

Characterizing Agentic Flooding of Government Services

Chris Schmitz¹, Lewis Hammond², Alan Chan³

¹Centre for Digital Governance, Hertie School
ch.schmitz@hertie-school.org

²Cooperative AI Foundation

³GovAI

Abstract

AI agents are making it easier for the public to interact with government, such as by helping them apply for benefits, understand complex policies, and make their opinions heard. Although improving service accessibility is beneficial, any resulting surges in demand could strain unprepared government services. We term such surges *agentic flooding* of government services (“flooding”) and provide three contributions. First, based on a collected dataset of 84 potential cases of flooding across 11 jurisdictions, we posit that flooding is likely occurring widely today, mostly through large language models (LLMs) generating text cheaply. Second, we evaluate what services are most exposed to flooding. We develop a risk matrix to analyze a service’s exposure, and suggest that near-term risk is highest for financially attractive, but complex services. Finally, we map possible government responses to flooding. Precedent suggests these responses will likely be sufficient to stop most cases of flooding, but the fastest to deploy – friction-inducing measures like fees – often trade off equitable access to public services. Accordingly, we close by recommending near-term actions that may allow governments to mitigate flooding without invoking this trade-off.

Dataset — <https://github.com/CLSChmitz/flooding-dataset>

Preprint. This work will appear in the proceedings of the 9th AAAI Conference on AI, Ethics, and Society (AIES), October 12–14, 2026.

1 Introduction

AI agents increasingly help the public interact with government. For example, they can already generate legally coherent text for an official complaint, or parse complex policy documents into plain language. Iscenko et al. (2026) classify about 1% of requests to Google Gemini as helping with government interaction. These capabilities reduce submission costs for citizens and make government more accessible, a desirable public value (Moore 1995).

However, such reductions in cost could also cause surges in the complexity or quantity of requests. Evidence of such surges is emerging: the Australian government has considered reintroducing fees for Freedom-of-Information

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Potential Instances of Agentic Flooding

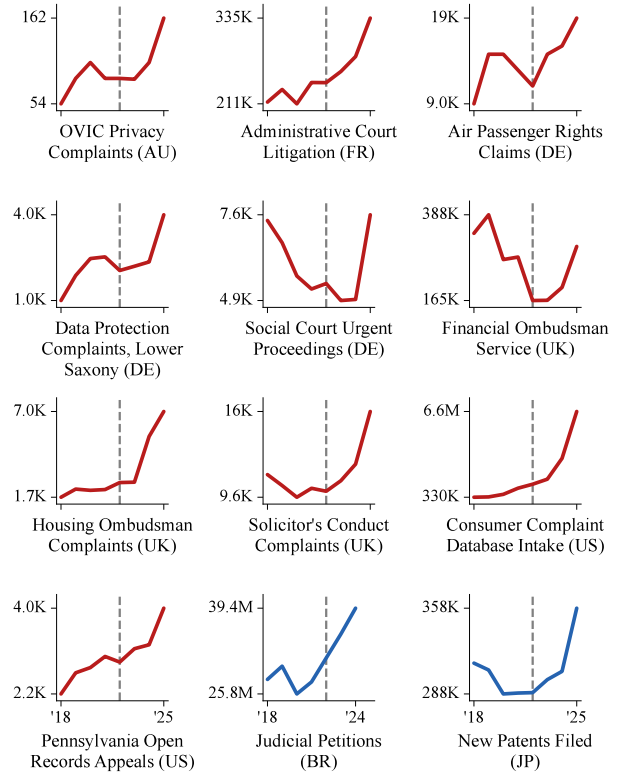


Figure 1: Annualized submission volumes (2018–2025) for 12 government services. For each, either government officials (red) or reputable secondary sources (blue) have asserted AI involvement in surges. Only some surges begin around the 2022 release of ChatGPT (dashed lines).

(FOI) requests citing a wave of AI-generated submissions (Canales 2025); some German social courts largely attribute a 55% year-on-year rise in caseload in 2025 to AI-generated claims (LTO 2026); Shah and Levy (2026) find a “dramatic increase” in both the share and average length of self-represented cases before US federal courts since 2022, presumably driven by AI-aided citizen self-representation.

Table 1: Types of “administrative burden” citizens face when interacting with government (Moynihan, Herd, and Harvey 2015), the AI agent capabilities that may reduce them, and technologies that enable these capabilities.

Burden Type	Specific Burden	Agent Capability	Enabling Technologies
Learning Costs	Understanding complex rules and procedures	Summarizing complex regulations and requirements in plain language	LLMs with large context
	Uncertainty about eligibility	Evaluating personal data against policy and entitlement rules	Context augmentation, directly or, e.g., via Model Context Protocol (MCP)
	Unawareness of services or entitlements	Scanning or continuously monitoring for entitlements; proactive alerts	Web search tools; pipelines with scheduled execution
Compliance Costs	Legal writing; filling forms	Context-specific, sophisticated writing	LLMs with user context
	Confusing government websites and service portals	Navigating and interacting with websites autonomously	Browser use tools; vision-language models
	Locating and assembling supporting documents	Retrieving, formatting, and attaching required supporting materials	File system access; multimodal document parsing
Psychological Costs	Managing response deadlines and follow-ups	Monitoring deadlines; alerting users or submitting replies autonomously	Asynchronous task queues; calendar API integration
	Having to re-explain case context to many contacts	Maintaining full case context across multi-step processes	Memory; context management
	Uncertainty about case progress and next steps	Polling for case updates and notifying the user only when action is required	Scheduled execution; access to web or API endpoints
	Repeated dehumanizing bureaucratic interactions	Handling rejections, follow-up calls, and status chasing autonomously	Tools for external communication, e.g. email and voice synthesis APIs

Key questions about such surges remain open. In particular, we need to know how widespread they are, how serious a risk they pose, and what governments can do about them. Given the rapid pace of AI capabilities advancements, there is a clear and pressing need for greater understanding.

This work provides a structured treatment of the phenomenon we term *agentic flooding* of government services (or “flooding” hereafter): surges in the volume or complexity of requests they receive, which (1) strain their capacity and (2) are enabled by agents reducing the cost of interacting with such services (§ 2). We address three questions:

Is flooding happening – and if so, in what forms? (§ 4) We compile a dataset (§ 3) of 84 recent surges in service demand, each of which officials or reputable secondary sources attribute to AI use (Figure 1). The data is consistent with agentic flooding occurring widely today, though our methodology does not permit causal or quantitative claims about AI’s impact. Almost all the cases we identify share a mechanism: LLMs generating large amounts of text, paired with humans navigating the rest of the process manually. More sophisticated use of agents, such as autonomous website navigation, is not yet evident.

How big is the risk? (§ 5) The risk of flooding depends on the progress and diffusion of AI capabilities, and varies significantly across countries and services. Analyzing our data, we posit that the near-term risk of flooding is high-

est for financially attractive services where complex submission requirements have historically suppressed demand, such as court claims and tax returns. These properties can be assessed in services now, and such assessments can be updated as new information about AI capabilities arrives. We construct a risk matrix to guide more detailed assessment of individual services.

How could governments respond? (§ 6) We map government response options, grouped under two strategies: suppressing demand (e.g. fees, rate limits, in-person requirements) and increasing capacity (e.g. staffing more people, deploying AI, or redesigning the service interface). Precedent suggests both approaches are sufficient to address flooding, but demand suppression is far faster to deploy. However, it decreases the quality and accessibility of government services, and can introduce procedural inequality.

Preparing proactively could help governments avoid this trade-off. To that end, we recommend three near-term actions (§ 7.1): auditing service vulnerability, integrating digital identity into the most exposed services, and clarifying the legality of resilience-building measures.

2 Agentic Flooding of Government Services

This section defines *agentic flooding* of government services (§ 2.1) and analyzes its potential causes (§ 2.2).

2.1 Definitions

Throughout this work, we use *AI agents* (or “agents”) to mean LLM-based systems with some degree of autonomy and tool access (Kasirzadeh and Gabriel 2025). Our definition includes current-generation consumer-facing LLMs, as they typically integrate tools such as web search (Sadeddine et al. 2025).

We use *government services* as a deliberately broad term, to mean both public services – such as welfare or infrastructure – and non-service channels of government interaction, e.g. public participation, judicial procedures, or freedom of information requests (Osborne 2018).

We use *the public* and *users* interchangeably to refer to all individuals interacting with government services.

With these terms, we can define agentic flooding.

Agentic flooding of a government service is a surge in the volume or complexity of requests it receives, which

- (i) is caused by AI agents interacting with it; and
- (ii) substantially strains the service.

We distinguish between two types of flooding: increases in the number of requests (*quantitative* flooding), and increases in the complexity of each request (*qualitative* flooding). For example, agents that help users discover eligible services may increase request volume, whereas agents that draft text may lengthen the average complaint letter. An instance of flooding could be both qualitative and quantitative. For example, LLM hallucinations could result in high volumes of complex, but incorrect, applications. Both of these effects, which Blundell, Bozio, and Laroque (2013) term the “extensive” and “intensive” margins of work, increase processing burden. However, some potential responses only address one type (§ 6). For example, introducing rate limits is unlikely to ease qualitative flooding.

2.2 Potential Causes of Flooding

AI agents could cause a surge in the volume or complexity of requests by (1) reducing the costs of interacting with the government service or (2) increasing the benefits of doing so. As such, both users and middle-man organizations, e.g. law firms, have incentives to deploy agents to interact with government services. Any resulting increases in demand could strain the service because many services are only designed to handle a limited amount of demand.

Agentic Capabilities Could Reduce Costs. AI agents could reduce *administrative burdens*: the costs of participating in bureaucratic procedure which users face when interacting with government (Halling and Baekgaard 2024). Moynihan, Herd, and Harvey (2015) group these costs into three types: learning, compliance, and psychological costs.

In Table 1, we map how agent capabilities could reduce all three costs. *Learning costs*, the cognitive effort of understanding government, are addressed most directly by agents’ ability to parse long contexts and personalize communication to users (Kasneeci et al. 2023). *Compliance costs*, the

practical effort of completing required steps, fall as agents grow more capable of high-quality digital work (Drouin et al. 2024). *Psychological costs*, the emotional toll of repeated bureaucratic contact, drop as agents handle more of that contact.

Reductions in administrative burdens often raise demand for services. For example, the introduction of an app to request municipal services in Boston led to a 33% increase in reports (O’Brien et al. 2016), and online voter registration options consistently increase enrolment numbers (Garnett 2022).

At a conceptual level, rational-actor models propose that users submit requests when the benefit they expect exceeds the cost of submission. (Nichols and Zeckhauser 1982; Zeckhauser 2019). Individual users rarely act financially optimally – their behavior is shaped by many other factors, such as psychological frictions or stigma (Currie 2004; Bhargava and Manoli 2015). But the rational-actor model often accurately describes aggregate behavior across a population (van Oorschot 1991). We would therefore expect requests to increase when AI agents decrease submission costs.

Agentic Capabilities Could Increase Benefits. Agentic capabilities could also increase the benefits of interacting with a service. For example, agents could rewrite Request for Information (RFI) responses to be more convincing, increasing the likelihood that the responses’ views are reflected in policy. AI agents could also cheaply help to ensure that all information relevant to an application is gathered, ensuring that applicants receive the maximal benefit. Analogously, tax accountants help to ensure that clients receive their maximum possible tax refund.

Incentives to Deploy Agentic Capabilities. Beyond the user-level incentives, there are strong commercial incentives to deploy agentic capabilities in ways that increase demand for government services. The business model is well-established: middle-man organizations offer to handle complex government interactions on behalf of users in exchange for a cut of the earned entitlements, growing the number of claims to government. The most prominent example is claims management companies, which have long mass-submitted benefit and compensation claims through templated submissions, such as for flight delay compensation and UK personal injury claims. Companies advertising their use of AI to similar effect already exist (DoNotPay 2026).

Further, adversarial actors may explicitly seek to flood services. For example, they may aim to destabilize critical infrastructure (Piedrahita et al. 2026), or to incapacitate agencies. Even good-faith users may be incentivized to abuse the system. Analogously, “vexatious litigants” file large numbers of meritless court claims, each of which they are entitled to file, but which collectively overwhelm the court (Rust 2024). Thus far, transaction costs have kept such cases rare enough for ad-hoc management; agent capabilities could make these cases much more frequent.

Increased Demand Can Cause Strain. Increased demand for government services is not in itself harmful, but agencies often cannot meet it. Many services run close to operational capacity because of budget constraints, operational inefficiency, or legal obligations to minimize overhead (GII 2026). Some governments know of and tolerate, or even deliberately maintain, administrative burdens which suppress take-up (Peeters and Widlak 2023).

Capacity strain can have both direct operational consequences – like growing backlogs or missed response deadlines – and broader fiscal and policy effects, such as requiring budget adjustments (§ 5). Collectively, these effects may erode user experience of public services, which strongly predicts trust in government (Van De Walle and Bouckaert 2003; Christensen and Laegreid 2005).

3 Evidence Collection

To compile a dataset of real-life cases where flooding may be occurring, we scan government services in 12 countries (§ 3.1). Because AI involvement in demand surges is difficult to prove, we only include cases where a government official or credible third party has asserted such involvement (§ 3.2). This enables a broad scan but trades off statistical representativeness. To collect cases, we employ an LLM-aided qualitative coding workflow (§ 3.3).

3.1 Scope

We scan for publicly accessible information about government services in 12 countries: Australia, Brazil, Denmark, Estonia, France, Germany, Japan, the Netherlands, Singapore, South Korea, the United Kingdom, and the United States. We choose countries where it is common to report publicly on administrative matters, aiming for variation in geography and in two core dimensions of public administration scholarship: administrative tradition (Painter and Peters 2010) and digital government maturity (OECD 2025b).

In each of these countries, we scan the same fixed list of 13 government domains. We set this list by combining public-facing lists of government services from three studied countries (UK, US, and France). Ten domains cover public services, e.g. benefits and social protection, citizenship, or jobs and pensions. Three cover government processes that are not public services, but in which flooding may also occur: participatory processes; regulatory complaints and reporting; and transparency and access to information. Appendix A describes the detailed methodology with which we set both scopes.

3.2 Case Selection and Attribution

Demand for public services is shaped by many interrelated factors including policy changes, spurious events, and broader social and economic developments; processing indicators like backlog sizes and wait times are similarly confounded (Siciliani, Borowitz, and Moran 2013). Deeper analysis may mitigate these problems in individual cases, especially where submission contents are public (Shah and Levy 2026), but it is not (yet) feasible at scale.

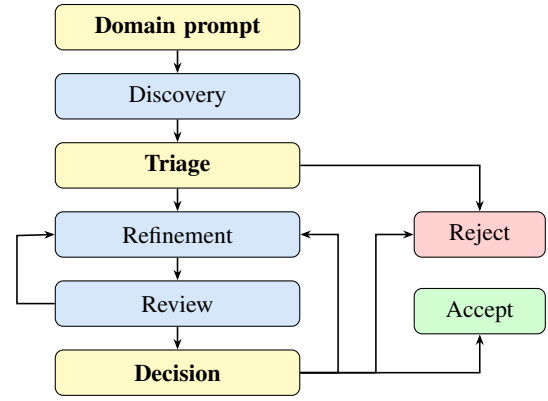


Figure 2: Overview of LLM-aided pipeline for collection of cases. Yellow: human step. Blue: LLM step.

As a way to partially address this issue, we select cases conservatively. In addition to evidence of the two criteria in our definition, we require that the affected government body itself – or a reputable, relevant source – attribute a change in demand patterns to AI use by the public. We pose three separate inclusion criteria per case:

1. **A plausible, specific mechanism** by which AI could reduce transaction costs for the service – e.g. that LLMs help write a complaint which can then be submitted online. We judge plausibility ourselves based on the design of the service.
2. **Evidence of change** in demand patterns for the service that is consistent with what the mechanism in (1) could cause, found in official statements or actions (or another high-quality source), and reflected in primary (self-collected) volume data if available. Trends in the volume data alone do not fulfill the criterion.
3. **External attribution** of that change to AI use, by either (a) a government source – e.g. a press statement, official communication, or policy update – or (b) a reputable secondary source – e.g. an established news outlet, professional legal or administrative publication, or peer-reviewed work.

Our methodology produces a broad dataset but does not afford causal, quantitative, or statistical claims. We further discuss limitations in § 8.

3.3 Methodology

Data Format. We code all cases into a standardized template, provided in the dataset repository and outlined in Appendix B. All figures for which we provide statistics are captured in a structured format in this schema. We do not perform additional analysis on free-text fields, except to highlight illustrative examples. For each case, we attempt to collect:

- Free-text descriptions of the service and its submission interface, the mechanism by which flooding may occur, and any impact and government response observed.

Table 2: Ten illustrative cases of *agentic flooding* in our dataset.

ID	Jurisdiction	Service Domain	Service Name	Flooding Type
A	Australia	FOI	Federal FOI Requests	Quantitative
B	Germany	Justice & Legal	Social Court Lawsuits	Qual. & Quant.
C	Germany	Public Participation	Environmental Impact Assessment (UVP) Public Comments	Qual. & Quant.
D	Brazil	Employment	Temporary Disability Benefit (INSS Auxílio-doença)	Qualitative
E	United Kingdom	Justice & Legal	Money Claims Online (MCOL) / Online Civil Money Claims	Qual. & Quant.
F	United Kingdom	Public Participation	Local Plan Consultations (Regulation 18 and 19)	Qual. & Quant.
G	Japan	Public Participation	Strategic Energy Plan Public Consultations	Qual. & Quant.
H	South Korea	Justice & Legal	Civil E-Litigation (Electronic Payment Orders)	Qualitative
I	Netherlands	Justice & Legal	Municipal WOZ Valuation Objections	Qual. & Quant.
J	United States	Citizenship	Voter Challenge and Voter Roll Purge Submissions	Qual. & Quant.

- Structured fields classifying the case type and mechanism along the typologies in §§ 2.1 and 2.2, whether it meets each of the three inclusion criteria, and who asserts AI involvement in the change in requests.
- Annualized volume statistics from 2018 to 2025 in a case-specific unit (e.g. requests filed, backlog size).
- Metadata, e.g. country, government body, and domain; and links to all sources used in compiling the case.

Case Collection Pipeline. We employ an LLM-aided research and qualitative coding workflow (Ye et al. 2024). As outlined in Figure 2, the pipeline has three stages. During *discovery*, a country-domain combination (e.g. “France, Health Services”) is scanned for candidate services. In *iteration*, each case cycles between LLM calls to improve case data and review case quality. In *finalization*, a human reviewer validates the case, makes required corrections, and chooses whether to include it in the dataset. Detailed prompts and scaffolding are listed in Appendix B.

Preventing Hallucinations. Our approach contains three safeguards against the risk of LLM hallucinations (Huang et al. 2025). First, we perform human verification of all cases, before both iteration and finalization. Second, we force structured responses from all LLM calls, and maintain each case in a standardized, JSON-based data format. This format enables us to verify key fields, such as the names of government agencies and numeric case counts, deterministically; it also allows metadata, such as iteration count and review comments, to be tracked and passed to each LLM call. Third, we extract accessed web URLs deterministically from the metadata of LLM API responses, rather than requiring the models themselves to generate them as part of the response. This guarantees all cited sources are real, human-verifiable websites, which we visit to confirm key figures.

4 Is Agentic Flooding Happening?

The collected data is consistent with flooding occurring today, across a wide range of government domains and jurisdictions. Almost all cases occur through one mechanism: LLM-generated text submitted by users who navigate the rest of the process manually.

Dataset Overview. We document 84 cases across 11 jurisdictions and 13 service domains. Full case data are available in the linked dataset repository. Table 2 contains ten illustrative cases to which the subsequent analysis refers.

These 84 cases are the product of a broad scan followed by strict inclusion criteria: of the 2288 candidate services named during discovery – roughly 190 per country – fewer than one in twenty met all three criteria in § 3.2. The biggest constraint on inclusion was requiring official or third-party attribution of the surge to AI.

Of the 84 cases, government officials assert AI involvement (with or without third-party corroboration) in 58 (69%); solely third-party sources assert it in the remaining 26 (31%). Public and judicial services are more represented than non-service interaction channels, such as participation procedures and transparency requests (73% and 27% of cases, respectively). The most frequently represented domains are Justice & Legal Services (23%, N=19), followed by Regulatory Complaints (12%, N=10) and Benefits & Social Protection (11%, N=9).

Strength of Evidence. The dataset provides evidence that AI use is increasing the volume and complexity of requests for at least some government services. The cases span jurisdictions, service types, and reporting outlets; in most cases, the affected government body itself asserts AI involvement. Any alternative explanation would have to account for independent misattribution across all of these criteria.

However, our data-gathering remains exploratory and diagnostic, and affords no causal or quantitative claims about either (a) the prevalence of flooding or (b) the role of agents. Demand is driven by many factors, and some volume series in Figure 1 begin rising before the release of ChatGPT. Moreover, because we require explicit public attribution, the dataset likely undercounts the cases that meet our definition of flooding (§ 2.1). § 8 discusses these limitations.

Capabilities Enabling Flooding. In most of our cases (87%), flooding occurs when LLMs help to cheaply generate legally sophisticated text. Most of the services in the dataset accept text, often of arbitrary length, e.g. FOI requests [Case A], planning consultations [Case F], or public comments [Case C].

There are two potential explanations for this pattern. First, text generation is more mature and accessible than agentic capabilities such as browser navigation (Zhou et al. 2024). Second, sampling bias: cases largely enter our dataset because officials or domain experts noticed an anomalous change in the pattern of submissions (69%, N=58). LLM-generated text can be identified by the content of the submission itself, whereas many other types of agent-aided submissions may look identical to human ones if the submission format is constrained – e.g. by short fields or structured online forms. Any effect such submissions have on request patterns could only be observed via the quantity of requests, which is also affected by various other factors (Siciliani, Borowitz, and Moran 2013). It is plausible that officials are thus far attributing such quantitative changes to AI less frequently.

Types of Flooding Observed. We code 60% (N=50) of cases as quantitative flooding and 90% (N=76) as qualitative, with 50% (N=42) exhibiting both types. These numbers are largely consistent with the above pattern of LLM text generation, which both allows *more* users to submit requests to free-text service interfaces – e.g. the less legally literate submitting litigation [Case H] – and makes these requests *longer*, as in [Case B], where some letters spanned over 4000 pages.

We observe cases of both adversarial and non-adversarial flooding in the dataset. While most cases show ordinary users acting in good faith, some involve individuals acting in bad faith for personal gain, e.g. attempting to get disability benefits with AI-generated medical certificates [Case D]. A few involve organized adversarial campaigns, notably mass-submitting FOI requests or voter roll challenges [Case J].

5 How Big Is the Risk?

The risks posed by flooding depend on how quickly AI capabilities progress and diffuse, and vary with the service and jurisdiction affected. A key question is therefore which services are most at risk. Analyzing our dataset, we posit that near-term risk is highest for financially lucrative services with complex applications and legal processing obligations, such as tax returns (§ 5.1). We propose a risk matrix (§ 5.2) to more rigorously evaluate the likelihood and severity of a specific service being flooded.

5.1 Which Services Are Most Vulnerable?

The services in our dataset share several commonalities. Most obviously, they overwhelmingly accept free-form text through open digital channels, and the agent capability required to flood them – LLM text generation – is freely accessible (§ 4).

The observed impacts of flooding so far are moderate. While we find evidence of overworked public servants, calls for additional budget, and calls to reform legal response obligations [Case E], the strain is modest compared to precedent, e.g. from mass comment campaigns (Balla et al. 2022). Where governments respond, e.g. by re-introducing FOI fees [Case A] and batch-dismissing consultation responses [Cases C, G], these responses also appear effective.

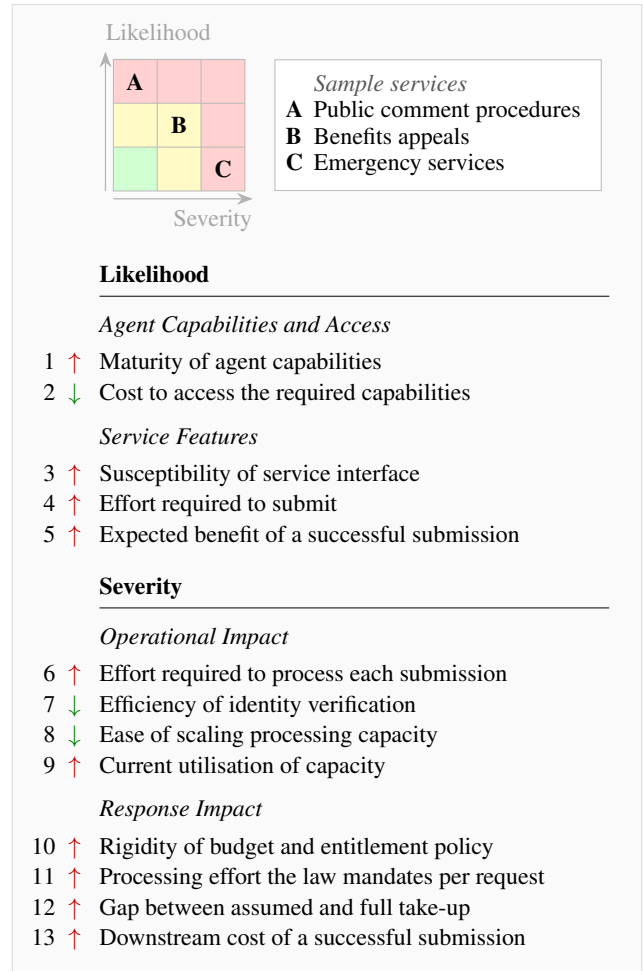


Figure 3: Risk matrix of factors which may predict the likelihood and severity of *agentic flooding* for a given government service, with three illustrative services mapped. Arrows indicate whether a higher value raises (↑) or lowers (↓) risk.

There are some higher-severity cases in our dataset, e.g. objections to tax valuations [Case I], social court lawsuits [Case B], and civil claims [Cases E, H]. These share two properties. First, each successful submission is individually consequential: it is financially valuable to the claimant if successful, or triggers legally mandated processing effort. Second, each service’s resilience has historically been owed to friction rather than design. With few structural limits on submission length or volume, the complexity of submitting – and the legal or professional knowledge it demanded – historically gated who submitted, and how well.

Given our inclusion criteria (§ 3), these patterns may also reflect which cases officials notice and report. Nonetheless the risk of flooding appears highest for services with these two properties: services that are financially attractive, and whose demand has so far been suppressed by friction or tacit knowledge. These could include tax administration and court systems. Notably, an initial assessment of services against these two properties does not require forecasting AI

progress. They can be analyzed today, as we discuss in § 7.1.

5.2 Risk Matrix for Flooding

Figure 3 describes a risk matrix (Kaplan and Garrick 1981) for the flooding of a given service, containing factors that may predict the *likelihood* that a service is flooded and the *severity* if it is. We construct this matrix by combining our data with theory and precedent from past demand surges (Balla et al. 2022; Verhoeven 2024; Wood and Lewis 2017). Such a matrix could be used to map a government’s exposure and prioritize interventions (§ 7.1).

Likelihood. Two sets of factors may predict the likelihood that a service is flooded: agent capabilities and access, and existing features of the service.

Agent Capabilities and Access. Flooding overall becomes more likely as agent capabilities improve (Factor 1). Some capabilities, e.g. browser use, are likely particularly useful for government interaction (§ 2.2).

A service is more susceptible to flooding if the relevant agent capabilities are broadly accessible (Factor 2). Our cases (§ 4) are driven almost entirely by LLM text generation, a capability now freely available from multiple providers. However, the most capable agents are paid services. Although inference is becoming cheaper overall, access costs or restrictions could prevent flooding, or restrict it to cases caused by wealthier users (Humlum and Vestergaard 2025; Sharp et al. 2025).

Service Features. The design of service interfaces can affect how readily certain capabilities translate into flooding (Factor 3) (Jordan, Peixoto, and Ramos-Maqueda 2026). For example, even if an agent can collect all the relevant information and generate an application, it can have trouble interacting with portals that have complex login flows or JavaScript-heavy front-ends (Drouin et al. 2024; Gur et al. 2023). Similarly, digital agents cannot attend in-person appointments.

Services requiring more effort to submit could be more exposed to flooding (Factor 4). We would expect higher submission costs to suppress a higher share of true demand (§ 2.2). For example, voluntary tax returns can take significant effort, disincentivizing those who only expect a small return. As agents reduce these costs, it is therefore more likely the resulting demand increase causes strain. Conversely, hard limits on submission volume – e.g. one annual tax return per citizen – could limit how much demand agents can add.

Finally, the expected benefit of submission likely predicts flooding itself, regardless of whether agents affect it (Factor 5). Falling submission costs alone can unlock latent demand (Currie 2004). For example, in [Case B] claimants more frequently respond to financially consequential claim denials because it is now easy, though there is no evidence their chance of success increases. We would expect this incentive to be stronger for services that offer higher financial reward.

Severity. Where flooding occurs, its impact could take two forms: first-order operational impacts from the demand surge itself, and second-order impacts and externalities from the responses governments adopt to address it.

Operational Impact. Operational impacts, such as backlogs, affect both the agency handling a service and its users. Their size depends on the efficiency and flexibility of the existing service (Factors 6–9), or its “surge capacity” (Hick, Barbera, and Kelen 2009; Bonnett et al. 2007). For public services, technical indicators of this capacity may include the use of structured and standardized data formats, which can enable “straight-through” processing of some cases without human intervention (Khanna 2008), or digital identity systems, which can avoid time-consuming personhood confirmation (Naghmouchi et al. 2025).

Response Impact. Beyond the direct impacts of flooding, the responses governments may choose often have trade-offs or negative externalities. We map these in § 6. For example, closing a digital submission channel could restrict access to the service. These trade-offs depend both on what response is chosen, and how intensely it is implemented.

Legal constraints can affect which response governments choose. Rigid budget and entitlement policies (Factor 10) and legally mandated processing effort, such as response requirements (Kwoka 2021) or right-to-explanation legislation (Buttaboni and Floridi 2026) (Factor 11), narrow the set of feasible responses.

More intense versions of responses have bigger externalities. For example, a more aggressive eligibility reduction for a welfare service affects more people. How strongly a government must intervene depends on the “slack” (O’Toole and Meier 2010) in the budgets and rules around a service: how much demand could change before they would require updating (Hendrick 2006). Slack is low where budgets assume take-up well below full entitlement (Factor 12), and where a successful submission has a high downstream cost, such as a benefit payout or a court proceeding (Factor 13).

6 How Could Governments Respond?

Governments explicitly respond to flooding in over half (56%) of cases in our dataset. However, these actions are usually tightly scoped or non-binding, like releasing AI use guidance. Precedent suggests a much broader range of possible future responses, which we map to two categories (Table 3): suppress demand for the affected services, or increase processing capacity to meet that demand. These responses would likely be effective, but each induces trade-offs. In particular, demand suppression restricts access to public services and can create procedural inequality.

6.1 Suppress Demand

Governments could suppress demand for services by making applications either more difficult or less attractive to complete, and have frequently done so in the past.

Add Friction. Governments could (re-)raise the cost of submission until enough users are deterred from filing, thus suppressing quantitative flooding. Measures taken during previous demand surges include: financial barriers like fees (Kwoka 2021), procedural barriers like digital identity verification (National Employment Law Project 2023), and direct access limits like per-claimant rate caps (Rust 2024). An obvious candidate measure for agentic flooding is blocking

	Approach	Sample Methods	Benefits	Drawbacks
Suppress Demand	<i>Add Friction</i>	• Fees • In-person appointments • Access limits, e.g. rate caps • Closing digital channels	• Fast to deploy • Few prerequisites • Proven to work	• Can drive procedural inequality • Worsens user experience • Some forms may erode as AI capabilities advance
	<i>Reduce the Expected Benefit</i>	• Reduce entitlement amounts • Narrow eligibility criteria • Fines for unsuccessful applications	• Robust to advances in agent capabilities • Policy option usually established	• Legislative decision, can trade off other priorities • Does not always address operational impact
Increase Capacity	<i>Redesign the Service</i>	• Structured interfaces with identity verification • Deterministic processing pipelines • Proactive “push” delivery	• Technical architecture is established • Can also improve citizen experience	• High upfront investment • Cross-agency coordination required • Often needs legislative change; can take years
	<i>Deploy Agents in Processing</i>	• Intake and triage • User correspondence • Autonomous processing of routine cases	• Scalable; likely effective • Improves with capabilities improvements	• Legal limits on automated decision-making • Provider dependency and sovereignty risks • Ethical and societal risks

Table 3: Overview of possible government responses to agentic flooding, grouped under two strategies.

bots from government websites, or restricting their permissions (Marro et al. 2026).

Most measures to introduce friction require little infrastructural change, have precedent, and can be deployed quickly, including when a surge is acute. In 14 cases (17%) in our dataset, governments have already responded with friction. For example, fees can be raised wherever a payment system is in place: Australia has considered reintroducing fees for FOI requests in response to a wave of AI-generated submissions [Case A]. Access can often be restricted with equally little lead time, as when Japanese authorities blocked submissions to a comment procedure by IP address [Case G]. Even where legal change is needed, it can be narrow: in [Case I], targeted reform stopped middle-man organizations from collecting fees on certain types of requests.

However, introducing friction decreases the accessibility of government services. Friction disproportionately deters poorer, less digitally literate, and otherwise vulnerable users, effectively shaping who accesses public services (Peeters and Widlak 2023). Fees and in-person requirements can thereby create procedural inequality and, where legal services are concerned, undermine access to justice. Friction also worsens user experience of public services and, by extension, trust in government (Van De Walle and Bouckaert 2003; Christensen and Laegreid 2005). Some measures, such as closing digital submission channels, could additionally prevent the accessibility improvements that AI agents promise (Yun, Yun, and Xue 2024).

Friction-inducing measures may also face practical challenges. For one, some forms of friction could become less effective as agent capabilities improve, such as how

CAPTCHAs no longer reliably identify human website visitors (Plesner, Vontobel, and Wattenhofer 2024). They may also be legally prohibited. For example, in many jurisdictions it is illegal to charge fees for applications to welfare services (SSA 2026). Lastly, friction may not deter adversarial actors or directly address qualitative flooding (e.g. longer text submissions).

Reduce the Expected Benefit. Governments could also lower what a user stands to gain from submitting to a service. They could reduce entitlement amounts, narrow eligibility, or increase the cost of unsuccessful applications. Such revisions often reduce service demand, e.g. reforms to poverty assistance in the US in the 90s (Grogger and Karoly 2005).

Entitlement revisions may be required regardless of whether they resolve flooding. Many services are budgeted assuming only limited take-up (Nichols and Zeckhauser 1982). For example, the UK Department for Work and Pensions explicitly anchors benefit budgets to historical take-up rates (GOV.UK 2026). Where budgets cannot be raised to meet full take-up, reducing benefits is a likely response.

A key challenge is that reducing benefits is not a purely executive decision, but a legislative decision. It may be harder to implement because of political priorities, and run counter to other aims, such as reducing poverty or helping the unemployed.

6.2 Increase Capacity

Governments could also respond to flooding by increasing their processing capacity. Most obviously, they may choose to scale existing resources, e.g. by hiring more staff. How-

ever, tight budget constraints are common and existing processes are often inefficient (Dunleavy et al. 2006). We list two alternative options.

Redesign the Service. Service redesign can include restructuring service interfaces, integrating identity verification, standardizing data formats, and providing some services proactively without applications (Dunleavy and Margetts 2015). Such redesign could make government more efficient and free up resources for addressing flooding. Moreover, as discussed in § 5.2, structured forms and identity verification could decrease both the likelihood and the severity of flooding.

The main disadvantage is the long lead times and proactive financial investment required for implementation. Structural redesign requires cross-agency coordination and, in many cases, legislative change (Pollitt 2017; Hood 1991). Bigger projects can take years or even decades, such as introducing central registers or data standards; many Western governments have famously taken decades to get digital infrastructure in place despite well-documented benefits. As such, redesign is likely unavailable as a short-term measure.

In 13 cases (15%), we code the measures governments take in response to flooding as redesign – though they are exclusively small changes, such as introducing a service to verify the legitimacy of case numbers [Case H], rather than broad structural reform.

Deploy Agents in Processing. Governments could use AI agents themselves – both to accelerate back-end processing, and to detect and prevent fraud. Governments are already taking first steps in this direction: in 21 cases (25%) in our dataset, they introduce AI tools explicitly in response to flooding, e.g. to analyze sentiment across large numbers of public comments [Case F] and to improve detection of AI-generated fake medical certificates [Case D]. In the vast majority of cases these are bounded tools handling one step in the process, consistent with established patterns of slow, piecemeal AI diffusion in public-sector organizations (Neumann, Guirguis, and Steiner 2024).

More comprehensive agent deployment lacks precedent, but evidence suggests it could scale processing throughput drastically. AI tools are already used in some countries for intake screening, triage, drafting of standard correspondence, and even autonomous handling of routine cases (Straub et al. 2024; OECD 2025a). Government work is likely suitable for agent deployment because bureaucratic processes and rules are generally well-documented and exhaustive (Schmitz, Rysstrøm, and Bätzner 2025). Benchmarks also suggest that agents are becoming increasingly competent in tasks similar to government work (Patwardhan et al. 2025).

However, there remain significant legal and practical risks regarding government use of agents. Laws often prevent automated government decision-making (Grimmelikhuijsen and Meijer 2022) and evaluating agents on government tasks remains difficult (Rysstrøm et al. 2026). Civil society groups regularly voice concerns about decision transparency, bias, and accountability diffusion (Ada Lovelace Institute 2021). Careless deployment could further expose government bod-

ies to provider lock-in or sovereign-data risk (Burwell and Propp 2022).

7 Discussion and Recommendations

We find broad evidence of agentic flooding, but our work does not suggest it poses a severe *operational* risk to governments. For one, responses to flooding so far are diverse and overall small in scale: piloting AI tools to help with individual steps, issuing guidance on AI use, or precise access limitations. Looking forward, while uncertainty remains about when particular agent capabilities arrive or diffuse widely, friction-inducing measures are likely to mitigate the operational impacts of flooding, as they have for previous demand surges following technological shifts (Verhoeven 2024; Balla et al. 2022).

However, adding friction worsens user experience of public services, potentially introduces procedural inequality, and prevents some of the accessibility gains that AI agents enable (Ilves et al. 2025). In contrast, capacity-building responses could address flooding without these trade-offs, but they require lead time, technical investment, and, in many cases, cross-departmental coordination; they are unlikely to be feasible once a surge is already underway.

One plausible trajectory is therefore that demand suppression again becomes a default response. Governments might reach for it because, when a surge occurs, it is the only intervention available on a short enough timeline. This pattern matches what Lindblom (1959) terms “muddling through” – iterative, short-term, and mostly reactive adaptation, which keeps government organizations operational, but does not usually optimize outcomes for the public.

7.1 Three Near-Term Recommendations

Preparing for flooding proactively may avoid governments having to (re)introduce friction to address it. We propose three near-term actions to this effect.

Auditing Exposure. Governments should systematically audit their services for vulnerability to flooding. Key factors predicting the susceptibility of a service can be assessed today, and updated with future information about AI capabilities (§ 5.2). A portfolio of services could be mapped and ranked against these factors, e.g. using our risk matrix, yielding a prioritized list of services and the response classes (§ 6) most applicable to each.

Developing a Digital Identity Strategy. Strong identity verification could effectively mitigate quantitative flooding because it allows automated enforcement of per-claimant rate limits and makes adversarial flooding more difficult. Digital identity also supports pre-population of known data, thus lowering the effort required for entitled users to apply, and for governments to process their case. Over 100 jurisdictions already have digital identity infrastructure, but its maturity varies drastically (Metz, Casher, and Clark 2024). Where feasible, integrating it into the most exposed services is likely to be a high-leverage action to preempt flooding.

Establishing Legal Certainty. Several of the response options in § 6 are either illegal or of unclear legality in some jurisdictions. For example, many uses of AI in government processing are constrained by AI and administrative law (Buttaboni and Floridi 2026), and it can be unclear where restricting free-form or digital submission channels, or raising fees, is lawful. This ambiguity may itself be a barrier to preventative action. Governments should commission internal legal reviews to clarify, for each response class, which forms of it are permissible under current law. This understanding would be valuable regardless of the order and intensity in which services are flooded.

8 Limitations and Future Work

The methodology we employ aims to collect evidence of an emerging phenomenon across jurisdictions and domains. However, it allows neither statistical claims about prevalence, nor causal claims about AI’s role in flooding. We suggest future research directions building on this work. To support such research, our dataset is published with this paper.

Collection Methodology. Our semi-automated pipeline only collects positive instances of potential flooding and does not measure what share of public services (per jurisdiction) are assessed. As such, the statistics presented in § 4–§ 6 cannot be taken to generalize within or beyond the studied countries, and we do not make comparative claims, e.g. of prevalence per country.

Further sources of uncertainty or bias include the selection of countries and domains studied, the selection of “candidate services” presented during discovery, differing traditions of publicly discussing administrative matters, and the differing performance of LLMs in different languages (all prompts were given in English).

Further, while we increase our confidence in included cases by requiring explicit third-party attribution of the surge to AI, that criterion likely excludes many cases that meet our definition of flooding, and could wrongly include some others. We also do not require time-series volume data to include a case, and only 25 possess it. Accordingly, we cannot make even correlative quantitative claims.

Finally, despite validating all cases with multiple instances of human review, we cannot exclude the risk of LLM hallucination. We only provide human-LLM agreement numbers, perform no inter-rater coding, and do not benchmark the performance of the LLM against humans.

Based on these limitations, a pressing direction for future work is more rigorous analysis of the relative prevalence of flooding for specific government services. Candidate methodologies may include quantitative analysis of volume time series, surveys of citizens about their use of AI agents, and analysis of agent usage data, where available. A further promising direction is more detailed analysis of the relevance of digital-government maturity for the risk and severity of flooding, e.g. analyzing how effective digital identity is at mitigation.

Identification of Risk Factors and Mitigations. Our contributions in § 5 and § 6 are largely based on precedent

and theory, rather than our data. We do not place our data in the risk matrix (§ 5.1). We also do not attempt to measure the success of any initial government responses we record, and the taxonomy of possible responses abstracts over jurisdiction-specific legal constraints.

This suggests a number of natural next steps for future work. The predictive value of the listed risk factors should be empirically validated within a single jurisdiction. Government responses should be measured empirically for both their effectiveness and the externalities they produce. Both the proposed risk matrix and response taxonomies should be continually updated as a result.

9 Related Work

Concurrent Work. Piedrahita et al. (2026) theorize “congested bureaucracy” as a result of AI-driven friction reduction; we validate and extend their framing here. Shah and Levy (2026) find evidence that suggests AI use is driving a “dramatic increase” in cases before US federal courts.

Agents and Friction Reduction. Shahidi et al. (2025) identify that agents may lower search, communication, and contracting costs in markets. Brynjolfsson and Hitzig (2025) theorize that agents may reduce the cost of knowledge formalization.

AI and Government Interaction. Jordan, Peixoto, and Ramos-Maqueda (2026) measure the accessibility of government service interfaces to AI agents. Yun, Yun, and Xue (2024) propose using LLMs to improve policy communication. Majithia et al. (2026) benchmark LLMs on answering citizen queries. Zhang and Nie (2025) find that AI tools improve some citizen-government interactions.

Administrative Burden. Madsen, Mikkelsen, and Moynihan (2022) review how the transaction costs faced by citizens are studied under overlapping terms, including “sludge”, “red tape”, and “ordeals”. Zeckhauser (2019) examines a case where burdens are maintained intentionally.

10 Conclusion

Agentic flooding is plausibly occurring across a broad range of government services, in the form of LLM-generated text submitted via permissive interfaces. Though it may intensify as agent capabilities improve, acute operational collapse appears unlikely. However, it is likely that – at least in some cases – budgets will need adjusting or user experience will suffer, particularly where governments introduce friction to reduce service demand.

On a more positive note, mitigating agentic flooding is an opportunity to transform government services more broadly: many governments recognize the value of AI-enabled government interaction, and the digital-government literature has long advocated similar structural reforms for the benefit of the public (Dunleavy et al. 2006). While friction-based responses lock this potential further out of reach, proactive interventions to build capacity and redesign services could instead help to realize the user-experience benefits that have long been possible.

Governments can act now. Our analysis suggests three near-term, proactive measures whose value does not depend on tracking the progress of AI capabilities: auditing service vulnerability, developing digital-identity integration strategies for the most exposed services, and clarifying the legal basis for resilience-building measures.

Acknowledgements

CS acknowledges funding by the Dieter Schwarz Foundation via the Hertie School. This work was partially completed during a seasonal fellowship at GovAI.

A Research Scope

Country selection. We select twelve countries to capture variation geographically and along two institutional dimensions. First, digital government maturity, as measured by the OECD Digital Government Index (OECD 2025b). We include three of the five highest-rated countries: South Korea, Australia, and the United Kingdom. Three further included countries score above 0.8, indicating middle-to-high maturity: Estonia, Denmark, and France. We include three that score under 0.8, indicating lower maturity, e.g. higher reliance on non-digital processes: Brazil, Japan, and the Netherlands. Finally, we include three countries for which OECD data is not available: the United States, Germany, and Singapore. Beside aggregate maturity, these states vary in specific technical implementation of digital government. For example, four studied countries (Estonia, Denmark, Singapore, South Korea) have mandatory digital identity systems.

Second, administrative tradition (Painter and Peters 2010): the list includes Anglo-American (UK, US, Australia), Napoleonic (France, Netherlands), Germanic (Germany), Scandinavian (Denmark), East Asian (Japan, South Korea, Singapore), Latin American (Brazil), and post-Soviet (Estonia) systems. Legally, these countries are split almost evenly between common-law and continental systems.

We select this list of countries to lower the chance of obvious confounders from these two dimensions, especially for our analyses of risk and prevalence of flooding. However, the list is evidently not representative, such that we do not make comparative or analytical claims about the distribution of flooding cases we find. Future work analyzing the relevance of these factors for flooding would likely be valuable, as we discuss in § 8.

Domain selection. Within each country we scan a fixed list of government interaction channels. Existing taxonomies of the functions of government, e.g. COFOG by the OECD (2025b), are not suitable for this because they map government responsibilities, not citizen interfaces.

Instead, we compile a list of domains to scan by comparing the top-level citizen service categories used by the national service portals of three of our sample countries: the UK, the US, and France (Table 4). We map ten public service categories that return across all three of these portals. However, as clarified in our definition (§ 2.1), some points of government interaction are not public services. We therefore append three domains for such channels: participatory processes, regulatory complaints and reporting, and

transparency and access to information. Of these, only complaints have a top-level equivalent on any of the three portals.

This methodology produces a plausible list of 13 government domains, validated in three of the 12 studied countries and enabling the broad and indicative scans for flooding we seek to make. However, it is unlikely to be optimal. More detailed refinement of this domain list could improve it, e.g. by comparing against existing taxonomies or scanning more countries. Accordingly, we name this list of domains as a possible source of bias in § 8.

B Case Collection Pipeline

This appendix documents the pipeline used to collect the dataset presented in the paper. We describe the three pipeline stages and the prompts used in each (§ B.1), the schema each case is coded against (§ B.2), and the scaffolding that orchestrates the pipeline (§ B.3). The full case records, prompts, data schema, and harness code are available in the linked dataset repository.

B.1 Pipeline Stages

As depicted in Figure 2, the pipeline has three stages: discovery, iteration, and finalization. An LLM is called during discovery and iteration, and all LLM calls return structured responses. We here describe each stage, provide summary statistics of cases processed in each, and give an overview of prompts and response schemas passed to the LLM.

Discovery From a domain prompt combining a country and a domain, e.g. “France, Health Services”, an LLM with web search access returns 20–30 candidate services to investigate. We survey 12 countries and the same 13 domains in each (§ A). Each candidate is returned as a stub, which is expanded into a case template deterministically. We then manually review the candidates and decide whether to investigate or remove each: of 2288 candidates, we pass 2210 (96%) on to iteration.

System prompt (opening)

You are a research assistant identifying candidate government services which may be experiencing "agentic flooding": surges of demand driven by AI use. You will receive a user input consisting of a country, and a domain of government services to investigate in that country. Your task is to return an initial list of 5–30 specific government services or interfaces which could be flooded in that domain. This is the first step in the pipeline, and these cases will be explored in detail afterward.

[...]

User prompt

Country to analyze: {country}
Government domain to analyze: {domain}.

Response schema

Response: DiscoveryOutput

DiscoveryCandidate:

Surveyed domains	United Kingdom GOV.UK https://www.gov.uk/browse	United States USA.gov https://www.usa.gov/	France Service-Public.fr https://www.service-public.gouv.fr/particuliers/vosdroits/theme
<i>Public Services</i>			
1 Benefits & Social Protection	Benefits Disabled people	Government benefits Disability services	Social – Santé
2 Health Services	–	Health	Social – Santé
3 Citizenship, Identity & Elections	Citizenship and living in the UK Births, deaths, marriages and care	Voting and elections	Papiers – Citoyenneté – Élections
4 Education	Education and learning Childcare and parenting	Education	Famille – Scolarité
5 Employment, Jobs & Pensions	Working, jobs and pensions Employing people Business and self-employed	Jobs, labor laws and unemployment Small business	Travail – Formation
6 Housing & Local Services	Housing and local services	Housing help	Logement
7 Immigration, Asylum & Visas	Visas and immigration Passports, travel and living abroad	Immigration and U.S. citizenship	Étranger – Europe
8 Justice & Legal Services	Crime, justice and the law	Laws and legal issues Scams and fraud	Justice
9 Tax & Revenue Administration	Money and tax	Taxes Money and credit	Argent – Impôts – Consommation
10 Transport, Driving & Licensing	Driving and transport	Travel	Transports – Mobilité
<i>Other Channels</i>			
11 Participatory Processes	–	–	–
12 Regulatory Complaints & Reporting	–	Complaints	–
13 FOI & Transparency	–	–	–
<i>Not Mapped</i>	Environment and countryside	Disasters and emergencies Military and veterans Innovation	Associations Loisirs – Sports – Culture

Table 4: Top-level citizen service categories on the national service portals of the UK, US, and France, and the domains we derive from them.

```

service_name: str
country: str
government_body: str
rationale: str

```

```

DiscoveryOutput:
  candidates: list[DiscoveryCandidate]

```

Iteration Each selected case then cycles between a refinement call and a review call, with no human involvement, until the review call recommends finalization or removal, or a maximum of five cycles is reached.

Refinement. The refinement call receives the current state of the case, a prompt to improve it by coding it against the case template (Dunivin 2025), and any instructions from the previous iteration, review, or human reviewer. It returns a set of JSON Patch operations to apply to the case file, a recommendation to either continue iterating or pass the case to review, and a description of next steps for further iteration, if any. The harness applies the patches deterministically to the stored case and proceeds according to the recommendation.

The LLM does not produce source URLs as part of its response. Sources are extracted deterministically from the web

search grounding metadata returned with each API response, and added to the case data. For some elements of the case data, the model returns a text description of the source used to aid the human reviewer. We append any URL returned by API response metadata, regardless of whether the model substantively used it. We manually verified all cases listed in the final dataset, but the broader source list in each case file often contains irrelevant links. We choose this approach as generating URLs as part of the response was very susceptible to hallucination (Huang et al. 2025), while line-specific grounding is not available when using structured outputs.

System prompt (opening)

You are a research assistant helping build a dataset of "agentic flooding" cases for an academic paper. Agentic flooding is when AI use by citizens drives a surge in demand on a government service, either in volume, in complexity, or both. The paper asks whether this is happening, where, and what governments could do about it.

Your job is to improve a single case record by performing web search to fill in or correct fields, and to judge whether the case meets each of three inclusion criteria.

[...]

User prompt

{Case JSON}

Response schema

Response: IterationOutput

```
Patch:
  path: str
  value: Any
  rationale: str

IterationOutput:
  patches: list[Patch]
  summary_of_changes: str
  next_steps: str
  recommendation: str
```

Review. The review call skeptically evaluates the current case and returns a recommendation of finalize, remove, or iterate, with optional field-level concerns. It does not modify the case. Its output becomes input either to a human decision, when it recommends finalization or removal, or to the next refinement call, when it recommends further iteration.

System prompt (opening)

You are a skeptical reviewer evaluating a candidate case for an academic dataset of "agentic flooding" | situations where AI use by citizens drives surges in demand on government services. A research assistant has iterated on this case across multiple passes. Your job is to decide what happens to it next: finalize it into the dataset, send it back for more work, or remove it as unsuitable.

[...]

Field	Type
<i>Identification</i>	
case_id	string
service_name	string
country	string
jurisdiction_detail	string
government_body	string
include	bool
tldr	string
analysis	string
<i>Types of Flooding (§ 2.1)</i>	
flooding.qualitative	bool
flooding.quantitative	bool
<i>Mechanism (inclusion criterion 1)</i>	
mechanism.meets_inclusion_criteria	bool
mechanism.inclusion_explanation	string
mechanism.ai_description	string
mechanism.primary_is_text_generation	bool
mechanism.interface_description	string
<i>Evidence of change (inclusion criterion 2)</i>	
evidence.meets_inclusion_criteria	bool
evidence.explanation	string
<i>AI attribution (inclusion criterion 3)</i>	
attribution.meets_inclusion_criteria	bool
attribution.explanation	string
attribution.source	string
attribution.description	string
<i>Volume time series</i>	
volume.unit	string
volume.annual.{2018..2025}	number
volume.source_notes.{2018..2025}	string
<i>Government response (§ 6)</i>	
response.action_explicitly_for_ai	bool
response.action_for_general_surge	bool
response.description	string
response.categories	string[]
<i>Sources</i>	
sources[].url	string
sources[].origin	string
<i>Pipeline metadata</i>	
pipeline.status	string
pipeline.discovery_country	string
pipeline.discovery_domain	string
pipeline.discovery_rationale	string
pipeline.iteration_count	integer
pipeline.last_summary_of_changes	string
pipeline.last_next_steps	string
pipeline.last_review_concerns	string
pipeline.human_review_comment	string
pipeline.last_error	string

Table 5: Fields collected for each case, grouped by section. Field descriptions are released with the dataset repository.

User prompt

```
{Case JSON}
```

Response schema

```
Response: ReviewOutput
```

```
ReviewConcern:  
  field_path: str  
  concern: str
```

```
ReviewOutput:  
  recommendation: str  
  rationale: str  
  specific_concerns: list[ReviewConcern]
```

Finalization We manually review every case the review call recommends for finalization or removal. Each is either added to the dataset, removed, or returned to iteration with concrete improvement instructions. Of 2210 cases, we take the recommended action in 2182 (98%); the most recommended action is removal (95% of cases).

B.2 Case Schema

Each case is stored as a JSON document. Table 5 lists the fields collected and their types, grouped by section. The full schema, including a description of every field, is available in the dataset repository. The schema mirrors the structure of our definition (§ 2). The three sections corresponding to our inclusion criteria (§ 3.2) – mechanism, evidence, attribution – each contain a `meets_inclusion_criteria` boolean, and a case is included in the dataset only if all three are true.

B.3 Scaffolding

Storage. Each case is stored as a JSON file on disk, named by its `case_id`, and versioned in the repository. An append-only log records the output of every iteration, including the patches applied, the LLM’s summary of changes, and the recommendation returned.

Human-in-the-loop. A minimal web interface allows a human reviewer (the first author) to make two decisions: (i) which candidates from discovery to advance to iteration, and (ii) for each case returned by the review stage, whether to finalize, exclude, or return to iteration with optional written feedback.

Models, cost, and reproducibility. All *discovery* and *review* calls used Gemini-3.1-Pro-Preview via the Gemini API; all *iteration* calls used Gemini-3.1-Flash-Lite. Web search grounding was enabled for *discovery* and *iteration*, but not for *review* calls. Temperature was set to 0 for all calls. Total token cost for producing the final dataset was approximately \$200. Full prompts and orchestration code are included in the linked dataset repository to support replication. Both the models used and the web search grounding component cannot be guaranteed to be deterministic; re-running the pipeline may not produce an identical dataset.

References

- Ada Lovelace Institute. 2021. Algorithmic Accountability for the Public Sector. Technical report, Ada Lovelace Institute.
- Balla, S. J.; Bull, R.; Dooling, B. C.; Hammond, E.; Herz, M.; Livermore, M.; and Noveck, B. S. 2022. Responding to Mass, Computer-Generated, and Malattributed Comments. *Administrative Law Review*, 74(1): 95–160.
- Bhargava, S.; and Manoli, D. 2015. Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review*, 105(11): 3489–3529.
- Blundell, R.; Bozio, A.; and Laroque, G. 2013. Extensive and Intensive Margins of Labour Supply: Work and Working Hours in the US, the UK and France*. *Fiscal Studies*, 34(1): 1–29.
- Bonnett, C. J.; Peery, B. N.; Cantrill, S. V.; Pons, P. T.; Haukoos, J. S.; McVane, K. E.; and Colwell, C. B. 2007. Surge Capacity: A Proposed Conceptual Framework. *The American Journal of Emergency Medicine*, 25(3): 297–306.
- Brynjolfsson, E.; and Hitzig, Z. 2025. *AI’s Use of Knowledge in Society*, chapter 12. University of Chicago Press.
- Burwell, F.; and Propp, K. 2022. Digital Sovereignty in Practice: The EU’s Push to Shape the New Global Economy. Technical report, Atlantic Council.
- Buttaboni, C.; and Floridi, L. 2026. A Regulatory Taxonomy of AI Opacity in the EU: Rethinking Transparency, Traceability, Interpretability, and Explainability. *AI and Ethics*, 6(1): 100.
- Canales, S. B. 2025. ‘Addiction to Secrecy’: Opposition and Crossbench Slam Labor’s ‘Undemocratic’ Changes to FoI – Including Charging Fees. *The Guardian*.
- Christensen, T.; and Laegreid, P. 2005. Trust in Government: The Relative Importance of Service Satisfaction, Political Factors, and Demography. *Public Performance & Management Review*, 28(4): 487–511.
- Currie, J. 2004. The Take Up of Social Benefits. Technical Report w10488, National Bureau of Economic Research, Cambridge, MA.
- DoNotPay. 2026. Save Time and Money with DoNotPay! <https://donotpay.com/>.
- Drouin, A.; Gasse, M.; Caccia, M.; Laradji, I. H.; Verme, M. D.; Marty, T.; Vazquez, D.; Chapados, N.; and Lacoste, A. 2024. WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks? In *Proceedings of the 41st International Conference on Machine Learning*, 11642–11662. PMLR.
- Dunivin, Z. O. 2025. Scaling Hermeneutics: A Guide to Qualitative Coding with LLMs for Reflexive Content Analysis. *EPJ Data Science*, 14(1).
- Dunleavy, P.; and Margetts, H. 2015. Design Principles for Essentially Digital Governance. In *American Political Science Association Annual Meeting*, 111. USA.
- Dunleavy, P.; Margetts, H.; Bastow, S.; and Tinkler, J. 2006. *Digital Era Governance: IT Corporations, the State, and e-Government*. Oxford University Press.

- Garnett, H. A. 2022. Registration Innovation: The Impact of Online Registration and Automatic Voter Registration in the United States. *Election Law Journal: Rules, Politics, and Policy*, 21(1): 34–45.
- GII. 2026. § 7 BHO - Einzelnorm. https://www.gesetze-im-internet.de/bho/_7.html.
- GOV.UK. 2026. Benefit Expenditure and Caseload Tables: Guidance and Methodology. <https://www.gov.uk/government/publications/benefit-expenditure-and-caseload-tables-guidance-and-methodology>.
- Grimmelikhuijsen, S.; and Meijer, A. 2022. Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, 5(3): 232–242.
- Groger, J. T.; and Karoly, L. A. 2005. *Welfare Reform: Effects of a Decade of Change*. Harvard University Press.
- Gur, I.; Furuta, H.; Huang, A. V.; Safdari, M.; Matsuo, Y.; Eck, D.; and Faust, A. 2023. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. In *The Twelfth International Conference on Learning Representations*.
- Halling, A.; and Baekgaard, M. 2024. Administrative Burden in Citizen–State Interactions: A Systematic Literature Review. *Journal of Public Administration Research and Theory*, 34(2): 180–195.
- Hendrick, R. 2006. The Role of Slack in Local Government Finances. *Public Budgeting & Finance*, 26(1): 14–46.
- Hick, J. L.; Barbera, J. A.; and Kelen, G. D. 2009. Refining Surge Capacity: Conventional, Contingency, and Crisis Capacity. *Disaster Medicine and Public Health Preparedness*, 3(S1): S59–S67.
- Hood, C. 1991. A Public Management for All Seasons? *Public Administration*, 69(1): 3–19.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Humlum, A.; and Vestergaard, E. 2025. The Unequal Adoption of ChatGPT Exacerbates Existing Inequalities among Workers. *Proceedings of the National Academy of Sciences*, 122(1): e2414972121.
- Ilves, L.; Kilian, M.; Peixoto, T.; and Velsberg, O. 2025. The Agentic State: How Agentic AI Will Revamp 10 Functional Layers of Government and Public Administration. Technical report, The Agentic State.
- Isenko, Z.; Strand, S.; Chen, Y.; Aimard, G.; Codreanu, M.; Sampathkumar, V.; Imas, A.; Jacobs, J.; Muiruri, E.; Mateos-Garcia, J.; Ng, J. J.; Javed, S.; Martin, J.; Ajmeri, O.; Calin, D.; Kim, A.; Millet, F. C.; and Manyika, J. 2026. Google’s AI & Economy ATLAS v1.0: Mapping Gemini Usage in the Economy. arXiv:2608.00038.
- Jordan, L.; Peixoto, T. C.; and Ramos-Maqueda, M. 2026. RADAR: Readiness for AI Discovery and Agentic Reach — Measuring the Accessibility of Digital Government Services to AI Systems. Preprint, The Agentic State.
- Kaplan, S.; and Garrick, B. J. 1981. On The Quantitative Definition of Risk. *Risk Analysis*, 1(1): 11–27.
- Kasirzadeh, A.; and Gabriel, I. 2025. Characterizing AI Agents for Alignment and Governance. arXiv:2504.21848.
- Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; and Kasneci, G. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103: 102274.
- Khanna, A., ed. 2008. *Straight through Processing for Financial Services: The Complete Guide*. Complete Technology Guides for Financial Services Series. Amsterdam Boston: Academic Press. ISBN 978-0-08-055484-6 978-0-12-466470-8.
- Kwoka, M. B. 2021. *Saving the Freedom of Information Act*. Cambridge: Cambridge University Press. ISBN 978-1-108-71089-3 978-1-108-69763-7.
- Lindblom, C. E. 1959. The Science of “Muddling Through”. *Public Administration Review*, 19(2): 79–88.
- LTO. 2026. Sozialgerichte: Überlastung Durch KISchriftensätze. *Legal Tribune Online*.
- Madsen, J. K.; Mikkelsen, K. S.; and Moynihan, D. P. 2022. Burdens, Sludge, Ordeals, Red Tape, Oh My!: A User’s Guide to the Study of Frictions. *Public Administration*, 100(2): 375–393.
- Majithia, N.; Shinde, R.; Chapman, Z.; Trital, P.; Decker, J.; Maskey, M.; Simperl, E.; and Shadbolt, N. 2026. The CitizenQuery Benchmark: A Novel Dataset and Evaluation Pipeline for Measuring LLM Performance in Citizen Query Tasks. <https://arxiv.org/abs/2602.04064v1>.
- Marro, S.; Chan, A.; Ren, X.; Hammond, L.; Wright, J.; Wanga, G.; Piccardi, T.; Campos, N.; South, T.; Yu, J.; Sengupta, S.; Sommerlade, E.; Pentland, A.; Torr, P.; and Pei, J. 2026. Permission Manifests for Web Agents. arXiv:2601.02371.
- Metz, A.; Casher, C.; and Clark, J. 2024. *ID4D Global Dataset Volume 2: Digital Identification Progress and Gaps*. Washington, DC: World Bank.
- Moore, M. H. 1995. *Creating Public Value: Strategic Management in Government*. Cambridge, Mass: Harvard University Press. ISBN 978-0-674-17557-0.
- Moynihan, D.; Herd, P.; and Harvey, H. 2015. Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions. *Journal of Public Administration Research and Theory*, 25(1): 43–69.
- Naghmouchi, M.; Laurent, M.; Levallois-Barth, C.; and Kaaniche, N. 2025. Perspectives on National Digital Identity Systems. *Blockchain: Research and Applications*, 100429.
- National Employment Law Project. 2023. ID Verification. Technical report, National Employment Law Project.
- Neumann, O.; Guirguis, K.; and Steiner, R. 2024. Exploring Artificial Intelligence Adoption in Public Organizations:

- A Comparative Case Study. *Public Management Review*, 26(1): 114–141.
- Nichols, A. L.; and Zeckhauser, R. J. 1982. Targeting Transfers through Restrictions on Recipients. *The American Economic Review*, 72(2): 372–377.
- O’Brien, D. T.; Offenhuber, D.; Baldwin-Philippi, J.; Sands, M.; and Gordon, E. 2016. Uncharted Territoriality in Co-production: The Motivations for 311 Reporting. *Journal of Public Administration Research and Theory*.
- OECD. 2025a. *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. OECD Publishing. ISBN 978-92-64-43767-8 978-92-64-68405-8 978-92-64-81828-6.
- OECD. 2025b. *Government at a Glance 2025*. Government at a Glance. OECD Publishing. ISBN 978-92-64-31390-3 978-92-64-36321-2 978-92-64-59508-8.
- Osborne, S. P. 2018. From Public Service-Dominant Logic to Public Service Logic: Are Public Service Organizations Capable of Co-Production and Value Co-Creation? *Public Management Review*, 20(2): 225–231.
- O’Toole, L. J.; and Meier, K. J. 2010. In Defense of Bureaucracy: Public Managerial Capacity, Slack and the Dampening of Environmental Shocks. *Public Management Review*, 12(3): 341–361.
- Painter, M.; and Peters, B. G., eds. 2010. *Tradition and Public Administration*. Palgrave Macmillan UK.
- Patwardhan, T.; Dias, R.; Proehl, E.; Kim, G.; Wang, M.; Watkins, O.; Fishman, S. P.; Aljube, M.; Thacker, P.; Fauconnet, L.; Kim, N. S.; Chao, P.; Miserendino, S.; Chabot, G.; Li, D.; Sharman, M.; Barr, A.; Glaese, A.; and Tworek, J. 2025. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. <https://arxiv.org/abs/2510.04374v1>.
- Peeters, R.; and Widlak, A. C. 2023. Administrative Exclusion in the Infrastructure-Level Bureaucracy: The Case of the Dutch Daycare Benefit Scandal. *Public Administration Review*, 83(4): 863–877.
- Piedrahita, D. G.; Banerjee, D.; Blin, K.; Cobben, P.; Corsi, G.; Huang, X. A.; Li, C.; Majumder, S.; Pandey, P. S.; Simko, S.; Strauss, I.; Zhang, T. J.; Anderson, A.; Bengio, Y.; Bethge, M.; Grosse, R.; Helbig, K.; Lie, D.; Mallah, R.; Mihalcea, R.; Nesbitt, S.; Perry, S.; Resnick, P.; Russell, S.; Sachan, M.; Schölkopf, B.; Tang, A.; and Jin, Z. 2026. AI Poses Risks to Democratic and Social Systems.
- Plesner, A.; Vontobel, T.; and Wattenhofer, R. 2024. Breaking reCAPTCHA v2. arXiv:2409.08831.
- Pollitt, C. 2017. *Public Management Reform: A Comparative Analysis - into the Age of Austerity*. New York, NY: Oxford University Press, fourth edition edition. ISBN 978-0-19-879517-9 978-0-19-879518-6.
- Rust, S. 2024. The Vexatious Litigant Problem. <https://houstonlawreview.org/article/127507-the-vexatious-litigant-problem>.
- Ryström, J.; Schmitz, C.; Korgul, K.; Batzner, J.; and Russell, C. 2026. Agent Benchmarks Fail Public Sector Requirements. In *IASEAI 2026*. arXiv.
- Sadeddine, Z.; Maxwell, W.; Varoquaux, G.; and Suchanek, F. M. 2025. Large Language Models as Search Engines: Societal Challenges.
- Schmitz, C.; Ryström, J.; and Batzner, J. 2025. Oversight Structures for Agentic AI in Public-Sector Organizations. In Kamalloo, E.; Gontier, N.; Lu, X. H.; Dziri, N.; Murty, S.; and Lacoste, A., eds., *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, 298–308. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-264-0.
- Shah, A. V.; and Levy, J. Y. 2026. Access to Justice in the Age of AI: Evidence from U.S. Federal Courts.
- Shahidi, P.; Rusak, G.; Manning, B. S.; Fradkin, A.; and Horton, J. J. 2025. The Coasean Singularity? Demand, Supply, and Market Design with AI Agents. In *The Economics of Transformative AI*. University of Chicago Press.
- Sharp, M.; Bilgin, O.; Gabriel, I.; and Hammond, L. 2025. Agentic Inequality. arXiv:2510.16853.
- Siciliani, L.; Borowitz, M.; and Moran, V., eds. 2013. *Waiting Time Policies in the Health Sector: What Works?* OECD Health Policy Studies. OECD. ISBN 978-92-64-17906-6 978-92-64-17908-0.
- SSA, ORDP. 2026. State Plans for Medical Assistance. https://www.ssa.gov/OP_Home/ssact/title19/1902.htm.
- Straub, V. J.; Hashem, Y.; Bright, J.; Bhagwanani, S.; Morgan, D.; Francis, J.; Esnaashari, S.; and Margetts, H. 2024. AI for Bureaucratic Productivity: Measuring the Potential of AI to Help Automate 143 Million UK Government Transactions. arXiv:2403.14712.
- Van De Walle, S.; and Bouckaert, G. 2003. Public Service Performance and Trust in Government: The Problem of Causality. *International Journal of Public Administration*, 26(8-9): 891–913.
- van Oorschot, W. 1991. Non-Take-Up of Social Security Benefits in Europe. *Journal of European Social Policy*, 1(1): 15–30.
- Verhoeven, T. 2024. ‘My How I Have Walked and Worked to Get Those Names’: Petitioning and the Women’s Suffrage Movement in the United States, 1908–1920. *Women’s History Review*, 33(5): 669–691.
- Wood, A. K.; and Lewis, D. E. 2017. Agency Performance Challenges and Agency Politicization. *Journal of Public Administration Research and Theory*, 27(4): 581–595.
- Ye, A.; Maiti, A.; Schmidt, M.; and Pedersen, S. J. 2024. A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet*, 16(5): 167.
- Yun, L.; Yun, S.; and Xue, H. 2024. Improving Citizen-Government Interactions with Generative Artificial Intelligence: Novel Human-Computer Interaction Strategies for Policy Understanding through Large Language Models. *PLOS One*, 19(12): e0311410.
- Zeckhauser, R. 2019. Strategic Sorting: The Role of Ordeals in Health Care. Technical Report w26041, National Bureau of Economic Research, Cambridge, MA.

Zhang, R.; and Nie, L. 2025. Enhancing Citizen-Government Communication with AI: Evaluating the Impact of AI-Assisted Interactions on Communication Quality and Satisfaction. <https://arxiv.org/abs/2501.10715v3>.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854.