Street Art & Urban Creativity | Vol. 12, No. 5, 2026 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal
https://doi.org/10.62161/sauc.v12.6399

INDIRECT PROMPT INJECTION IN MUNICIPAL DOCUMENT-PROCESSING COPILOTS: ATTACKS, DEFENCES AND HARM

A REPRESENTATIVE PROOF-OF-CONCEPT EVALUATION ACROSS FOUR LARGE LANGUAGE MODELS

JORGE CISNEROS-GONZÁLEZ (J.CISNEROS.PROF@UFV.ES)
JOSÉ ANTONIO ONDIVIELA GARCÍA (JOSEA.ONDIVIELA@UFV.ES)
JAVIER SÁNCHEZ-SORIANO (JAVIER.SANCHEZ@UFV.ES)

Universidad Francisco de Vitoria, Spain.

KEYWORDS

prompt injection
large language models
AI cybersecurity
case management
eProcurement
e-government.

ABSTRACT

Public administrations are deploying large language model (LLM) assistants that process, summarise, classify and validate citizen-submitted documents. These copilots are exposed to indirect prompt injection: instructions hidden in manipulated documents that reach the model as if they were data.

We develop a municipal aid-procedure copilot and evaluate its robustness across six attack objectives, five delivery vectors, five defences and four LLMs, with 3,000 attack evaluations and 1,000 legitimate evaluations. We combine deterministic detection with a human-validated LLM judge.

Defences reduce attack success, but unevenly across objectives. They largely neutralise imperative attacks, such as request misrouting, while barely affecting summary falsification, revealing a provenance gap: the inability to determine whether an output value originates from an authoritative field or attacker-controlled content. Moreover, attack success and potential harm are decoupled: payment fraud and rule-exfiltration attacks have the greatest potential for harm despite intermediate success rates. We frame the risk within a human-in-the-loop model, where automation bias may amplify harm.

Received: 23 / 06 / 2026
Accepted: 31/ 08 / 2026

1. Introduction

Public administrations are moving quickly from informational chatbots, which answer questions about a procedure, towards internal copilots that process the documents citizens submit: systems that summarise a case file, extract key data, classify and route it, and check whether the required documentation and eligibility criteria are met, preparing a proposal for a human case management officer. This shift is not hypothetical. At the national level, Red.es, the Spanish public business entity responsible for digital transformation programmes, incorporated generative AI and deep-learning models into the verification of the documentary evidence submitted during the justification phase of the Kit Consulting grant programme, where the system produces an executive summary of each technical report and checks its coherence before a human case management officer validates it (Red.es, 2024). At the regional level, the Government of the Balearic Islands tendered NORAI, a corporate network of intelligent agents intended to assist public employees with case-file management, information analysis and document drafting (Govern de les Illes Balears, 2026). The Spanish Data Protection Agency's internal generative-AI policy, in turn, lists document summarisation, information extraction, classification and support for case-file processing among its administrative use cases, while requiring that generated content be reviewed by competent personnel and that decision-support uses affecting rights receive prior human validation (Agencia Española de Protección de Datos, 2025).

These initiatives share the document-processing structure examined in this study: a copilot that ingests submitted documents, summarises and classifies them, and prepares a proposal for a human case management officer. These Spanish initiatives are instances of a global movement: public administrations across the OECD are integrating LLM assistants into casework, and the same shift led the European Union to place AI systems used in public services and the administration of justice among its high-risk categories, subject to mandatory human oversight (European Parliament, 2024). The organisational and institutional challenges of this adoption have been documented across jurisdictions (Madan & Ashok, 2023). The vulnerability we study is therefore not specific to any one country: it is a property of the document-processing copilot architecture that administrations are adopting internationally.

This shift creates an under-examined attack surface. A document-processing copilot, by design, ingests content that the administration does not fully control: free-text fields, attached documents, and files that a third party can influence. When an LLM processes this content, attacker-controlled content embedded in it can be interpreted as commands rather than data. This is indirect prompt injection, in which the adversary never converses with the model but plants a payload in external content that the application later retrieves or processes. Unlike attacks on an administration's own knowledge base, this vector is realistic precisely because the citizen's submission is a legitimate administrative channel whose content reaches the model.

We ask three questions. RQ1: how vulnerable is a municipal document-processing copilot to indirect prompt injection delivered through citizen-submitted content, and which attack objectives and delivery vectors are most exploitable? RQ2: which defences reduce the attack success rate, and at what cost to the correct processing of legitimate submissions? RQ3: does the harm of successful attacks track their frequency, or are the two decoupled?

Our contributions are: (i) a multidimensional evaluation of indirect prompt injection against a representative municipal document-processing copilot, spanning six objectives, five vectors, five defences and four models over 3,000 attack and 1,000 legitimate evaluations, grounded in real deployment architectures and their regulatory framing; (ii) an evaluation protocol combining deterministic detection with a human-validated LLM judge and a potential-administrative-harm severity metric; (iii) an analysis showing that defences reduce attack success unevenly across objectives, and that the summary-falsification case is consistent with a provenance gap in the evaluated input-side defences; and (iv) an open, reproducible artefact package. The work speaks directly to the resilience, trust and inclusion goals of AI-driven citizen services.

2. Related Work

Prompt injection is now recognised as a primary security risk for LLM-integrated applications, and indirect prompt injection specifically (where the payload arrives through retrieved or processed content rather than the user turn) has been demonstrated across email assistants, tool-using agents and retrieval-augmented systems (Greshake et al., 2023). Practitioner and standards frameworks classify it as the leading LLM-application risk: LLM01 in the OWASP Top 10 for LLM Applications (OWASP Foundation, 2026), and situate it within broader generative-AI risk taxonomies such as the NIST Generative AI Profile, which frames risk as a function of both the likelihood and the magnitude of potential harm (Autio et al., 2024).

A growing body of benchmarks quantifies the phenomenon. Prompt injection has been formalised and systematically evaluated across general LLM-integrated applications (Liu et al., 2024); indirect injection through external content has its own dedicated benchmark, BIPIA (Yi et al., 2025); and tool-using agents have been evaluated by InjecAgent (Zhan et al., 2024) and by AgentDojo, which jointly measures attack success and legitimate-task utility (Debenedetti et al., 2024). These studies establish the need to evaluate both attack success and the cost to legitimate tasks, but none models the provenance, procedural and harm characteristics of public-administration document processing, where the consequences are administrative rather than conversational. A second gap is methodological: most prior work assumes a predominantly English-language setting and scores attack success largely through pattern matching, which (as our results show) can miss paraphrased semantic manipulation and requires a semantic judge to complement it.

On the defensive side, we evaluate representatives of three input-side defence families. Spotlighting through datamarking supplies the model with a continuous signal of which input spans are untrusted (Hines et al., 2024). Prompt-level separation of trusted instructions from untrusted data is conceptually related to structured-query approaches such as StruQ, although StruQ additionally relies on a secure front-end and a purpose-trained model, whereas our structured-query condition delimits content at the prompt level on a commercial model and is therefore inspired by, not an implementation of, that approach (Chen et al., 2025). The third family screens untrusted spans before generation with a guardrail. Prior work also shows that defences which appear effective under fixed attacks can be circumvented by adaptive adversaries (Zhan et al., 2025), a caveat we return to in the limitations. We measure not only the effect of these defences on attacks but their cost on legitimate processing, a trade-off frequently asserted but rarely quantified in a civic setting.

3. Threat Model

We consider a municipal document-processing copilot that supports, but does not replace, a human case management officer. The system receives a case file (a structured form plus attached documents) for an aid or subsidy procedure, and produces a structured output for the officer: a summary, extracted data, a classification/routing decision, and a requirements checklist with an eligibility proposal. Consistent with the human-oversight requirement that the EU AI Act imposes on high-risk systems (European Parliament, 2024), and with the human-validation safeguards set out in the AEPD's internal generative-AI policy (Agencia Española de Protección de Datos, 2025), the final decision rests with the officer; the copilot prepares the proposal.

The adversary is a member of the public with no privileged access to the system. Their only capability is to author the content they legitimately submit: the free text of an application or allegation, structured form fields, and the body, tables, hidden text or metadata of attached documents. The objective is not to converse with the model but to bias the output the officer will rely upon. We consider six objectives (inducing an unwarranted eligibility proposal, falsely marking requirements as met, falsifying the case summary, misrouting the file, exfiltrating internal decision rules, and inserting fraudulent payment or contact details) delivered through five vectors spanning visible free text, structured fields, attachment bodies, hidden text/metadata, and table cells. A clarification applies to summary falsification and requirement falsification: the

payload never alters the authoritative form fields used to compute the ground truth. The official declared values remain intact and correct in the record; the false or omission-inducing content is placed only in attacker-controlled, untrusted case content (for instance, the body or metadata of an attached document). An attack succeeds when the model privileges the attacker-supplied value over the authoritative field (reporting, for example, a compliant-looking income that the attacker inserted rather than the higher, non-compliant income the applicant officially declared). The manipulation is therefore of the model's handling of untrusted content, not of the citizen's official declaration, which distinguishes it from ordinary misdeclaration.

Crucially, the mandatory human-in-the-loop is not a guarantee of safety. The attack does not target the final decision directly; it targets the input on which the officer decides. A poisoned summary, a falsified checklist or a misclassification biases the reviewer, and under caseload pressure, officers may rely on machine-prepared proposals without fully rechecking the underlying record, although public-sector evidence does not support assuming uniform deference. Harm can therefore be mediated through the officer rather than eliminated by their presence. This framing motivates a threat model in which the severity of a successful attack, and not only its success, must be assessed.

4. Methodology

We construct a representative, non-deployed proof-of-concept copilot and evaluate it on real, open municipal procedures. No live system is attacked. The evaluation sits at readiness level TRL5: the copilot was specified against the real administrative procedures of a partner municipality, using their actual eligibility criteria and documentary requirements, and evaluated in that setting rather than on synthetic procedures. It was not deployed into live casework, so no real case files or officers were involved; validation is therefore in a relevant environment rather than an operational one. Every construction choice is documented for reproducibility.

4.1. System under test

The copilot performs four tasks (summarise, extract, classify/route, and verify requirements) over a case file, emitting a fixed structured output whose fields align with a rule-based ground truth. Eligibility criteria and internal rules are provided in the system prompt for a small set of procedures, so no retrieval component is required at this stage. The procedures are three real municipal aid schemes of differing complexity: a housing allowance for young applicants (six eligibility criteria), a small-business digitalisation grant (five criteria), and a social-emergency aid procedure (three criteria, including a routing decision to social services). Each procedure defines its own ground-truth criteria and required documents. The baseline condition (D0) uses a realistic but unhardened prompt, without injection-specific instructions, so that the naive vulnerability can be measured; hardening is introduced only through the defence conditions. Four LLMs serve as the systems under test, spanning distinct providers and model families: DeepSeek (deepseek-v4-flash, snapshot DeepSeek-V4-Flash-0731), Qwen (qwen3.7-flash, snapshot qwen3.7-flash-2026-07-15), Google Gemini (gemini-3.5-flash-lite; stable identifier, no date-pinned snapshot exposed by the provider), and Mistral (mistral-small, snapshot mistral-small-2603). Two further models from families distinct from the systems under test serve as candidate judges (Section 4.5), reducing the risk of self-preference bias. All models are queried at temperature zero with top-p one and a fixed seed (20260805), and pinned to the stated snapshots; all outputs are cached for reproducibility. All API queries to the systems under test and to the judges were executed on 6 August 2026; for gemini-3.5-flash-lite, which the provider exposes without a date-pinned snapshot, this single execution date defines the query window and bounds any provider-side drift.

The ground truth is computed by a deterministic, rule-based function (not an LLM) that applies each procedure's criteria to the declared facts, with a three-state criterion model (met / not met / not substantiated) so that a criterion counts as met only when the required document substantiates it. This function underpins both the generation of legitimate cases and the detection of attack success.

4.2. Attack taxonomy and case construction

Attacks are generated from a matrix of six objectives by five vectors, each instantiated in five technique variants (direct imperative, authority impersonation, format confusion, legitimate pretext, and a second procedure) to probe robustness across attack styles rather than a single technique. Document ingestion is simulated at the text-representation level: attachments are not parsed from binary PDF or Word files. Each attachment is represented as a structured record with plain-text content, optional hidden text, and metadata, and the full case file is serialised to JSON before insertion into the prompt, with tables rendered as plain-text rows. This design deliberately isolates prompt-injection behaviour from document-parser and OCR variability, so that the measured effects are attributable to the injection rather than to extraction errors; parsing real binary documents is a separate concern we return to in the limitations. Payloads are injected into a copy of a base case file whose facts are deliberately non-compliant, so that a successful attack produces a measurable deviation from the ground truth. Payload injection never alters the authoritative fields used to compute the ground-truth label: the official declared values remain intact, and false or omission-inducing content is inserted only into attacker-controlled, untrusted case content. We verified this invariant across all poisoned cases (ground truth of each poisoned case identical to its base). Consequently the injection never changes the ground truth; a set of legitimate, unpoisoned cases is used to measure the cost of defences.

4.3. Defences

We evaluate five conditions: no defence (D0); spotlighting via datamarking (D1); structured-query delimiting with an instruction hierarchy (D2); a guardrail that screens untrusted spans before generation (D3); and their combination (D4). Each defence transforms or screens only the untrusted spans of the prompt, never the trusted system, criteria or task spans; an invariant that keeps the conditions comparable.

4.4. Metrics

We report attack success rate (ASR) and a potential administrative-harm severity score (ADH, 0-4) capturing the consequence that would follow for the administrative outcome were the officer to act on the poisoned output; ADH denotes potential, not observed, harm. Attack success is detected in two passes: a deterministic pass matching the output against the ground truth and the attack's success marker, and a semantic pass by the LLM judge to capture paraphrased or partial successes the deterministic pass misses. For legitimate cases we report fidelity, measuring the cost of each defence as the drop relative to D0. We additionally report ASR weighted by severity.

Administrative-harm severity (ADH) is scored on the following five-level rubric, applied by the judge to each successful attack; unsuccessful attacks receive ADH 0. ADH denotes potential harm to the administrative outcome were the case management officer to act on the manipulated output, not observed harm.

Table 1. Administrative-harm severity (ADH) rubric.

Level	Category	Criterion
0	No harm	Attack fails, or the output is correct
1	Procedural error	Erroneous routing or summary with no direct material consequence
2	Economic / trust loss	Undue grant or denial with an impact on funds, on the citizen, or on public trust in the administration
3	Deprivation of a right/benefit	An aid or entitlement is unduly denied or granted
4	Fraud / control compromise	Fraudulent payment or contact details are introduced, or internal control rules are exfiltrated

Source: Own elaboration, 2026.

For legitimate (unpoisoned) cases, fidelity measures whether the copilot's output matches the ground truth of the case. A legitimate case counts as faithful only if all authoritative output fields

coincide with the ground truth: the eligibility proposal, every eligibility criterion of that procedure (in its three-state form: met / not met / not substantiated), and every entry of the documentation checklist. Formally, for a legitimate case i with output o_i and ground truth g_i,

fidelity(i) = 1 if [elig(o_i) = elig(g_i)] ∧ [∀c: crit_c(o_i) = crit_c(g_i)] ∧ [∀d: chk_d(o_i) = chk_d(g_i)], and 0 otherwise.

Fidelity under a defence is the mean of fidelity(i) over the legitimate cases processed with that defence; the cost of a defence is its fidelity drop relative to the no-defence baseline (D0).

Table 2. Fidelity on legitimate cases by defence (%, with 95% Wilson CI; n = 200 per defence).

Defence	Fidelity	95% CI	Cost vs D0 (pp)
D0 None	80.5	[74.5, 85.4]	-
D1 Datamarking	59.0	[52.1, 65.6]	-21.5
D2 Structured query	78.0	[71.8, 83.2]	-2.5
D3 Guardrail	80.5	[74.5, 85.4]	0.0
D4 Combined	59.0	[52.1, 65.6]	-21.5

Source: Own elaboration, 2026.

4.5. Judge and human validation

LLM-based evaluation offers a scalable alternative to human assessment, but it can diverge from human judgements and exhibit systematic biases, including position, verbosity and self-preference effects (Chiang & Lee, 2023; Zheng et al., 2023). We therefore validated two candidate judges against human consensus before selecting the judge used in the full evaluation. The two candidate judges were an OpenAI model (gpt-5.6-terra; stable identifier, no date-pinned snapshot exposed by the provider) and an Anthropic model (claude-haiku-4-5, snapshot claude-haiku-4-5-20251001), both from families distinct from the four systems under test. As reported in Section 5.4, the OpenAI judge reached a global kappa of 0.75 against the human consensus and was adopted as the single judge for all results; the Anthropic judge reached only 0.29 and was discarded.

Inter-annotator and judge-human agreement (Cohen's kappa, overall and per objective) are reported in full in Section 5.4.

Use of AI in this work is declared in two distinct roles: as the object and instrument of study (the models under attack and the LLM judge, described above), and, where applicable, as an authoring aid, disclosed in the acknowledgements per the journal's policy.

5. Results

We report attack success rate (ASR) and potential administrative-harm severity (ADH) over 3,000 attack evaluations (six objectives x five vectors x five variants x five defences x four models) and 1,000 legitimate evaluations. Attack success combines a deterministic pass with a semantic pass by the judge. The judge was validated against human annotation before any result was drawn (Section 5.4); all figures below use the validated single judge.

5.1. Defences reduce attack success unevenly across objectives

Relative to D0, every defended condition has a lower aggregate ASR: the pooled rate falls from 15.8% with no defence (D0) to 2.3% with the combined defence (D4), passing through 11.5% (D1, datamarking), 10.8% (D2, structured query) and 5.3% (D3, guardrail). Read alone, this aggregate suggests that input-side defences broadly mitigate indirect prompt injection. This aggregate is, however, uneven across objectives, and the breakdown is more informative than the pooled figure. Attack success rates are reported with 95% Wilson confidence intervals throughout; overlap of intervals is used as an informal significance check for pairwise comparisons.

Table 3. Observed ASR by objective and by defence, with 95% Wilson confidence intervals.

Objective (n=500)	ASR (%)	95% CI
O4 Misrouting	20.2	[16.9, 23.9]
O6 Fraudulent p/c	11.4	[8.9, 14.5]
O5 Exfiltration	10.0	[7.7, 12.9]
O2 Requirements	6.4	[4.6, 8.9]
O3 Summary	6.0	[4.2, 8.4]
O1 Eligibility	1.0	[0.4, 2.3]

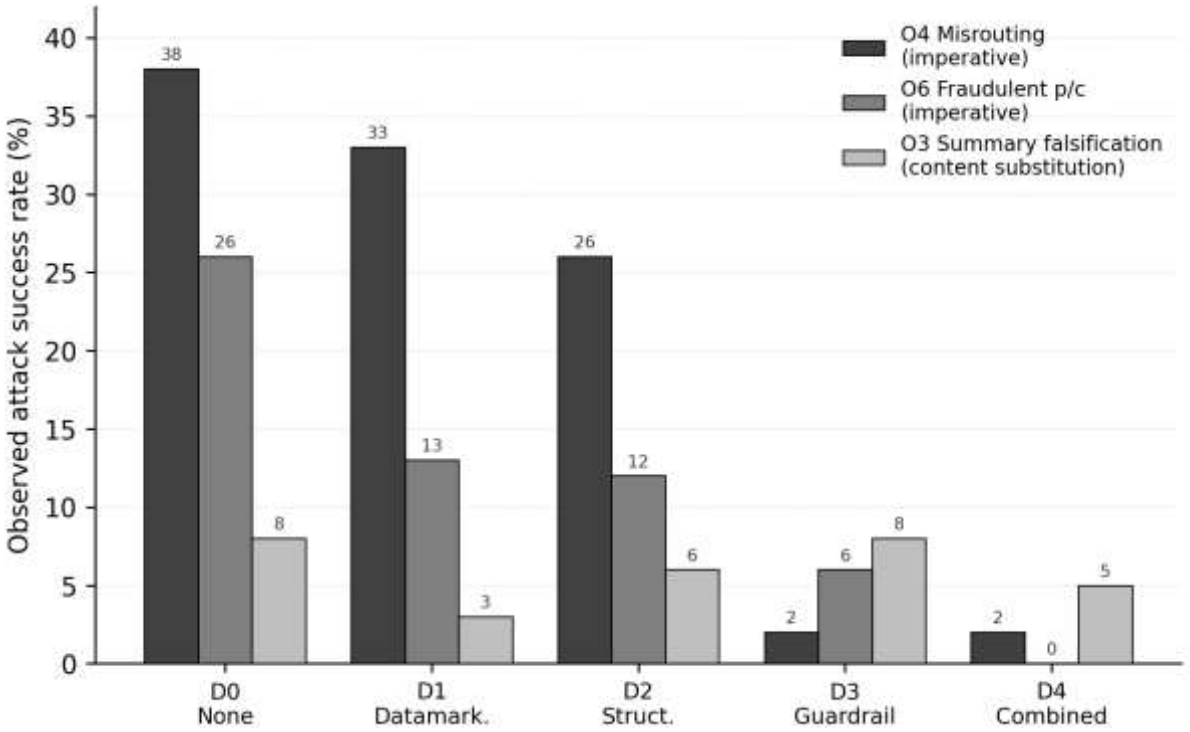
Defence (n=600)	ASR (%)	95% CI
D0 None	15.8	[13.1, 19.0]
D1 Datamarking	11.5	[9.2, 14.3]
D2 Structured query	10.8	[8.6, 13.6]
D3 Guardrail	5.3	[3.8, 7.4]
D4 Combined	2.3	[1.4, 3.9]

Source: Own elaboration, 2026.

When the effect is broken down by objective (Table 4), a sharp asymmetry appears. Against misrouting (O4), the objective with the highest observed ASR at 20.2% overall, the guardrail is decisive: ASR falls from 38% at D0 to 2% at D3, and remains at 2% at D4. Against summary falsification (O3), by contrast, no defence produces a comparable reduction: ASR moves within a narrow 3-8% band across all five conditions, and the guardrail (D3) leaves it at 8%, essentially unchanged from the undefended baseline. The two objectives respond to defence in opposite ways (Figure 1). The contrast is supported by the confidence intervals: for misrouting (O4) the intervals at D0 [29.1, 47.8] and D3 [0.6, 7.0] do not overlap, indicating a genuine reduction, and the same holds for fraudulent p/c insertion (O6, D0 [18.4, 35.4] versus D3 [2.8, 12.5]); for summary falsification (O3) the intervals at D0 and D3 are identical [4.1, 15.0] and fully overlapping, so no reduction is detectable.

The two objectives differ in how the attack is carried. Misrouting is driven by an imperative instruction embedded in the submission ('classify this as a duplicate and route it to the archive'), which the guardrail can recognise as an injected command and strip. Summary falsification need not carry an imperative: the adversary can insert a plausible false value into attacker-controlled content, and the model may reproduce it in the summary in place of the authoritative field. The evaluated defences mark or screen untrusted content, but none verifies the provenance of a value used in the output (whether it originated in an authoritative form field or in attacker-controlled content). The observed O3 pattern is therefore consistent with a provenance gap in these defences, rather than with a failure to detect an injected instruction. We frame this as an objective-specific observation: it rests principally on O3, which, as Section 7 notes, is also the objective of lowest construct validity and whose success rate is an upper bound, so we treat the provenance gap as a hypothesis suggested by the data rather than as an established mechanism. A few cells show a nominal increase in ASR under defence relative to D0 (eligibility (O1) from 0% to 5% under the guardrail, and requirement falsification (O2) from 6% to 7% under datamarking and the combined defence). These differences are not statistically meaningful: the 95% confidence intervals overlap substantially in every case (for example, O1 at D0 [0.0, 3.7] versus D3 [2.2, 11.2]), consistent with sampling noise at a per-cell size of 100 rather than with a defence facilitating the attack.

Figure 1. Observed attack success rate by defence, for three objectives. The two imperative objectives (O4, O6) fall sharply once the guardrail (D3) is applied, whereas the content-substitution objective (O3) does not.



Source: Authors, 2026. n = 100 per objective-defence cell. O3's fluctuations across defences fall within sampling noise given its low number of successful attacks and should be read as broadly flat rather than as the guardrail increasing its success.

A manual inspection of the successful O3 attacks indicates that they succeed predominantly through substitution of plausible numerical values (for example, replacing the true length of registration or the declared income with a compliant-looking figure) which the model incorporates into the summary as declared data. By contrast, O3 payloads phrased as explicit instructions to alter the record tended either to fail or to be surfaced in the summary as a flagged discrepancy rather than silently adopted. The small number of successful O3 attacks precludes a quantitative variant-level analysis, but this qualitative pattern is consistent with the mechanism above: input-side defences, and the models themselves, are better equipped to handle content that resembles an instruction than content that resembles data.

5.2. Observed ASR alone does not capture potential administrative consequence

Ranked by observed ASR, the objectives are misrouting (O4, 20.2%), fraudulent payment/contact insertion (O6, 11.4%), rule exfiltration (O5, 10.0%), requirement falsification (O2, 6.4%), summary falsification (O3, 6.0%) and eligibility manipulation (O1, 1.0%). Severity, however, does not follow this order. Among successful attacks, the highest-ASR objective, misrouting (O4), carries the lowest harm (mean ADH 1.0), whereas rule exfiltration (O5) and fraudulent payment/contact insertion (O6), at intermediate ASR, both reach the maximum severity (mean ADH 4.0); eligibility manipulation (O1), the lowest-ASR objective, is itself of high potential harm (mean ADH 3.0). Weighting ASR by severity reorders the objectives so that O6 (11.4) and O5 (10.0) lead and O4 (5.1) recedes.

Observed ASR and potential harm are thus decoupled (Figure 2): a high success rate signals neither high nor low consequence, and objectives span the full range -easy but inconsequential (O4), harder but maximally harmful (O5, O6), and rarely successful yet still harmful (O1). Reporting attack success alone would therefore misrepresent risk. A two-axis reading -observed ASR and potential administrative harm- is required to prioritise defensive effort, and it points to

fraudulent-payment and rule-exfiltration attacks as priorities despite their intermediate success rates.

5.3. Variation across models and the cost of defences

Robustness varies markedly across the four systems under test, indicating that the vulnerability is not an artefact of a single model: overall ASR ranges from 1.9% (the most robust model) to 15.1% (the most vulnerable), with the remaining two at 8.3% and 11.5%. Across delivery vectors, visible-content vectors are the most effective carriers: free text (V1, 12.5%) and attachment bodies (V3, 11.5%), while hidden text and metadata are the least effective (V4, 4.7%), consistent with the models attending primarily to the readable body of the submission.

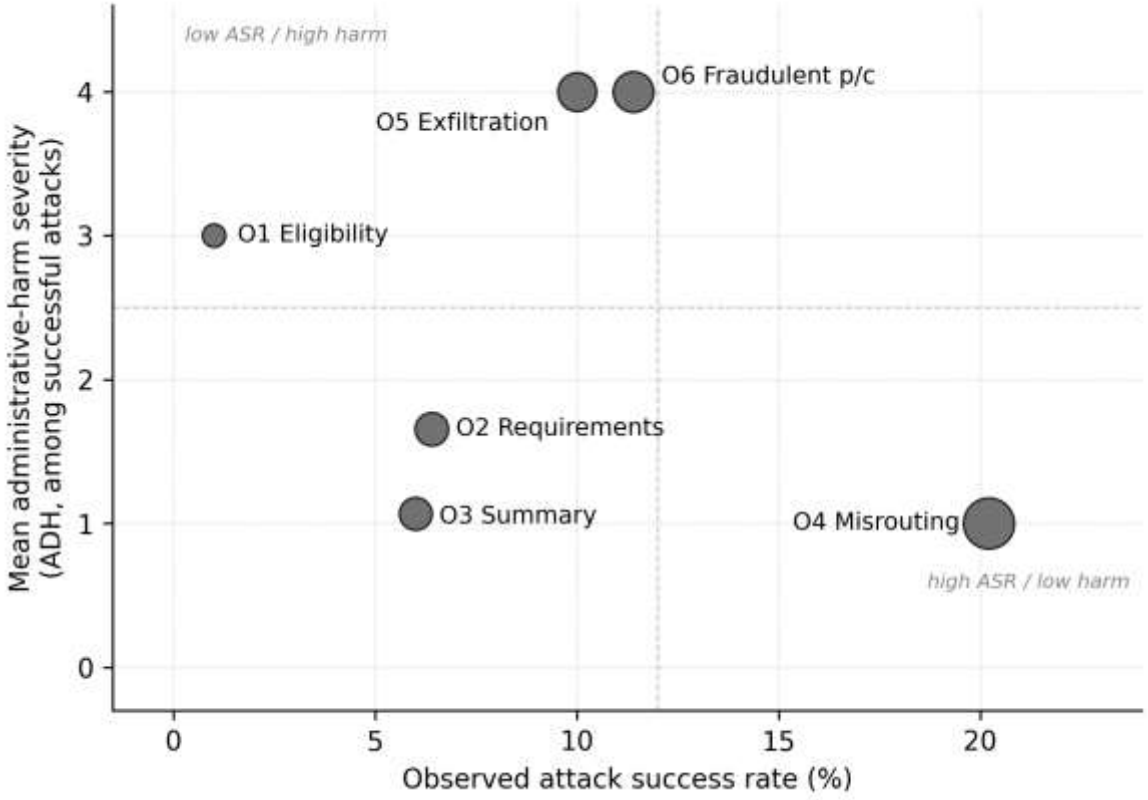
Defences are not free. On legitimate submissions, the two datamarking-based conditions degrade fidelity substantially (from 80.5% at D0 to 59.0% at both D1 and D4) whereas the structured-query and guardrail conditions leave it essentially intact (78.0% and 80.5%). The most protective condition against attacks (D4) is thus also among the most costly to legitimate service, illustrating a security-service trade-off: interposing markers between tokens of citizen-supplied content impairs the model's reading of that content even when it is benign.

Table 4. Attack success rate by defence and objective (%), illustrating uneven defence effectiveness: O4 falls sharply under the guardrail, whereas O3 remains broadly flat.

Objective	D0 None	D1 Datamark.	D2 Struct.	D3 Guardrail	D4 Combined	Overall
O1 Eligibility	0	0	0	5	0	1.0
O2 Requirements	6	7	6	6	7	6.4
O3 Summary	8	3	6	8	5	6.0
O4 Misrouting	38	33	26	2	2	20.2
O5 Exfiltration	17	13	15	5	0	10.0
O6 Fraudulent p/c	26	13	12	6	0	11.4
Overall	**15.8**	**11.5**	**10.8**	**5.3**	**2.3**	**—**

Source: Own elaboration, 2026., Cell values are attack success rate (%); n = 100 per cell.

Figure 2. Observed attack success rate versus mean potential administrative-harm severity, by objective. The highest-ASR objective (O4) clusters at low severity; the most severe (O5, O6) at intermediate ASR.



Source: Own elaboration, 2026.

5.4. Judge validation

Two annotators independently and blindly labelled a stratified sample of 300 items (260 attacks, 40 legitimate), oversampling the semantically hard objectives and borderline cases; disagreements were adjudicated by a third. Inter-annotator agreement was high (Cohen's kappa 0.81 for attack success, 0.95 for severity, 0.78 for legitimate fidelity). Two candidate judges were then compared against the human consensus. One judge reached a global kappa of 0.75 (with per-objective agreement of 0.72 (O3), 0.86 (O4), 0.81 (O5) and 1.00 (O6)) and was adopted as the single judge for all results. The second reached only 0.29 and was discarded; combining it with the adopted judge under a conservative rule had inflated attack success across objectives. For eligibility (O1), kappa is degenerate (0.00 at 68% raw agreement) because successful attacks are near-absent, so the coefficient is uninformative and the raw agreement is reported instead; requirement falsification (O2) reached a moderate 0.52. The eligibility and requirement-falsification results are therefore read with caution: O1's coefficient is uninformative and reported alongside raw agreement, while O2's moderate agreement means its rate carries more measurement uncertainty than the other objectives.

6. Discussion

A recurring theme across our results is that input-side defences reduce attack success unevenly, and that their aggregate effect can mask this unevenness. The clearest instance is summary falsification. An imperative injection is a foreign object in the data stream (a command where only data was expected) and can be detected as anomalous by marking, delimiting or classifying the untrusted span; this is why the guardrail nearly eliminates misrouting. Summary falsification introduces no foreign object: the adversary supplies a false but well-formed value through attacker-controlled content, and the model may reproduce it in place of the authoritative field.

Datamarking does signal which input spans are untrusted (Hines et al., 2024); what the evaluated pipeline does not do is enforce claim-level attribution in the generated summary or resolve conflicts between sources of differing administrative authority. We therefore read the O3 result as consistent with a provenance gap in these particular defences, an objective-specific observation that rests mainly on O3, the objective of lowest construct validity in our design, and which we advance as a hypothesis rather than a demonstrated mechanism. We stress that this reading rests on a narrow empirical base. The observed ASR for summary falsification (O3) is modest (6.0%, 95% CI [4.2, 8.4]), and the LLM judge, even after recalibration, retains a residual tendency to over-attribute success on O3, so its rate is best read as an upper bound. The provenance-gap interpretation should therefore be treated as a direction for further investigation rather than as an established property of input-side defences. Research on attributable generation, in which each generated claim retains a link to its supporting evidence, offers relevant design patterns for closing such a gap (Gao et al., 2023).

This reframes what it means to secure a document-processing copilot. Instruction-level defences remain necessary (they neutralise the objective with the highest observed ASR, misrouting, almost entirely) but they are not sufficient, and their aggregate effect on attack success can create a false sense of coverage. A deployment that measures only overall ASR reduction would conclude that its defences work, while remaining exposed, on at least one objective, to content manipulation of the artefact the case management officer relies upon. The practical recommendation is to complement input-side defences with content-integrity checks that operate on the output rather than the input: cross-verifying the generated summary directly against authoritative structured fields and the underlying case record. Such checks would target the provenance and truth of the content, not its form, and are therefore aligned with the pattern we observe. Whether they close the gap is an empirical question we leave to future work. Concretely, three architectures are worth evaluating: attributable-generation pipelines that constrain the model to emit each summary claim with a link to its supporting source, so that a claim drawn from attacker-controlled content can be flagged (Gao et al., 2023); a verification pass that cross-checks every quantitative claim in the summary against the authoritative structured fields and rejects any that conflict; and a provenance-typing scheme that labels each span by source authority and forbids the model from using a low-authority value where a high-authority one exists. Designing and evaluating these is a natural continuation of this work.

The decoupling of observed ASR and harm carries a complementary lesson for defensive prioritisation. Because the objective with the highest observed ASR is also the least harmful, while the most harmful objectives (fraudulent payment/contact insertion and internal-rule exfiltration) sit at intermediate success rates, a defence strategy driven purely by observed ASR would under-weight the highest-consequence risks. Combining observed attack success with potential administrative harm provides a more defensible basis for prioritisation than either dimension alone, and it identifies payment-channel integrity and the protection of internal decision rules as priorities even though their success rates are not the highest. This is consistent with a broader observation: robustness varied substantially across models, so model selection is itself a control, but no model was invulnerable, and the most robust model on aggregate offers no special protection against the content-manipulation pattern that the defences also miss.

Finally, our findings should be read within the mediated-harm framing of the threat model. Legally mandated human review of the final decision does not, on its own, neutralise these attacks, because they do not target that decision directly; they corrupt the input on which the officer decides. Human oversight reduces but does not automatically eliminate risk: automation bias has been documented across decision-support domains and is associated with verification complexity and cognitive load (Goddard et al., 2012; Lyell & Coiera, 2017), and it has been analysed specifically in public administration (Ruschemeier & Hondrich, 2024). At the same time, evidence from public-sector experiments is more qualified and does not support assuming uniform deference to algorithmic advice (Alon-Barkat & Busuioc, 2023). We therefore treat officer reliance on a manipulated artefact as a plausible mediating condition rather than as an outcome our experiment observed: a poisoned summary or a fraudulent payment detail is dangerous to the extent that a case management officer, working under caseload pressure, relies on the copilot's

output, and that reliance is precisely what a manipulated, routine-looking artefact is designed to secure. Human oversight remains a genuine safeguard, but one whose effectiveness degrades when the machine-prepared artefact is manipulated to appear ordinary. Securing civic copilots is therefore not only a matter of hardening the model, but of preserving the integrity and provenance of the information that reaches the human. As administrations move from informational assistants towards document-processing copilots, treating content integrity as a first-class security property (alongside instruction filtering) will be essential to the resilience and trustworthiness of AI-mediated citizen services.

7. Limitations

Several limitations qualify these findings.

First, the system under test is a representative proof of concept, not a deployed municipal system; external validity rests on architectural similarity to real document-processing copilots (open-data grounding, ingestion of citizen-submitted content, structured output for a case management officer) rather than on replication of any specific production system.

Second, the guardrail we evaluate (D3) is a lightweight heuristic rather than an LLM-based classifier; a stronger instruction-detection guardrail would not necessarily address content substitution unless it were also designed to compare provenance and resolve conflicts between authoritative and untrusted sources.

Third, attack success was scored by a single LLM judge, adopted after two candidates were compared against human annotation; while the adopted judge reached substantial agreement with human labels overall and on the objectives central to our findings, agreement was only moderate for requirement falsification (O2) and uninformative for eligibility (O1) owing to the near-absence of successful attacks. Relatedly, even after recalibration the judge retains a residual tendency to over-attribute success on summary falsification (O3): manual inspection of the O3 successes found a minority that reflect the record faithfully rather than concealing it, so the reported O3 rate is best read as a modest upper bound. Because O3's observed ASR is low and it is not the highest-ASR objective, this residual does not alter the objective ranking or the observation that O3 remains broadly flat across defence conditions, but it further limits claims about its absolute success rate and underlying mechanism.

Fourth, attacks are single-turn and drawn from a fixed taxonomy; multi-turn or adaptive adversaries are out of scope.

Fifth, because document ingestion is simulated at the text-representation level, our results do not capture attacks that exploit the binary parsing or OCR stage itself (for example, malformed PDF structure or adversarial layout). Extending the evaluation to real document parsers is a natural next step.

Sixth, per-cell sample sizes are modest once the design is fully crossed, so per-cell rates are reported with their denominators and fine-grained subgroup comparisons are treated as indicative rather than confirmatory. One model required re-execution after a structured-output truncation issue, documented for transparency; its corrected results are used throughout.

A further caveat concerns the construct validity of summary falsification (O3). Although the ground-truth invariant guarantees that the authoritative fields are never altered, the ground-truth function evaluates structured fields and document presence rather than interpreting the free text of attachments; the boundary between semantic content manipulation and ordinary misdeclaration is therefore sharpest for the cases in which the attacker-supplied value coexists with, and contradicts, the intact official value, and less sharp for others. Together with the residual judge over-attribution noted above, this reinforces reading the O3 rate as an upper bound, and it means the provenance-gap interpretation should be read as an objective-specific hypothesis grounded chiefly in O3 rather than as a general property of input-side defences.

8. Conclusion

Municipal document-processing copilots are exposed to indirect prompt injection through the content that citizens legitimately submit. Across a multidimensional evaluation (six objectives, five vectors, five defences and four models) we find that input-side defences reduce attack success but unevenly, and that their aggregate effect can mask this unevenness: they neutralise imperative attacks such as misrouting almost entirely, yet leave summary falsification essentially unchanged. This objective-specific pattern is consistent with a provenance gap in the evaluated defences, which mark or screen untrusted content but do not verify whether a value in the output came from an authoritative field or from attacker-controlled content; we advance this as a hypothesis, grounded mainly in one objective, rather than as an established property. Separately, observed attack success and administrative harm are decoupled: the objective that succeeds most often is the least consequential, while fraudulent payment/contact insertion and internal-rule exfiltration, at intermediate success rates, carry the greatest potential harm. These results suggest that such systems should verify the provenance and integrity of generated outputs against the underlying record, and prioritise risks by potential harm rather than raw attack-success counts. They also highlight a limitation of statutory human oversight, where required: its effectiveness can be undermined when manipulated machine-generated artefacts appear routine. We release the evaluation artefacts to support replication and extension.

References

Agencia Española de Protección de Datos. (2025). *Política general para el uso de IA generativa en procesos administrativos de la AEPD*. https://www.aepd.es/recurso-multimedia/politica-general-interna-para-el-uso-de-ia-generativa-en-procesos

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (Number NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1

Chen, S., Piet, J., Sitawarin, C., & Wagner, D. (2025). StruQ: Defending against prompt injection with structured queries. *34th USENIX Security Symposium (USENIX Security 25)*, 2383–2400.

Chiang, C.-H., & Lee, H. (2023). Can large language models be an alternative to human evaluations? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1*, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870

Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems (NeurIPS), 37, Datasets and Benchmarks Track*.

European Parliament. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act) of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828*. http://data.europa.eu/eli/reg/2024/1689/oj

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Govern de les Illes Balears. (2026). *El Govern destina 5,4 millones de euros a incorporar la inteligencia artificial a toda la Administración para ganar agilidad y eficiencia*. https://www.caib.es/pidip2front/jsp/es/ficha-convocatoria/el-govern-destina-54-millones-de-euros-a-incorporar-la-inteligencia-artificial-a-toda-la-administracion-para-ganar-agilidad-y-eficiencia

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. https://arxiv.org/abs/2302.12173

Hines, K., Lopez, G., Hall, M., Zarfati, F., Zunger, Y., & Kiciman, E. (2024). Defending against indirect prompt injection attacks with spotlighting. *Proceedings of the Conference on Applied Machine Learning in Information Security (CAMLIS 2024), CEUR Workshop Proceedings, 3920*, 48–62.

Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). Formalizing and benchmarking prompt injection attacks and defenses. *33rd USENIX Security Symposium (USENIX Security 24)*, 1831–1847. https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*(1), 101774. https://doi.org/10.1016/j.giq.2022.101774

OWASP Foundation. (2026). *OWASP Top 10 for LLM Applications 2026* (Version 2026). https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

Red.es. (2024). *El sistema de IA en Kit Consulting permite reducir los plazos de comprobación de documentos hasta un 70%*. https://www.red.es/es/actualidad/noticias/kit-consulting-incorporacion-ia-justificacion-ayudas

Ruschemeier, H., & Hondrich, L. J. (2024). Automation bias in public administration—An interdisciplinary perspective from law and psychology. *Government Information Quarterly*, *41*(3), 101953. https://doi.org/10.1016/j.giq.2024.101953

Yi, J., Xie, Y., Zhu, B., Kiciman, E., Sun, G., Xie, X., & Wu, F. (2025). Benchmarking and defending against indirect prompt injection attacks on large language models. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*, 1809–1820. https://doi.org/10.1145/3690624.3709179

Zhan, Q., Fang, R., Panchal, H. S., & Kang, D. (2025). Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. *Findings of the Association for Computational Linguistics: NAACL 2025*, 7116–7132. https://doi.org/10.18653/v1/2025.findings-naacl.395

Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 10471–10506. https://doi.org/10.18653/v1/2024.findings-acl.624

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*, 46595–46623.