described in plain technical terms: a speech-recognition model to

assess spoken Estonian and a licence-plate reader for the Kraków police both sat in the band farthest from AI. A review agent, reading the notices, recognised both and set the platform aside. Hence the design of this report: meaning first, to order the review; an agent next, to read what meaning ranks high and a random sample of the rest; a person last, to audit the agent on samples of known probability. Keywords alone err in both directions (chapter 4). A language model reading every notice reads hundreds of thousands that need no reading, and still needs a human sample to know its error. This design is one way to oversee AI that reads procurement, not the only one; it is published in full, with code and review sheets, so that others can test it.

The same lesson bears on regulation. Using AI to monitor procurement, a "meta" use of AI for procurement, is, in Sanchez-Graells's reading, not addressed in the proposed EU Public Procurement Act (European Commission, 2026), and he argues that it should be treated as high-risk under the AI Act ([Sanchez-Graells, 2026](#)). Without a label, a supervisor could not even find which public buyers use AI to run or monitor their procurement: such a tool is bought, like any other AI in this report, under a title that does not say so. If it is made high-risk, overseeing it will take more than a person signing off an agent's codes. The AI Act requires high-risk systems to be designed so that those who oversee them stay aware of automation bias, the tendency to rely too much on their output (Article 14(4)(b)), and leaves the deployer to assign that oversight (Article 26). This report observed that tendency: with the agent's code in view, doubt nearly disappeared, and the difference holds when the blind sample is weighted to the design of the gold standard (annex B). A sample read blind by a person who does not see the system's answer, and a cross-check of the two readings, are a cheap way to meet that duty. AI tools can complement human review here, not replace it, as Sanchez-Graells concludes for AI against procurement corruption ([Sanchez-Graells, 2024b](#)).

No design removes this bias. People err on their own, and defer when a system answers first, as studies of automation and of annotation assisted by language models find ([Parasuraman and Manzey, 2010](#); [Schroeder, Roy and Kabbara, 2025](#); [Beck et al., 2025](#)). Oversight of AI that reads procurement is a combined system with bias on both sides. This report does not offer a solution. It shows, on a real case, where the difficulties arise when one tries to oversee such a tool, and practices that bring them into view: an agent proposes and a person confirms; a sample is read blind; the two readings are crossed and the cases of conflict studied. They are recommended practices, not guarantees, and they say more about how to oversee than the AI Act does. Better data, such as the label of R1, would give a firmer floor to start from; it would not remove the bias of the people and systems that read them.

What the AI Act assumes, and the registers cannot give

The AI Act regulates AI systems; it does not say how anyone is to find them. Four of its assumptions fail on the data of this report.

- **That AI is a category one can count.** The Act's definition (Article 3(1)) decides whether a system is AI, not whether a purchase is. To sort purchases, this report had to add its own rules for compute, courses, staffing and governance (annex A). A count of public AI made with the Act's definition alone rests on choices no one has declared.
- **That deployers know what they deploy.** Public bodies must assess, oversee and register their high-risk AI, but nothing in the Act or in procurement law makes a purchase say that it contains AI. In TED, keywords may miss up to about half of the AI bought (table 4.2), and AI built into products is not described at all.
- **That human oversight works because it is required.** The Act asks providers to design high-risk systems so that people can oversee them and stay aware of automation bias (Article 14), and deployers to assign that oversight to competent people (Article 26). It does not say how oversight is to be done or checked. This report tested it where it should work best, with a reviewer who wrote the codebook and knew the risk: once a firm answer was in view, doubt fell from 22% of purchases to about 1% (annex B). Oversight required in general terms, with no design for how it is done, is another *empty shell*, the term Sanchez-Graells (2026) uses for the model clauses.
- **That governance can be contracted.** The model contractual clauses are voluntary, and governance bought as a contract of its own appears in about 1 in 25 AI purchases, an estimate that rests on seven purchases (chapter 5).

The first two gaps can be closed at the source, by the label of R1, while the EU Public Procurement Act is still being negotiated. The other two need oversight that is designed and measured, as this report does for its own use of AI, and as R4 asks of every count.

Recommendations

The recommendations are the author's proposals, each with who should act and by when. R1 and R2 make AI visible, and R3 and R4 work best once it is. R3 makes the governance of that AI easy to buy. R4 makes every count of public AI checkable.

R1. Procurement notices should carry a mandatory AI label. The label should say what role AI has in the purchase (the object bought, a component of another product, infrastructure, literacy or governance) and what it does (diagnosis, triage, text processing, prediction); for high-risk systems, it should also give the AI Act category (the Annex III area, or the regulated product). A yes-or-no flag is not enough. In practice, this means a set of values in eForms and, in a future revision of the CPV, entries for AI. The label is filled in at three points: the buyer sets it in the contract notice, the supplier declares in its offer what AI the product contains, and the award notice records it.

Suppliers can do this: they know what their products contain, and providers of high-risk systems already document them under the Act. But a supplier has a reason to under-declare, so the label is checked, not trusted, in two ways. First, a notice that carries no label, but that a screening model ranks as probable AI, becomes a lead for an auditor, such as a procurement supervisor or a court of auditors; the screening model can be a classifier on the meaning of notices, as in chapter 5. Second, for high-risk systems, the label can be checked against the EU database of Article 71: the database records high-risk systems in use, while the label records every purchase that contains AI, when it is bought. The legal vehicle exists: the eForms are laid down by a Commission implementing regulation ([Implementing Regulation \(EU\) 2019/1780](#)), which can add the fields without new legislation. But a field can be left empty: the proposed EU Public Procurement Act, now before the co-legislators, is the place to make the label an obligation for every buyer and every supplier; it already lets the Commission extend the information that notices must carry by delegated act (recital 60). *Who*: the European Parliament and the Council (the EU Public Procurement Act), the European Commission and the Publications Office (eForms, CPV), and national platforms below the EU thresholds. *When*: the obligation in the EU Public Procurement Act during its negotiation; the eForms fields in force before 2 December 2027; entries for AI at the next revision of the CPV.

R2. Registers should record what was bought, for how much, and where it is registered.

First, what was bought: the notice should describe the purchase in enough words to be read, since registers are now read by machines first, from the title and the structured fields. This includes each category of a framework agreement or dynamic purchasing system, whose umbrella titles often say nothing of what will be bought. Second, for how much: each order should record what it costs, not the ceiling it draws on, in a field checked against the ceiling of its framework. Third, where it is registered: for high-risk systems in Annex III, the notice should name the system's registration in the EU database of Article 71, where public deployers must register their use of a system before using it (Articles 26(8) and 49(3)). *Who*: the Publications Office and national registers and platforms (ANAC, DECP, PLACSP). *When*: the next eForms release, and no later than 2 December 2027.

R3. Governance should be easy to buy. Every contract for high-risk AI should budget for its governance (impact assessment, audit and training), as a lot or a clause of the same contract, using as a minimum the EU model contractual clauses for AI procurement ([Community of Practice on Public Procurement of AI, 2025](#)). Those clauses are a floor, not a guarantee: they are voluntary, and Sanchez-Graells (2026) proposes mandatory clauses instead. AI framework agreements should include a governance lot, so that every buyer who calls off an AI system under a framework can call off its assessment under the same framework; in the UK, just over a third of AI contracts since 2018 were call-offs ([Tussell, 2026a](#)). In Spain, by number, most AI and most governance is bought in small contracts, by bodies without legal or technical staff. Central purchasing bodies should therefore also

offer governance as call-off orders that a municipality or a hospital can use without a tender of its own. Practical guidance for such buyers exists ([Sklar, 2025](#)). When a buyer cannot tell from the offer whether a system is AI, it should ask the supplier to state it in writing before the award. *Who*: contracting authorities buying high-risk AI, starting with health, and national and regional central purchasing bodies. *When*: in every such tender from now; governance lots at the next renewal of each AI framework.

R4. Every measurement of public AI should publish its validation: the definition, the procedure, the hand-checked sample, and the precision and recall. It costs little: a random sample of 100 keyword hits, checked by hand, states their precision within about ten percentage points. Recall also needs a random sample of records without AI words, as in table 4.2. Where AI does the reading, part of that sample should be read blind, by a person who does not see the system's codes (annex B). *Who*: the European AI Office, the OECD, national statistical offices, procurement supervisors and commercial trackers. *When*: the next edition of each measurement, and before 2 December 2027 for official figures.

8. Robustness and limitations

Every figure has a direction in which it may be wrong (table 8.1), and every result that depends on a few cases is given with and without them. The list after the table sets out each limit.

Figure	What it is	Which way it may err
Keyword counts	Records that name AI	Too high by the false hits, too low by what they miss; in TED the missed AI may outweigh the false hits; the interval includes the keyword count (chapter 4)
Estimated AI purchases (tables 4.1, 4.3)	Range from firm to counting doubtful as AI, with a sampling interval	Ranges too narrow: the review with the agent's code in view settled doubts that a blind reader leaves open (annex B; up to 6 more points of keyword precision)

Figure	What it is	Which way it may err
Recall (tables 4.2, A.2)	Share of the AI in the text that keywords, or review to a given depth, find	Slightly high: the lowest closeness band, read at 340 records per source, was read by the agent and checked by the author, not read blind; AI described in no recognisable way stays out
AI built into products (table A.6)	Sensitivity bound on product categories	An upper bound, not a count
"Not AI" codes	What the notice shows	Low as a measure of AI in products: the notice may not describe it
Counts of procedures	Grouping by buyer and text	High: the grouping leaves apart some records of one procedure (table A.4)
Governance, 1 in 100	Found by search	A floor
Governance, about 4%	Random sample, weighted (1.4-7.8%)	Wide: seven purchases, IT sets of TED and Spain only
Values	As recorded	High where framework ceilings are recorded as orders; missing in most TED and German records

Table 8.1. The direction of error of each kind of figure.

- **Framework agreements.** Their amounts are multi-year ceilings. One EUR 75 M French framework is 83% of the value of governance bought; without it, governance falls from 2.9% to 0.5% of the value of AI purchases. That is why the count, not the value, is the headline figure (chapter 5).
- **Values.** In Italy, three records carry EUR 2,299 M of the EUR 2,568 M recorded for contracts that name AI. Each records the ceiling of a framework or lot instead of the cost of its own order; this is a recording practice of a few buyers, not of the register (annex A). Without them, the total is EUR 269 M.
- **Small bases.** With 40 governance purchases, sector comparisons describe; they do not test (annex A).
- **The base of chapter 5.** Governance shares are shares among purchases that a classifier marks as AI, and part of that base is not AI (chapter 5).

- **Short series.** 2026 is included up to October, so no growth rate is computed.
- **Coverage of the search.** All records of Germany, the Netherlands and France, and Spain's tenders, are searched by meaning; in TED and Italy, only the IT and health sets are. AI bought elsewhere in TED and Italy that avoids AI words therefore stays out.
- **Reach.** Small purchases are fully covered only in Spain and Italy; France is covered from EUR 40,000, Germany and the Netherlands as published nationally, and TED only above the EU thresholds (table 3.2).
- **Languages.** The model is multilingual, but its accuracy may vary by language; the gold standard is stratified by source, which only partly separates languages.
- **One human reviewer.** The author reviewed every row of the gold standard, and agreement with the agent is measured blind (annex B); a second human reviewer is still needed to measure agreement between people, and to cross the two tests independently.
- **The review agent.** It is a proprietary model that no one can rerun once it is retired ([Ollion et al., 2024](#)), so the open embedding model and the human gold standard are the anchors that remain. Under a changed prompt or a smaller model, it changes 4% to 7% of its codes, almost all between doubtful and firm (annex B).
- **The lowest closeness band.** The gold standard found no AI in the lowest closeness band of any source (0 of 40 records each), which could not exclude a rate of up to about 7% in a band that holds three quarters of each IT set. A further 300 records per source were therefore drawn at random: the review agent read all 1,800, and the author reviewed the 30 it did not code as not AI. Five are AI, 0 to 2 per source, and because the band is large they lower recall, most in TED (table 4.2). This check was read with the agent's code in view, not blind; the 1,770 records the agent coded as not AI were not reread, and the blind check shows that the agent also misses AI (annex B), so the AI found there is a floor. Its size was set by the rule of three; a power calculation can set it in advance ([Card et al., 2020](#)).
- **What is not published cannot be seen, and a purchase is not a use:** no figure says whether the AI was deployed, worked or was useful.
- **Dates.** Data run to October 2026 (Italy, to December 2025; the Netherlands, to September 2026).

Author and competing interests

Pablo Ramos-Sánchez is an independent researcher. Arquestra is the author's personal website and blog; it has no commercial interest and is not a registered trademark. Its tools, the AI Procurement Observatory, the audit tool mentioned in annex A and a search of open tenders, are free and public. The design for overseeing AI proposed in chapter 7 is the one these tools use; it is offered as one possible design and published in full, so that

others can test it. No funding was received for this study. Use of AI: the author designed the study, chose the sources, wrote the codebook that defines AI, reviewed every purchase of the gold standard and wrote the report. As the method requires, a review agent (Claude, claude-opus-5-5) applied the codebook to the samples (annex A); AI assistants also helped with code, literature search and editing. The author is responsible for the content. Questions and data requests: pabloramos1395@gmail.com.

References

- Ada Lovelace Institute (2024). [Buying AI: Is the public sector equipped to procure technology in the public interest?](#)
- Alice Labs (2026). [Public sector AI procurement in Sweden.](#)
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. and Zrnic, T. (2023). [Prediction-powered inference](#). Science 382(6671).
- Atkinson and O'Bryan (2026). [Government AI Use as a Monitoring Primitive](#). arXiv:2607.04543.
- Barrie, C., Palaiologou, E. and Törnberg, P. (2024). [Prompt stability scoring for text annotation with large language models](#). arXiv:2407.02039.
- Beck, J., Eckman, S., Kern, C. and Kreuter, F. (2025). [Bias in the loop: how humans evaluate AI-generated suggestions](#). arXiv:2509.08514.
- BidSkim: White, S. (2026). [Artificial intelligence in UK public procurement, 2022 to 2026](#). BidSkim Research.
- Brookings: Denford, Dawson and Desouza (2026). [Where does federal AI spending stand in 2026?](#)
- Callaghan, M. W. and Müller-Hansen, F. (2020). [Statistical stopping criteria for automated screening in systematic reviews](#). Systematic Reviews 9, 273.
- Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K. and Jurafsky, D. (2020). [With little power comes great responsibility](#). EMNLP.
- Chen, J. et al. (2024). BGE M3-Embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. [arXiv:2402.03216](#) (model BAAI/bge-m3).
- Cochran, W. G. (1977). [Sampling Techniques](#), 3rd edition. Wiley.
- Community of Practice on Public Procurement of AI (2025). [Updated EU AI model contractual clauses](#) (MCC-AI-High-Risk and MCC-AI-Light).
- CSIS: Atalan, Tiersten-Nyman and Jensen (2026). [Understanding the U.S. Federal Government's AI Spending](#).

- Dawid, A. P. and Skene, A. M. (1979). [Maximum likelihood estimation of observer error-rates using the EM algorithm](#). Journal of the Royal Statistical Society, Series C 28(1), 20-28.
- Desai, Card and Jacobs (2026). [Validating LLMs in social science: Epistemic threats and emerging norms](#). arXiv:2607.07915.
- DIGIWHIST and Opentender (2018). [Digital platform for transparency of public tenders](#). Hertie School.
- Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E. and Stewart, B. M. (2022). [How to make causal inferences using texts](#). Science Advances 8(42), eabg2652.
- Egami, N., Hinck, M., Stewart, B. M. and Wei, H. (2023). [Design-based supervised learning](#). NeurIPS.
- European Commission (2025). [Guidelines on the definition of an artificial intelligence system under Regulation \(EU\) 2024/1689](#).
- European Commission (2026). [Proposal for a Regulation of the European Parliament and of the Council on public contracts and concessions \(Public Procurement Act\)](#). COM(2026) 590 final, 2026/0265(COD), 9 September 2026.
- European Commission (n.d.). [Single Market Scoreboard: public procurement](#) and [Public Procurement Data Space](#).
- Fazekas, M. and Czibik, Á. (2021). [Measuring regional quality of government: the public spending quality index based on government contracting data](#). Regional Studies 55(8), 1459-1472.
- Fazekas and Kocsis (2020). [Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data](#). British Journal of Political Science 50(1), 155-164.
- Fazekas, M., Tóth, B., Abdou, A. et al. (2026). [Public procurement cartels: a large-sample testing of screens using machine learning](#). International Journal of Industrial Organization.
- Fazekas, Tóth, Alshaibani and Abdou (2024). [GTI Global Public Procurement Dataset](#).
- Grimmer, J., Roberts, M. E. and Stewart, B. M. (2022). [Text as Data: A New Framework for Machine Learning and the Social Sciences](#). Princeton University Press.
- Grossman, M. R. and Cormack, G. V. (2011). [Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review](#). Richmond Journal of Law and Technology 17(3).
- Gutiérrez David and Quintana Cortés (2023). [Public procurement of AI for the EU healthcare systems](#). ERDAL 4(1), 87-139.
- Gwet, K. L. (2008). [Computing inter-rater reliability and its variance in the presence of high agreement](#). British Journal of Mathematical and Statistical Psychology 61(1), 29-48.

- Halterman, A. and Keith, K. A. (2025). [Codebook LLMs: evaluating LLMs as measurement tools for political science concepts](#). Political Analysis (2025); arXiv:2407.10747.
- Hayes, A. F. and Krippendorff, K. (2007). [Answering the call for a standard reliability measure for coding data](#). Communication Methods and Measures 1(1), 77-89.
- ISO (1999). [ISO 2859-1: Sampling procedures for inspection by attributes, part 1](#).
- King, G., Lam, P. and Roberts, M. E. (2017). [Computer-assisted keyword and document set discovery from unstructured text](#). American Journal of Political Science 61(4), 971-988.
- KPMG Law (2021). [The dusty CPV jungle](#).
- Lepers, Krieger, Pellens and Prüfer (2026). [Identifying Circular Public Procurement with LLMs](#). ZEW Discussion Paper 26-019.
- Lewis, D. D. and Gale, W. A. (1994). [A sequential algorithm for training text classifiers](#). SIGIR.
- licitaciones.io, via Europa Press (2026). [La Administración pública española dispara su inversión en IA](#).
- Lombard, M., Snyder-Duch, J. and Bracken, C. C. (2002). [Content analysis in mass communication: assessment and reporting of intercoder reliability](#). Human Communication Research 28(4), 587-604.
- Madras, D., Pitassi, T. and Zemel, R. (2018). [Predict responsibly: improving fairness and accuracy by learning to defer](#). NeurIPS.
- McKinney and Tarnay (2024). [Using AI for green procurement](#). Open Contracting Partnership.
- OECD (2024). Governing with Artificial Intelligence: Are governments ready? OECD Artificial Intelligence Papers, No. 20. [doi:10.1787/26324bc2-en](#)
- OECD (2026). [Digital Government Outlook 2026](#), ch. 4.
- OIReScon (2025). [Informe Anual de Supervisión de la contratación pública de España 2025](#), executive summary.
- Ollion, É., Shen, R., Macanovic, A. and Chatelain, A. (2024). [The dangers of using proprietary LLMs for research](#). Nature Machine Intelligence 6, 4-5.
- Open Contracting Partnership (2023). [Data Registry](#).
- Open Contracting Partnership (n.d.). [OCDS for eForms: strategic procurement codelist](#).
- Parasuraman, R. and Manzey, D. H. (2010). [Complacency and bias in human use of automation: an attentional integration](#). Human Factors 52(3), 381-410.
- [Regulation \(EU\) 2017/745](#) (Medical Devices Regulation).

- [Regulation \(EU\) 2019/1780](#) (Commission Implementing Regulation establishing the standard forms for public procurement notices, eForms).
- [Regulation \(EU\) 2019/2144](#) (General Safety Regulation for motor vehicles).
- [Regulation \(EU\) 2024/1689](#) (AI Act).
- [Regulation \(EU\) 2025/2152](#) (Commission Delegated Regulation on the EU procurement thresholds from 1 January 2026).
- [Regulation \(EU\) 2026/1744](#) (Digital Omnibus on AI), OJ 24 July 2026.
- Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (ed.), *The SMART Retrieval System*, 313-323. Prentice-Hall.
- Rolshoven et al. (2025). [Identifying Sustainability in Public Tendering](#). ECAI-2025 Green-Aware AI Workshop, CEUR-WS 4165.
- Samoili, S. et al. (2020). [AI Watch. Defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence](#). Joint Research Centre, EUR 30117.
- Shamir, A. (2026). [Public procurement and AI governance: developing a strategic framework for middle powers](#). RAND Europe.
- Sanchez-Graells, A. (2024a). [Digital Technologies and Public Procurement: Gatekeeping and Experimentation in Digital Public Governance](#). Oxford University Press. doi:10.1093/oso/9780198866770.001.0001.
- Sanchez-Graells, A. (2024b). [Procurement corruption and artificial intelligence: between the potential of enabling data architectures and the constraints of due process requirements](#). In S. Williams and J. Tillipman (eds), *Routledge Handbook of Public Procurement Corruption*, 29-41. Routledge.
- Sanchez-Graells, A. (2026). [AI's absence from the proposed EU Public Procurement Act: suggested improvements](#). How to Crack a Nut (blog), 5 October 2026.
- Schroeder, H., Roy, D. and Kabbara, J. (2025). [Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks](#). Findings of the Association for Computational Linguistics: ACL 2025, 25771-25795.
- Sklar, K. (2025). [Buying AI: Tools and Tips for Public Procurement](#). Open Contracting Partnership.
- Sklar, K. (2026). [Is your government ready to buy AI?](#) Open Contracting Partnership.
- Sombetzki, P. (2026). [Statement to the Bundestag Digital Committee on the German AI Act implementation bill](#). AlgorithmWatch, 23 March 2026.
- Tussell (2026a). [UK Public Sector AI Procurement Tracker](#).
- Tussell (2026b). [Public sector AI market analysis 2026](#).
- Webber, W. (2013). [Approximate recall confidence intervals](#). *ACM Transactions on Information Systems* 31(1).

- World Bank (2021). [ProACT](#); Cocciolo, Samaddar and Fazekas, [Government Analytics Using Procurement Data](#).
- World Bank (2026). [AI Procurement Contracts](#), WDR 2026 dataset; and [reproducibility package](#), DOI 10.60572/6n9m-dk62.
- World Bank: Lundy et al. (2026). [How Is Government Using Artificial Intelligence?](#)
- Zaidan, O., Eisner, J. and Piatko, C. (2007). [Using "annotator rationales" to improve machine learning for text categorization](#). NAACL-HLT.

Annex A. Method in detail

This annex gives the detail of the four steps of chapter 3: first the sources and the sets searched, then step 1 (find), step 2 (classify), step 3 (count) and step 4 (compare). Three further analyses follow, which are not part of the four steps: AI outside IT without AI words (table A.6), governance by sector (table A.7), and the observatory's estimated classifications. Annex B gives the checks on the classification.

Sources, thresholds and the sets searched by meaning

The sources are used under their licences: TED, free reuse with attribution; PLACSP, open data of the Spanish Ministry of Finance; ANAC, CC BY-SA 4.0; DECP, Licence Ouverte 2.0; the German and Dutch registers through the Open Contracting Partnership's Data Registry in the Open Contracting Data Standard (OCDS), CC0 1.0. The EU thresholds for supplies and services were EUR 143,000 for central government and EUR 221,000 for other buyers in 2024-2025, and EUR 140,000 and EUR 216,000 from 2026 ([Regulation \(EU\) 2025/2152](#)).

Keywords scan every record of each national source and, in TED, the IT and health notices. The search by meaning covers two sets, IT in every source and health in Spain, TED and Italy (table A.1). It also covers the whole registers of Germany, the Netherlands and France, and Spain's tenders, which are sampled for table A.6. Every record that names AI enters the IT set, wherever its CPV code puts it.

Set	What enters	Why	Size
Information technology	CPV divisions 48 (software) and 72 (IT services), plus any record that names AI	The denominator of "AI within IT" (table 4.3)	Germany 25,871; France 25,845; the Netherlands 3,840; TED 209,139; Italy 136,959; Spain 60,223

Set	What enters	Why	Size
Health	Purchases by health buyers or with health CPV codes	The sector whose AI is most often high-risk (chapter 6)	Spain 137,665 procedures; TED 341,954 notices; Italy 937,104 contracts

Table A.1. Where meaning is searched, and why. Units differ: Spain's health set counts procedures (its 271,571 health records correspond to 137,665 procedures), TED's notices and Italy's contracts.

Step 1. Find: meaning instead of words

Each record is turned into a point in a space of meaning with BAAI/bge-m3, multilingual and with open weights ([Chen et al., 2024](#)), from its title and the start of its description. Two texts that mean something similar end up close, whatever words they use: "automatic clinical note scribe" lands next to "speech recognition for medical reports". The same text always gives the same vector, up to numerical precision: vectors computed on a laptop graphics card and on a cloud service agree with cosine similarity of at least 0.99 on the sample checked. Embedding ran on one laptop graphics card (an NVIDIA GeForce RTX 4060, 8 GB): Italy's 937,104 health contracts took just over two hours.

AI is described with published **seed sentences**: descriptions of AI ("software that automatically detects lesions in images") and of ordinary IT ("maintenance of the hospital information system"). Each record's score is a **margin**: its similarity to the closest AI seed minus its similarity to the closest ordinary-IT seed. Each record confirmed by hand as AI becomes a seed for the next round, and each confirmed as not AI strengthens the other side, classic relevance feedback (Rocchio, 1971). Three defences keep it from drifting: only hand-checked records become seeds; each round keeps a random sample of what it did not retrieve; and the loop stops when a round adds no new confirmed positives. No gold-standard case is ever used as a seed: if a case teaches, it cannot test. The seeds, the score cut-off and the model version are frozen and published with the review sheets.

Among records without AI words, the top 1% by meaning holds AI 6 to 35 times more often than records taken at random, yet only 1% to 13% of that top 1% is AI (table 4.2), because AI that a reader can recognise without AI words is rare. Almost all of it is an AI system, not a course or a governance service, which nearly always name AI. Ranking by meaning therefore says what to review, not what to count. Recall at each depth of review is in table A.2.

Source	Keywords only	Plus top 1% by meaning	Plus top 5%	Plus top 25%
TED (EU)	55% (3,627)	58% (5,682)	63% (13,902)	79% (55,005)

Source	Keywords only	Plus top 1% by meaning	Plus top 5%	Plus top 25%
Spain	82% (3,870)	84% (4,433)	87% (6,687)	100% (17,958)
Italy	78% (3,908)	79% (5,238)	89% (10,560)	89% (37,170)
France	96% (888)	97% (1,137)	100% (2,135)	100% (7,127)
Germany	85% (1,680)	88% (1,921)	94% (2,889)	94% (7,727)
Netherlands	66% (138)	67% (175)	72% (323)	89% (1,063)

Table A.2. Recall by depth of review (firm codes), with the records that would need reviewing in brackets. 100% means the sample found no AI further down. The lowest closeness band was read at 340 records per source: the review agent read all of them and the author reviewed every one it did not code as not AI. The AI found there keeps the last column below 100% in TED, Italy, Germany and the Netherlands.

Step 2. Classify: the codebook

The AI Act first. An AI system is "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Article 3(1)), read with the Commission's guidelines, which leave out basic data processing, classical heuristics and simple prediction systems. A purchase that fails this test is not AI, whatever its words. The Act says nothing about a GPU server, a course or an audit; for those the codebook adds its own rules, published with every boundary case and its reason.

The rules. A purchase is an AI system when what is bought contains a model learnt from data that produces outputs for new cases inside a product or service; a commercial tool with AI inside counts. A statistical model fitted once for a study is not AI unless the contract names machine learning: inference on new cases is the line. Automation without a learnt model is not AI. A service whose supplier may use AI to deliver it is not an AI purchase unless the contract asks for AI. Every AI purchase is given a role: the object bought, a component of another product, infrastructure, literacy or governance. The boundaries:

- **Courses.** A course whose content is AI law, ethics, standards or compliance is governance; a general AI course is literacy, recorded apart, because the Act requires AI literacy (Article 4).

- *Compute.* GPU servers, AI supercomputers and AI gigafactories are not AI and are reported apart as infrastructure: hardware infers nothing on its own. A cloud platform bought to host AI applications is infrastructure unless the notice says it provides models; where it cannot tell, it is doubtful.
- *Consulting.* Advice on how to adopt or deploy AI is governance, because deciding how AI will be used is governing it. Purchases by which a public authority governs AI in general (support to apply the Act, an AI observatory, a standard for testing AI) are governance too, shown apart.
- *Staffing.* Hiring staff for any role is not AI, even an AI-specific one such as a machine-learning engineer: what is bought is people.
- *Topic.* AI that is only the topic of a purchase does not count; an event about AI counts as literacy only when it trains people, and its expenses (a speaker's hotel, catering) never count.

A boundary that a codebook does not fix is decided differently by every reader: two review agents with the same draft codebook once split on GPU servers bought for AI research, one coding them as AI and the other as compute. That is why each boundary is written down, and why R1 asks for a label that fixes it at the source.

Doubt has a direction. Some texts cannot be decided: "a forecasting dashboard, predictive model" may be a learnt model or a simple statistical rule. Doubtful cases are recorded with the way they lean, and two uses are kept apart. To measure, doubt is not forced either way: counts are given as a range. To govern, a precautionary rule leans doubt towards AI where three things meet: the notice carries no label, its text describes a function the Act's definition covers (detection, recognition, matching, prediction, assistance), and the field is one where an error is costly, such as health, employment, education, critical infrastructure or essential public services. Calling an AI system "not AI" removes it from every check designed for it, while the opposite error costs a question to the supplier. The rule only moves such cases into doubt; it never makes a doubt count as AI in the firm estimate, and it raises only the upper end of each range.

The cascade. The review allocates human time by uncertainty ([Lewis and Gale, 1994](#); [Madras, Pitassi and Zemel, 2018](#)): keyword hits and the records closest in meaning form the shortlist; each candidate is compared with its nearest reviewed neighbours, and clear non-AI drops out; the review agent (Claude, model identifier claude-opus-5-5, October 2026) reads each remaining candidate in full and codes it under the codebook, with a reason; it groups the doubtful cases by reason and proposes an answer for each group; and the author reviews every code and decides what is left. The agent never sets the seeds, the cut-off or the codebook. Every decision in the review sheets carries who took it, when and on what reason.

Error patterns found in review, each with its correction:

- AI as topic, not object: the role code.
- AI named only by technique, such as natural language processing: technique terms in the word lists and seeds.
- Research text that sits near AI in meaning: research projects as ordinary seeds.
- Near-identical notices from one buyer crowding the shortlist: identical texts grouped before review, the shortlist capped by buyer.
- Statistics and automation without a learnt model: the codebook.
- Several notices per procedure: count by procedure where it matters.
- A language missing from the word list, such as Hungarian and Romanian: word lists checked for every language.
- "Automatic" that is not AI, such as automatic doors: ordinary seeds for automation.

Step 3. Count: values, and the unit of count

Amount fields. Amounts are missing in a pattern, not at random, and the pattern differs by register (table A.3). Values in chapter 4 use only amounts of EUR 1 or more and are given with and without the largest records.

Source	Records that name AI	Amount recorded	No amount	Token amount (under EUR 1)	Recorded value	Without records of EUR 50 M or more	Median value
TED (EU)	3,627	40%	59%	1%	EUR 143.5 bn	EUR 6.5 bn	EUR 1.8 M
Spain	3,870	98%	2%	1%	EUR 891 M	EUR 466 M	EUR 5,918
Italy	3,908	100%	0%	0%	EUR 2,568 M	EUR 269 M	EUR 2,500
France	888	50%	0%	50%	EUR 1,055 M	EUR 640 M	EUR 144,000
Germany	1,680	14%	85%	1%	EUR 1,669 M	EUR 1,355 M	EUR 529,332
Netherlands	138	36%	62%	1%	EUR 36 M	EUR 36 M	EUR 349,734

Table A.3. The amount field and the recorded value of the records that name AI, using only amounts of EUR 1 or more. In TED, 96% of the value sits in 123 notices of EUR 50 million or more, mostly the estimated value of large IT frameworks that mention AI among many other things. In France, 441 of the 442 records without a usable amount carry a token EUR 0.1 or 0.2, almost all of them admissions to one dynamic purchasing system (chapter 4).

Ceilings recorded as orders. In Italy, three records carry EUR 2,299 M of the EUR 2,568 M recorded for contracts that name AI. The largest, EUR 1,605 M for services for a virtual assistant, is an order under a national framework that reports the framework, not the order: its lot field names the framework and repeats its amount, while the same framework holds 264 orders from 120 buyers with ordinary amounts, from EUR 3,000 to EUR 7.5 million. The other two are an order for project management, which carries the EUR 400 M of a "digital transformation" framework, and an order of books, among them one on artificial intelligence, recorded at EUR 293.9 M. It is the recording practice of a few buyers, not of the register. Before any value is published, framework ceilings are set apart, impossible amounts removed, and totals given with and without the largest records.

Records and procedures. Records are grouped into procedures when they share buyer and text, after normalising case, accents and punctuation. Notices, awards, corrections, lots and suppliers admitted to one framework or dynamic purchasing system count as one procedure; yearly renewals, separate orders and separate building trades do not. The grouping was checked on two samples of the IT records that name AI, drawn with a fixed seed: 40 groups of two or more records, to see whether what it merges is one procedure, and 33 pairs with the same buyer and almost the same text (at least 80% of words shared) that it leaves apart, to see what it misses (table A.4). A 34th pair was left out: its two records came from different Greek buyers whose names the normalisation, which keeps only Latin letters, had reduced to nothing. Across TED's records that name AI, this affects one group of four records.

Counting once per procedure can change a rate. In France, the dynamic purchasing system of chapter 4 weighs on the sample: 62 of the 114 French rows read (in the gold standard and the blind check) are in it, and 19% of those are AI, against 83% of the other rows; counted once per procedure (each row read weighted by one over the records of its procedure), the French rate is 81% instead of 48%. Germany shows the same on a smaller scale: one district's training courses, repeated by lot in 125 records, move Germany's keyword rate, on the rows read, from 51% counted by record to 56% counted once per procedure. The repetition was first flagged by running the report's data through the author's free audit tool ([Arquestra Audit](#)).

Register	Groups merged: one procedure	Pairs left apart: one procedure
TED	15 of 15	8 of 9
Spain	none merged: each record carries its file number	1 of 5
Italy	1 of 5 (1 unclear)	0 of 5

Register	Groups merged: one procedure	Pairs left apart: one procedure
France	10 of 10	5 of 5
Germany	8 of 8	1 of 5
Netherlands	2 of 2	4 of 4
All	36 of 40 (90%)	19 of 33 (58%)

Table A.4. Checking the grouping into procedures. It rarely merges what is separate, except in Italy, where small orders for one event or one funding line share a title; it often leaves apart what is one procedure, so counts of procedures are upper bounds. In France, where the per-procedure rate differs most from the per-record one, every merge checked is right. Unit: groups and pairs, not records.

Step 4. Compare: the common cut

Table A.5 applies the common cut: it cuts every IT set at EUR 140,000, the lower EU threshold for central government over the period (see the thresholds by year in the first section of this annex), and keeps only records with a usable amount at or above it. The estimate uses the gold standard restricted to the sampled purchases above the cut (estimation for a subgroup, or domain: each band's share of AI above the cut, times its size), with 95% intervals from resampling within bands. A cut at EUR 216,000, the EU threshold for other buyers, moves every estimate within its interval, but in Italy it rests on one or two sampled purchases, and in Spain on three or four.

IT set from EUR 140,000	Records with a usable amount	Of which above the cut	Naming AI	AI, estimated (firm / counting doubtful as AI) [95% interval]	Sampled AI above the cut
TED (EU)	90,088 of 209,139	84,489	1.6%	1.1% [0.7-1.6]	25 to 28 of 280
Spain	59,478 of 60,223	13,123	2.4%	2.0% to 2.5% [0.0-5.7]	3 to 5 of 129
Italy	136,121 of 136,959	11,504	1.1%	0.9% to 4.8% [0.0-13.3]	2 to 3 of 48
France	23,224 of 25,845	12,640	1.8%	1.0% to 1.2% [0.5-1.8]	14 to 16 of 323
Germany	4,022 of 25,871	3,288	5.7%	4.5% to 6.1% [1.9-10.8]	10 to 11 of 82
Netherlands	1,932 of 3,840	1,472	2.1%	1.0% to 1.1% [0.5-1.7]	11 to 12 of 218

Table A.5. The IT sets cut at EUR 140,000. Records without a usable amount are left out of both the count and the denominator, which matters most for Germany and TED (table A.3).

AI outside IT, without AI words

Purchases outside the IT divisions of the CPV that name no AI were sampled by closeness band in four registers, separately from the gold standard (table A.6). The table is a sensitivity test of the definition, not a count: its upper bounds count product categories that may carry AI, almost all new vehicles, which since July 2024 must carry intelligent speed assistance and, for cars and vans, emergency braking ([Regulation \(EU\) 2019/2144](#)). If such products counted, the AI that German public buyers buy without naming it could outnumber the AI they name many times over; in France, whose register starts at EUR 40,000 and holds few vehicles, it could at most match it.

Outside IT, no AI words	Records	AI, firm	Upper bound, counting probable embedded AI (mostly vehicles)	Records that name AI in IT, for scale
Spain (tenders)	450,461	none found	up to about 10,500	3,870
France (from EUR 40,000)	776,511	none found	up to about 840	888
Germany	585,441	about 590 [0-1,460]	up to about 24,800	1,680
Netherlands	45,418	about 8	up to about 2,800	138

Table A.6. AI bought outside IT without being named. 300 reviewed purchases per register (301 in Spain) in four closeness bands. Germany's firm estimate rests on very few sampled purchases, as its interval from 0 shows. Not a ranking of countries.

Governance by sector: the exact tests

Sector	Share of purchases marked as AI	Governance purchases (of 40)
Research and universities	19.3%	2 (5.0%)
Cities and regions	14.0%	7 (17.5%)
Education and training	13.0%	4 (10.0%)
Health and care	8.4%	2 (5.0%)
Central government and digital	8.2%	11 (27.5%)

Sector	Share of purchases marked as AI	Governance purchases (of 40)
Transport and mobility	5.0%	0
Justice, law and police	3.4%	2 (5.0%)

Table A.7. Who buys AI and who buys its governance: the seven largest named sectors (of the eleven tested below) of the 3,836 purchases marked as AI and the 40 governance purchases of chapter 5 (figure 5). The sector of AI purchases is given by a classifier; that of governance purchases was checked by hand.

Central government buys 11 of the 40 governance purchases (27.5%), against 8.2% of AI purchases; research buys 2 of 40, against 19.3%; health buys 2, against 3.4 expected from its share. Against each sector's share of AI purchases (two-sided binomial exact test, eleven sectors holding at least 3% of AI purchases), central government ($p < 0.001$) and research ($p = 0.02$) differ at the 5% level; with eleven sectors, chance alone would put one below that level about four times in ten, and only central government survives a correction for multiple comparisons. This rests on the sector assignments: those of AI purchases come from a classifier, those of governance purchases were checked by hand; with the classifier's assignments on both sides, central government has 7 of 40 ($p = 0.04$), which would not survive the correction.

The base is small. Only research, cities and regions, and education are expected to hold five or more of the 40, and about 250 governance purchases would be needed to detect a sector buying governance at half its share three times in four. One reading, still untested: bodies that use AI as a tool for their work may be less likely to see it as something to govern.

The observatory's estimated classifications

The Arquestra observatory gives every purchase that was not reviewed an estimated classification (likely AI, uncertain, less likely AI), from a logistic regression on the meaning vectors trained on reviewed purchases; it is a different model from the classifier of chapter 5. Checked by five-fold cross-validation on the 595 gold-standard purchases the observatory holds (those that name AI or were found by meaning, so not a random sample of all purchases), "likely AI" was AI in 97 of 100 cases and "less likely AI" was AI in 17 of 100, with 45% uncertain. Keeping each procedure's records in the same fold changes little (97 and 16), and leaving each register out in turn, which tests transfer to an unseen register, gives 98 and 21, with 55% uncertain. Estimated classifications are shown apart from reviewed codes, with their probability.

Annex B. Checking the classification

This annex gives the checks on step 2: how the gold standard was reviewed, a blind check of the readers, the agreement measures, and the stability of the review agent. In short, agreement read blind is moderate, well below agreement in the review as it was run. Crossing the two tests shows what each adds: the blind reading keeps doubt open, and, on the cases where the two split, rereading under the codebook favoured the review with the agent's code in view, a rereading subject to the same bias it examines.

The gold standard, as reviewed

Every one of the 2,317 purchases in the gold standard (one row each, called rows below) was coded by the review agent, which grouped the cases by reason and proposed an answer for each group; the author then reviewed every row, the doubtful ones one by one and the firm ones by group, with the agent's proposal in view. The author changed 18 of the 2,188 rows the agent had coded firmly (0.8%), all towards AI: 10 to AI and 8 to doubtful under the precautionary rule (annex A). Of the agent's 129 doubtful rows, the author decided 106 and kept 23 as doubtful. With the 8 moved to doubtful above, 31 stay doubtful: 26 that the author judged to lean towards AI, 4 away from it and 1 he could not tell. Agreement on this review was Cohen's kappa 0.97 on the rows the agent coded firmly and 0.84 on all 2,317 (94.6%). These figures measure the review as it was run, a person checking an agent's proposals, on the codes before the recode described next, and they are the figures to compare with the blind check below. After that review, two boundaries of the codebook were fixed in writing (compute and staffing are not AI), which recoded 72 rows (3%), almost all from AI to not AI: 69 by the written rule and 3 decided by the author under it; every estimate in this report uses the final codes.

A blind, sequential check

Two judgements made one after the other are not independent: the 0.97 measures how far the reviewer followed the agent as much as how often both were right. Reading every row again blind is not the answer either, because human attention is the scarce resource: at about 140 notices an hour, re-reading the gold standard would take some 17 hours. A blind check of a sample, designed for that constraint, completes the review.

The design combines four practices (table B.1). Purchases are split into topic groups formed without any verdict and sampled at random within each group with known probabilities, as in stratified survey sampling ([Cochran, 1977](#)). Review effort goes where the readers disagree, following double acceptance sampling from quality control ([ISO 2859-1](#)). A random sample of what the method discards estimates what it misses, as in the elusion tests of technology-assisted review ([Grossman and Cormack, 2011](#); [Webber, 2013](#); [Callaghan and Müller-Hansen, 2020](#)); the closeness bands without AI words in table 4.2

play that role. And automatic codes are corrected with the human sample, which gives unbiased estimates even when the agent errs ([Egami et al., 2023](#); [Angelopoulos et al., 2023](#)).

Step	Question it answers	Practice it comes from
Topic groups formed without the agent's verdicts	How to review without reading everything, and without anchoring?	Stratified sampling
Random sample per group, known probabilities	How to generalise from the sample to the whole?	Stratified sampling (Horvitz-Thompson weights)
More sample where the readers disagree	Where to spend human attention?	Neyman allocation, double acceptance sampling
Random sample of what the method discards	How much AI does it miss?	Elusion test (technology-assisted review)
Automatic codes corrected with the human sample	How far to trust the final figures?	Design-based supervised learning, prediction-powered inference

Table B.1. The validation design and where each step comes from.

The check uses 186 purchases that no earlier review had seen, drawn with the design of the gold standard (15 with AI words and 18 without per source, by closeness band; fewer in the Netherlands, where the closest band had no unseen purchases left), translated into English by an agent that saw no codes, and then grouped by topic. An independent agent coded all of them under the codebook; the author coded a sample without seeing those codes, seeing only the translated notice, the buyer and the CPV code, never the closeness band or the score from the ranking by meaning.

The acceptance rule was fixed before any agreement was seen: in each group, 5 purchases at random; no firm disagreement accepts the group; one adds 5 more, accepted if the 10 hold at most one; two or more send the group to tightened review, in blocks of 5, until disagreement falls to 10% or the group is exhausted. Doubts, recorded with their lean, never add to the sample; that choice was made after 15 rows, before any agreement was computed. The first part was read group by group, the rest in an order mixed at random across groups.

Two groups were accepted at the first sample of 5 (IT services, and computing hardware); every other group was small enough to read whole or went to tightened review (table B.2). The author read 136 of the 186. Because hard groups were read in full and easy ones only in part, the blind figures describe the harder cases more than the population drawn.

Topic group	Read by the author	Outcome of the sequential rule	Firm disagreements	Doubts
Software and licences	59 of 59	Tightened, read whole	7	12
Data, research and analytics services	20 of 20	Tightened, read whole	2	10
Communication, media and events	12 of 12	Tightened, read whole	3	1
Medical and laboratory equipment	7 of 7	Tightened, read whole	3	1
Training and education	16 of 16	Tightened, read whole	2	0
IT services and consulting	5 of 41	Accepted after 5	0	2
Computing hardware and data centres	5 of 19	Accepted after 5	0	1
Vehicles and transport	3 of 3	Small group, read whole	0	2
Office, facilities and construction	5 of 5	Small group, read whole	0	2
Security, surveillance and cameras	2 of 2	Small group, read whole	0	1
Other	2 of 2	Small group, read whole	0	1

Table B.2. Where the readers split in the blind check. Doubts include seven umbrella categories the text cannot settle. Each row describes a group, not a rate to generalise.

The hard groups are not the obvious ones. Hardware, where AI is often assumed, passed at the first sample; software licences and data services, where AI hides in umbrella categories, renewals and brand names, needed the whole group read. The readers split least where a notice names a product and most where it names a field.

Three outcomes

Each purchase falls into one of three outcomes. **Agreement:** the text is enough. **Firm disagreement** (AI against not AI): one reader erred, and the cause is recorded. **Doubt:** at least one reader cannot decide, the limit of the text. Of the 136 purchases the author read, 86 are agreements (63%), 17 firm disagreements (13%) and 33 doubts (24%), once seven firm disagreements on umbrella categories (generic lot titles, such as "Data and AI needs", that do not say what will be bought) are counted as doubts. Doubts vary by source, from 8% in Italy to 42% in France (36% in TED), on about 23 purchases per source, and short texts were no more doubtful than long ones. The author also judged, during the review, that several of his "not AI" codes should have been doubts, so doubts here are a lower bound. In the gold standard, reviewed with the agent's code in view, 31 of 2,317 rows stayed doubtful (1.3%). Part of that gap could be the sample, which reads the hard groups in full, but it holds when the blind sample is weighted to the design of the gold standard by source and closeness band: 22% of the author's blind codes are doubts, against 1.2% of the gold standard's codes in the same cells (topic groups cannot be matched, since the gold standard has none). The rest is the review itself: a reader shown a firm proposal settles cases that they would otherwise leave open, as experiments with annotation assisted by language models find ([Schroeder, Roy and Kabbara, 2025] (<https://doi.org/10.18653/v1/2025.findings-acl.1323>); [Beck et al., 2025] (<https://arxiv.org/abs/2509.08514>)). The ranges of chapter 4 are therefore narrower than a blind reading would give; table 4.2 says by how much for keyword precision.

Cause	Cases	Who said AI	What the notices buy	Reading under the codebook
The notice cannot settle it	7	Author	Six categories of a dynamic purchasing system for "services and support relating to Data and AI needs"; a federal "AI platform" offering cloud services for AI applications	Doubt: the limit of the text

Cause	Cases	Who said AI	What the notices buy	Reading under the codebook
A word pulls the reader	4	Author	A multilingual video-conferencing system with human interpretation; "Research Intelligence" bibliometrics software; a document-management consultation that mentions AI; business intelligence dashboards	Not AI: the word is not the object
AI around the purchase	3	Author	Electrodes, laboratory consumables and an NMR spectrometer for research projects that use machine learning	Not AI: it serves an AI project but contains none
AI only as a topic	2	Author	A hotel night for a speaker at an AI seminar; a public outreach event with a seminar on AI	Not AI
A codebook category misread	5	Author	ITIL coaching, a course catalogue, digital teaching aids, IT lots that include training	Not AI: the category meant training in AI, read as any training
AI named in technical terms	3	Agent	A Neural Designer licence, automatic transcription for subtitling, a contract to build a machine-learning model	AI: the agent read terms the author passed over

Table B.3. Firm disagreements in the blind check, by cause: 24 of 136 purchases. The first row, the seven umbrella categories, is counted as doubt in the text, which leaves 17 firm disagreements. Each cause was assigned by rereading the case under the codebook; this adjudication is not independent of the two readers.

Crossing the two tests

For every disagreement, the agent set out both codes and the reason each reader gave, and the author reread the case under the codebook (table B.3). In all 17 firm disagreements the agent's code, which is the gold standard's, stood: in 14 the author's

blind code had over-included, pulled by words such as "intelligence" or "multilingual", by AI as the topic of an event, by material bought for an AI research project, or by a codebook category misread as "any training"; in 3 he had missed AI named in technical terms. On the cases where the two tests split, rereading under the codebook therefore favoured the review with the agent's code in view. Because both readers took part in this adjudication, it shows where and why they split rather than proving who was right: the combined review is itself subject to the bias it measures. The direction it suggests, the author including too much, is also what the latent-class model below finds, with the caveat given there. Every firm disagreement and every doubt is also given a reason from a fixed list (a renewal that does not say of what, a brand name alone, a generic "intelligent" system, hardware only, a text too short, AI only as a topic or in the buyer's name, an umbrella category, a product likely to carry embedded AI), as annotator rationales do ([Zaidan, Eisner and Piatko, 2007](#)), so that a later reader can return to every case with the same reason. These reasons describe the notice; the causes in table B.3 describe why the readers split.

Agreement measures

Measure	Sample	Value
Author against agent, with the agent's proposal in view, rows the agent coded firmly	2,188 rows of the gold standard	kappa 0.97
Same, all rows	2,317	kappa 0.84 (94.6% agreement)
Blind, firm answers	110 of the 136 read by both	78% agreement; kappa 0.57; Krippendorff's alpha 0.56 [0.40-0.71]; Gwet's AC1 0.57
Blind, firm answers without the seven umbrella cases	103	83% agreement; kappa 0.67
Blind, with doubt as a third category	136	66% agreement; kappa 0.45; alpha 0.42; AC1 0.52
Third reader (a smaller model of the same family) against the agent	Firm answers among the 136	alpha 0.88
Third reader against the author	Firm answers among the 136	alpha 0.45
Same role given, when both blind readers code AI	39	31 (79%)

Table B.4. Agreement measures, as content-analysis guidance asks ([Lombard, Snyder-Duch and Bracken, 2002](#)). Alpha and AC1 are two further measures that are robust to one category dominating ([Hayes and Krippendorff, 2007](#); [Gwet, 2008](#)); here they give the same picture, because the two firm categories are close to balanced.

The drop from 0.97 to 0.57 is an upper bound on the effect of reading with a verdict in view, not a measure of it, since the two figures come from different samples and procedures. The two readers also differ in temper: the author coded 62 purchases as AI, 51 as not and 23 as doubtful; the agent 43, 86 and 7. With three readers (the third a smaller model of the same family as the agent, table B.4), a latent-class model ([Dawid and Skene, 1979](#)) reads the disagreement as the author including too much, coding about a quarter of what is not AI as AI. But it assumes that readers err independently, and two models of one family do not; a second person, or an open model of another family, is what would settle it. For the figures, the cross-check and the model agree: the blind reading adds AI that the codebook does not support, so the gold standard keeps its codes, and what the blind check changes is the width of the ranges, not their centre (table 4.2).

Stability of the review agent

The review agent was tested for stability, the weakness most often raised against coding with language models ([Barrie, Palaiologou and Törnberg, 2024](#); [Halterman and Keith, 2025](#)). On 120 purchases drawn from the gold standard (10 keyword hits and 10 records without AI words per source), five independent runs coded each purchase without seeing any earlier code: the codebook as published, twice; a paraphrase with the categories in another order; a minimal version with only the AI Act's definition; and the published codebook with a smaller model of the same family. The two runs of the same prompt gave identical codes. The paraphrase changed 4% of the codes (alpha 0.91 against the first run), the minimal version 7% (0.86) and the smaller model 6% (0.88, on these 120 purchases, not the blind sample of table B.4); almost every change moved a purchase between doubtful and a firm code, and the share of keyword hits coded as AI stayed between 43% and 53% (48% in the gold standard). The codebook is applied consistently, but this is not independent evidence, since the gold standard began from the same model's codes; the blind check against a person is the independent test.

From leans to probabilities

A human lean mixes two things that should be kept apart: how likely the purchase is to be AI, and how costly it would be to miss it. A model can state the first as a probability, from the notice and the base rate among similar notices already reviewed, and the precautionary rule can then be applied on top, openly, as a cut-off set lower where the risk is higher. That probability must be calibrated against the gold standard like any other code; measuring it across models is part of the follow-up study (chapter 7).