Agent Benchmarks Fail Public Sector Requirements

Jonathan Rystrøm[1,*] **Chris Schmitz**[2] **Karolina Korgul**[1] **Jan Batzner**[3,4]

Chris Russell[1]

[1]Oxford Internet Institute, University of Oxford, Oxford, United Kingdom
[2]Centre for Digital Governance, Hertie School, Berlin, Germany
[3]Weizenbaum Institute, Berlin, Germany
[4]Technical University of Munich, Munich, Germany
jonathan.rystrom@oii.ox.ac.uk

Abstract

Deploying Large Language Model-based agents (LLM agents) in the public sector requires assuring that they meet the stringent legal, procedural, and structural requirements of public-sector institutions. Practitioners and researchers often turn to benchmarks for such assessments. However, it remains unclear what criteria benchmarks must meet to ensure they adequately reflect public-sector requirements, or how many existing benchmarks do so. In this paper, we first define such criteria based on a first-principles survey of public administration literature: benchmarks must be *process-based*, *realistic*, *public-sector-specific* and report *metrics* that reflect the unique requirements of the public sector. We analyse more than 1,300 benchmark papers for these criteria using an expert-validated LLM-assisted pipeline. Our results show that no single benchmark meets all of the criteria. Our findings provide a call to action for both researchers to develop public sector-relevant benchmarks and for public-sector officials to apply these criteria when evaluating their own agentic use cases.

1 Introduction

In liberal democracies, public-sector organisations (PSOs) are expected to be neutral, efficient, unbiased entities faithfully enacting democratically negotiated laws [7]. To ensure this, most jurisdictions place stringent requirements upon them: process legitimacy, political neutrality, transparency, and impersonality, to name a few [see, e.g., 32, 82]. Breaches of these requirements can significantly undermine citizen trust in government [49]; where breaches become systematic or intentional, they may cause shifts in the "narrow corridor" balance between state and societal power [2], threatening democratic stability.

The introduction of *Large Language Model-based Agents (LLM Agents)* in public-sector institutions poses an unprecedented threat to these requirements. The promise (and peril) of LLM agents is their capability to accomplish more complex tasks with less oversight [44]. For instance, LLM Agents could improve efficiency by assisting with compliance [96] or compiling evidence for complex policy documents [60]. In both scenarios, LLM agents inherently make discretionary micro-decisions on which information to include and how to present it, which threatens the legitimacy and autonomy of the system [72]. In the long term, this discretion risks shifting power from human decision makers to unsteerable, opaque systems [16]. Adapting these systems responsibly in public-sector contexts, therefore, requires robust guarantees that the resulting sociotechnical construct continues to comply

Article presented at the Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI 2026).

with these uniquely strict requirements [52]. Such guarantees are enabled by reliable measurement of LLM agents' capabilities and risks [73].

The primary technical means for evaluation of model capabilities is *benchmarking*.[1] Benchmarks are sets of tasks used to evaluate AI systems. While much machine learning research is structured around developing systems that perform well on the most popular benchmarks [87], it is increasingly debated whether they provide much information of real-world use [36]. Many benchmarks measure abstract capabilities on artificial tasks detached from the real-world contexts they aim to emulate [66, 10]. This creates a *benchmark-deployment* validity gap of an unknown size, which makes it difficult for practitioners across fields to make informed decisions. Public-sector deployments contend with a particularly severe form of this problem: it is unclear whether any given benchmark provides useful information, let alone certainty, on the conformity of an agentic system to public-sector requirements.

Statement of Contributions: In this work, we build the foundation for bridging this gap. We do so by combining the literature on digitalisation and bureaucracy [e.g., 13] with theories of socio-technical evaluations of AI systems [74, 84]. This disconnect between current benchmarking practices and public-sector needs motivates our investigation into what constitutes appropriate evaluation for this domain.

To this end, we ask:

1. What criteria must a benchmark meet to guide safe and responsible adoption of LLM agents in the public sector?
2. To what extent do current benchmarks for LLM agents meet these criteria?

To answer the first question, we survey public administration literature to compile a list of operationalisable criteria against which to evaluate benchmarks (see § 3). We arrive at the following list: Benchmarks should be *task-centric*; based on *realistic* tasks and data; concerned with *public sector-relevant* tasks; and report relevant *metrics* including measures of *cost* and *fairness*.

Based on these theory-grounded criteria, we systematically analyse 1304 agentic LLM benchmark papers using a validated LLM-assisted pipeline (see § 4). We show that no existing benchmark meets all criteria, with the biggest gaps existing for public-sector relevance and more comprehensive metrics (§ 5). We suggest ways forward both for the technical design of public-sector agent benchmarks and the holistic sociotechnical evaluation of agent deployment in PSOs.

2 Background

2.1 Agentic Large Language Models

Defining LLM Agents: Kasirzadeh and Gabriel [44] characterise AI agents as "systems that have the ability to perform increasingly complex and impactful goal-directed action across multiple domains, with limited external control". Their framework for understanding agentic LLMs consists of four key dimensions: autonomy, efficacy, goal complexity, and generality:

First, **autonomy** can be understood as the capacity to perform actions without external direction or control [44], with the key characteristic that they can "take multiple consequential actions in rapid succession [...] before a human notices" [15] without "requiring step-by-step guidance" [44]. Second, **efficacy** refers to an agent's ability to directly interact with its environment [44]. Particularly the "access to external tools or services" [15] and "the ability to remove humans from the loop" define agentic systems. Third, **goal complexity** represents agents' ability to form and pursue increasingly sophisticated goals, decomposing complex objectives into subgoals, and reliably adapt, plan, and execute over long time-horizons [44, 16, 15]. Finally, **generality** refers to an AI agent's ability to operate effectively across different roles, contexts, cognitive tasks, or "economically valuable tasks" [15], ranging between "highly specialised AI agents that are focused on specific tasks to general-purpose agents that can shift between different domains" [44].

AI Agents and the Public Sector: Public-sector tasks generally exhibit several features that make them amenable to automation: (a) clear documentation, (b) consistent structure, and (c) repetition [73, 13]. Previous scholarship has demonstrated the successes of LLM Agents on such narrowly

[1] Benchmarking is a component of larger oversight procedures; see Mökander et al. [59], Schmitz et al. [73].

defined and clearly documented tasks, like office management [34], manual healthcare delivery [94], and software development [41]. However, what 'success' means in the context of each paper is often poorly defined. We return to this question in § 3.2.

Beyond narrow tasks, agents can act as interfaces and co-pilots. For instance, Yun et al. [93] argue that agents can help explain complex policies to citizens, and Zhu et al. [96, from Meta] argue that LLM agents can assist with compliance. Again, whether these would work in practice requires strong benchmarks. We return to this in the next section.

2.2 Benchmarking

Benchmarks are commonly defined as standardised collections of *datasets* (representing tasks or abilities) with *scoring metrics* [representing performance on the task or ability; 66, 67]. Such benchmarks create a shared framework to compare different methods and measure progress. As such, benchmarks form the backbone of LLM evaluations.

By operationalising an abstract capability (the phenomenon) into concrete problems and performance measures, benchmarks provide objective comparisons across architectures and training regimes [10, 5]. Crucially, a benchmark's *construct validity* – the degree to which it truly measures the intended phenomenon – determines whether high performance genuinely reflects a system's mastery of the underlying ability [5]. For public-sector applications, where decisions may affect citizens' rights and welfare, trusting benchmark scores without ensuring validity can evidently be dangerous [66].

Drawing on psychometric principles, Bean et al. [5] identify several essential features of a good benchmark:

- ✓ **Clear phenomenon definition:** Benchmarks must precisely define the targeted capability, whether "multi-step reasoning", "compliance checking", or "policy summarisation", and, when relevant, decompose it into sub-skills.
- ✓ **Representative tasks:** Task sets should sample from realistic, domain-relevant scenarios (e.g., authentic case documents or citizen queries), rather than rely solely on contrived or synthetic examples.
- ✓ **Appropriate metrics:** Benchmarks should go beyond top-line measures like accuracy. We will discuss this in § 3.2.
- ✓ **Rigorous methodology:** Robust statistical analyses, confidence intervals, significance tests, and contamination checks guard against overfitting to familiar data and ensure that improvements are meaningful rather than noise.

Unfortunately, many widely used benchmarks exhibit gaps. **Synthetic or toy tasks** are common, with nearly half of recent benchmarks augmenting small real-world cores with large volumes of LLM-generated examples, which risks inflated scores based on artefactual regularities [5]. **Limited ecological validity** is another issue, as only about 30% of benchmarks include evaluations in realistic simulators or field-like conditions, meaning that models may fail when confronted with the messiness of real public-sector workflows [53]. **Narrow metrics** further constrain interpretability, since over 80 % of benchmarks report only exact-match accuracy, omitting cost and other metrics critical for governmental use cases [43]. Finally, **sparse statistical rigour** undermines reliability, with fewer than 20% of studies conducting statistical tests of differences, making it difficult to determine whether reported gains reflect genuine improvement or random fluctuation [55].

For public institutions, these deficiencies translate into *deployment risk*: an agent that "succeeds" on a simplistic benchmark may flounder when processing ambiguous citizen correspondence, exhibit brittleness to phrasing variations, or disproportionately disadvantage vulnerable groups. To mitigate these risks, future benchmarks for LLM-based agents in public-sector contexts should adhere to the construct-validity framework, including defining phenomena precisely, sampling tasks representatively, reporting multi-dimensional metrics, and applying rigorous statistical methods [5, 87]. Only by "measuring what matters" can policymakers gain the clear-eyed insights needed to deploy LLM agents responsibly in service of the public good.

What explains the disconnect between public-sector requirements and benchmarking practice? We hypothesise that it is owed to the design of these benchmarks, which is often driven by technical "paths of least resistance", conventions in computer science, and data availability, rather than their

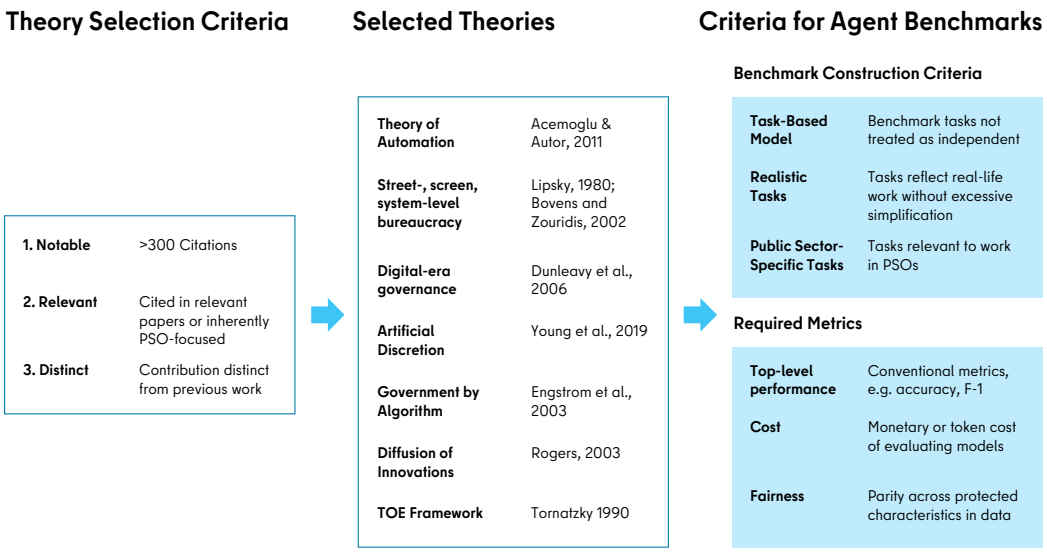


Figure 1: Overview of process for formulation of public sector-relevant agent benchmark criteria.

usefulness for real-world applications [87]. Benchmarks provide static measurements of the capacity of models to perform certain tasks that neither meaningfully advance nor test these hypotheses relevant to real-world use [36].

3 Criteria for a Good Public-Sector Benchmark

In this section, we concretise the notion of a "good" public-sector agent benchmark in the form of six criteria. As established above, the increasing benchmark-deployment validity gap is casting attention on the ecological validity, or real-world representativeness, of LLM and agent benchmarks. While the problems with existing benchmarks are well-mapped [27], there are few concrete characterisations of benchmarks that remedy these issues. As a result, the usefulness of benchmarks for tasks beyond progress measurement, such as for industry practitioners deciding on models, is often limited [36].

Accordingly, this section moves in the other direction. It derives a set of criteria for a "good" public-sector agent benchmark from theoretical work across public administration, theories of automation, and technical benchmarking literature. As posed in RQ1, we aim to find criteria that make benchmarks theoretically and practically meaningful for the public sector. In describing these ideal-type criteria, we do not consider the effort required or even the technical feasibility of such a benchmark.

Public-admininstration theories, which are often descriptive or even normative, should not be treated as blueprints to be reproduced or as claims to be empirically validated through benchmarking. Rather, their use in this exercise is that they produce intermediate abstractions which plausibly generalize across public-sector contexts. Our objective is to make use of these abstractions to strengthen the context-specific *construct validity* of agent benchmarks. As such, we do not aim to set an upper bound on benchmark quality – which is likely to remain case- and deployment-specific – but to raise the floor of what qualifies as informative benchmarks in a public-sector context.

To select relevant social scientific theories, we adopted a purposeful sampling approach [63], selecting for theories that are (i) well-established and notable, (ii) salient in research on public-sector AI use, and (iii) robust and distinct. This resulted in a list of seven theoretical lenses, which we synthesised into six benchmark criteria, grouped into two categories: requirements for the structure of a benchmark, and requirements for the quantitative metrics the benchmark reports. The process is summarised in Fig. 1 and detailed in the appendix, and the remainder of this section introduces and justifies these criteria.

3.1 Benchmark Construction

Task-based Model: The most obvious link between the studied theories is the conceptual deconstruction of organisations, systems, and processes into their constituent *tasks*. Tasks are often defined as "the smallest unit of activity with a meaningful outcome" [e.g., by the widely-used O*NET database; 21]. In automation theory, for example, Acemoglu and Autor [1] describe tasks as the "fundamental unit of work". Bovens and Zouridis [9] describe the incremental task-level digitisation of public-sector processes as the vector by which the transition to screen- and system-level bureaucracy occurs, e.g., the entry of forms by public officials. The digital-era governance framework rests deeply on modularity, both for its "do once" design principle and the end-to-end digitalisation of processes [24]. Rogers et al. [69] does not introduce a task-based lens to the diffusion of innovations, but does refer to this same modularity as an enabler of the iterative processes of innovation and organisation adaption that enable diffusion. Literature on artificial discretion and decision-making in PSOs frequently isolates discretionary and administrative tasks [92, 75]. Practical work applying these theories to AI, including in the public sector, also makes use of this decomposition [77, 37, 31].

The value of these theories is in the re-aggregation of these tasks into superstructures. Acemoglu and Autor [1] aggregate tasks into *jobs* to enable task-based measurement of automation exposure. Both the concepts of screen- and system-level bureaucracy [9] and digital-era governance [23] aggregate tasks into *processes*. This enables (i) the introduction of dependence relationships between tasks, where the outputs of one task are required for subsequent ones, and (ii) the modular reuse and integration of tasks and their outputs across workflows.

A good benchmark should therefore:

- Evaluate a model or system on individually meaningful tasks or units.
- Aggregate these tasks into structures, such as processes, which create dependence relationships between them.

An example of a good process-based paper is AutoPenBench [30], which studies penetration testing through five well-defined, interconnected stages from discovery through to flag capturing.

Realistic Tasks: Each studied theory presupposes, frequently implicitly, that its composite tasks are actual pieces of work that are performed in its respective context. Variations of this requirement are readily apparent: diffusion of innovations [69] defines an innovation's relative advantage as its ability to fulfil real-world needs; automation theory requires tasks to be value-producing [1]; literature on AI adoption and government by algorithm categorizes AI use cases in terms of concrete previously human-performed tasks [26, 19, 61].

As several sources show, this assumption is frequently broken by benchmarks which, for example, use machine-generated or isolated questions, mock or simplify environments, or ignore real-life contextual complexity [36, 74].

Beyond the realism of individual tasks, studied theories presuppose that the *collective* of all tasks is itself a realistic representation of the unit being studied, such as an occupation or a process. Two variations of this emerge. First, the distribution of tasks being studied must reflect the real-life distribution of tasks. For example, work on algorithmic discretion frequently highlights that discretionary decision-making itself requires only a small fraction of a public servant's time, compared to administrative and coordinative tasks [13]. Second, the totality of tasks must indeed constitute the entire aggregate unit – with, e.g. no coordinative, translational, or administrative steps skipped in the reconstruction. This is prominent in literature reviewing past failures of algorithmic governance, such as the Dutch benefits or Michigan MIDAS cases, which frequently point to an overemphasis on the decision-making steps of processes over the preparatory and subsequent ones [64, 25].

A good benchmark should therefore:

- Ensure the internal realism of each evaluated task, such as the context, tools, required output, and constraints.
- Ensure the totality of tasks tested captures the entirety of the phenomenon being evaluated in a representative distribution.

A good example of a realistic benchmark is Chen et al. [18], which studies end-to-end scientific data analysis. It includes real, expert validated tasks.

Public Sector-Specific Tasks: The studied theories reflect the unique conditions of PSOs, which are overwhelmingly Weberian bureaucracies [28]. Though PSOs can be structurally similar to organisations in the private sector and perform comparable tasks [65], they work under exceptionally strict requirements of process justice and documentation [32] upheld, for example, by strong traditions of administrative law encoding public values [12]. These constraints can produce a meaningfully different "jagged frontier" of AI performance [20], even though the well-documented processes and strong regularization may in theory create *more* favourable conditions for agentic automation and oversight [73]. As such, the complex interactions between system-level rules and street-level bureaucracy create unique conditions which require explicit study [51, 3].

A good benchmark should therefore measure performance on tasks that either specifically occur in the public sector, or underlie the same requirements. A good example is AgentsCourt [38], which measures judicial decision-making.

3.2 Reporting relevant metrics

Useful benchmarks require useful metrics. Metrics define *what* agents are evaluated on [5]. While all benchmarks report at least one top-level performance metric (e.g., 'Success rate' or 'Accuracy'), this is insufficient for making decisions about whether to implement an agent to solve a specific task [54]. All public-sector applications must meet statutory requirements set by the law [83] as well as expectations from citizens and politicians [32]. As a minimum to meet these requirements, we argue that PSO-relevant benchmarks must also report sector-relevant metrics of the following: **Cost** and **Fairness**. We argue for these below:

Cost: The first additional metric to define is how expensive an agent system is to run, i.e., the *cost*. As shown by Kapoor et al. [43], the cost of different agent systems can differ by orders of magnitude. The cost of an agent system has two components: 1) the cost per token, and 2) the number of tokens required per task. The cost per token principally depends on the size of the underlying LLM [70]. Crucially, this represents a trade-off: larger models are often more capable and can thus solve a given task in fewer tokens [89]. This trade-off can be analysed by showing the *Pareto curve* of cost versus performance.

Cost is an essential component for public-sector decision-making, since it directly informs whether the agent system would even be an economically viable option for a given task [independent of considerations of accountability or legitimacy; cf. 72].

Fairness: Another metric type – which is much more challenging to define – is *fairness* [83]. At its core, fairness in AI systems is about ensuring that different groups (e.g., people with different genders, ages, or ethnicities) are treated fairly. Of course, this definition strongly depends on how we define 'fairly', which is a question that has puzzled philosophers for centuries [8]. The concrete definition of fairness depends on the context of not just the AI system, but the socio-technical system in which it is embedded [74]. Sometimes, fairness should be operationalised as equal performance between different groups [e.g., in the case of resume classification; 35], whereas in other scenarios it is more important the AI system achieves some threshold of performance for all groups [e.g., in medical diagnostics; 58]. See Mitchell et al. [56] for an overview of fairness metrics.

Regardless of the thorniness of its definition, measuring fairness is essential for public-sector systems. Fairness and accountability are fundamental values across many public sectors [42, 81]. All applications will have context-specific statutory requirements and norms, where fairness metrics play an important role in human evaluation [83]. Having a norm of reporting context-specific fairness metrics is a necessary step for moving towards organisational accountability [57]. While it is beyond the scope of this paper to evaluate if benchmarks use *appropriate* fairness metrics given the ethical and legal context, we will evaluate *whether* benchmarks use fairness metrics.

An example of a paper measuring fairness is *MAPS* [39], which reports agent performance across 11 different languages, which enables analysis of disparities and challenges.

Other metrics: Fairness and cost are not the only relevant additional metrics to report. For instance, Kapoor et al. [43] mentions *robustness* (in the sense of robustness to distribution shifts) as important for agent applications. Another kind of robustness is *adversarial robustness* [14], which is whether small perturbation to the input can lead to different decisions. Systems with poor adversarial

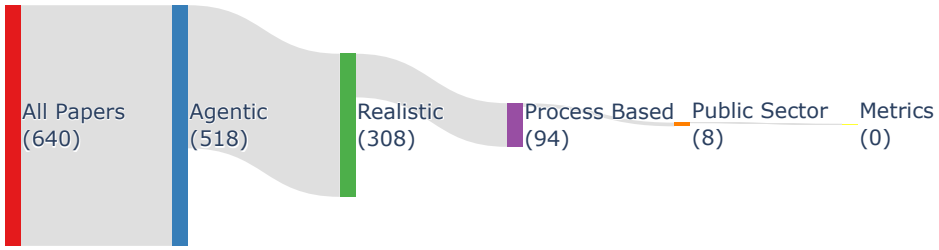


Figure 2: Flow of how well the surveyed papers meet our criteria (see § 3). No single paper meets all of the criteria, with the most challenging categories being public-sector relevance and relevant metrics.

robustness will be prone to attacks and be gameable, which can hurt individual fairness by introducing shortcuts.

This requires carefully considering which changes can affect the agent once deployed. Since LLMs have inherent randomness, robustness is important to measure [6, 45]. However, in preliminary experiments, we found high variation in definitions of robustness, which made it infeasible to include specific guidance.

Furthermore, each specific public-sector use case will have specific socio-technical constraints that should be accounted for [74]. The two metrics we focus on in this paper ('cost' and 'fairness' metrics) should therefore be seen as a starting point rather than a comprehensive checklist.

4 Methods

Here, we describe how we survey the literature to construct an overview of the state of public-sector relevant agent benchmarks. We implement an iterative LLM-assisted distant reading methodology [40]. First, we conduct a keyword search on ArXiv as well as various machine learning conferences (ICLR, NeurIPS, and ACL Anthologies) using the keywords "LLM", "Benchmark", and "Agent". We then deduplicate the papers using fuzzy matching on the titles using a threshold of 90%, which removes 29 duplicates [4]. This results in 1312 unique papers. We download the full text of all these papers from ArXiv or the relevant proceedings. On a high level, we iteratively refine the classification of these papers using a combination of LLM and expert labelling.

4.1 LLM labelling and operationalisation

The principal step of our iterative flow is to operationalise the criteria discussed in § 3 in a way that can be processed by an LLM. Here, we use zero-shot structured data extraction [17], i.e., defining a structured schema (using JSON), which constrains the outputs of the LLM. Crucially, this allows us to define the targets (i.e., the criteria) in natural language.

To increase performance, we use a chain-of-thought-inspired workflow [86]. This means that before the LLM outputs a label for the criteria, it writes a brief analysis which provides reasoning traces that we can use to qualitatively assess alignment [11]. The exact structure of the data model can be found in the accompanying repository, which contains the prompts and schemas used for structured extraction[2]. To create a fully specified response set, we allow each criterion to be labelled as 'Yes', 'No', or 'Unknown' [33]. Note, that in the first iteration, we employ naive descriptions and system prompts to kick off the iterative refinement.

We use the open weights `gpt-oss-120b` for the screening with respect to whether the paper implements a benchmark and adheres to our definition of agenticness (see § 2.1). Since the model is open-weights it increases reproducibility relative to closed models [68]. We use a temperature of 0 to ensure reproducibility further.

[2] `https://github.com/jhrystrom/public-sector-agent-benchmarks`

Table 1: Krippendorf's α and positive rate for LLMs and annotators

		Process Based	Realistic	Public Sector	Cost	Fairness
α_K	Human-Human	62.6%	73.3%	73.2%	64.4%	47.8%
	Human-LLM	29.2%	57.7%	79.3%	100.0%	78.5%
Rate	Human	8.8%	46.9%	23.5%	8.1%	8.8%
	LLM	27.3%	63.6%	15.2%	6.1%	9.1%

4.2 Expert labelling

To ensure the reliability of the LLM labelling, we manually annotate a subset of papers. We select the subset based on the first iteration of the LLM labelling such that we have at least two papers per category (e.g., two public sector/non-public sector, measures cost/doesn't measure cost, realistic/not realistic etc.) This results in a total of 54 sampled papers. We equally divide the sampled papers between the authors such that each paper has two assigned labellers. This allows us to calculate inter-rater reliability on a per-attribute basis [46] as well as increase the reliability of the labelled data by selecting only items with agreement [33].

We label the papers based on a survey with questions corresponding to each criteria as well as two screening attributes: whether the paper contributes a novel benchmark and whether the paper measures an agent as per our definition (see § 2.1). The full survey can be found in the accompanying repository[3]. We first pilot the survey to ensure we have rough agreement on the constructs within the labelling group.

The author group went through three iterations of refining definitions before running the final LLM-assisted analysis. At the end of each iteration, we manually analysed disagreements and did conflict resolution loosely following the steps outlined in Oortwijn et al. [62].

This dataset allows us to design a prompt that accurately captures the concepts. Our initial prompt was exactly the instructions used for annotation. Following Guerdan et al. [33], we design a prompt that meets two criteria: high alignment with expert labellers and similar selection rates. This ensures that we capture both similar judgements and similar rate statistics. We iterate the prompt by adding and modifying inclusion and exclusion criteria until we reach sufficient alignment.

We show the inter-rater reliability statistics in Table 1. We report both Krippendorf's alpha (α_K) and the *positive rate*, i.e., the fraction of papers rated as belonging to the given category. Conventionally, an $\alpha_K > 60\%$ is seen as moderate reliability [cf. 48]. However, one should be careful about dogmatic interpretations [29]—especially with the complex constructs in this paper [88]. Still, the human-LLM reliability is generally high (with the exception of process-based). For the rates, it's important to note that the LLM has generally higher rates than the humans.

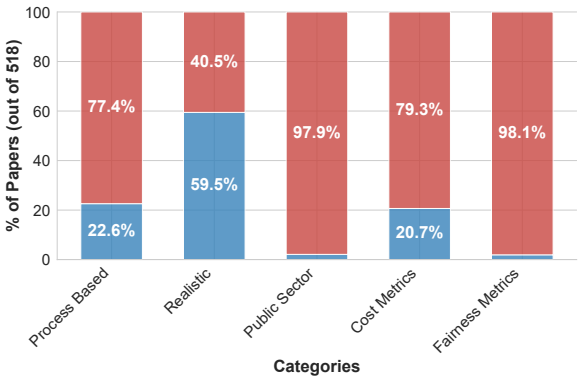
5 State of the Literature

In this section, we analyse the results of our LLM-assisted literature review. First, we analyse the high-level patterns our review has uncovered. Then, we carefully analyse the few papers that best meet our criteria.

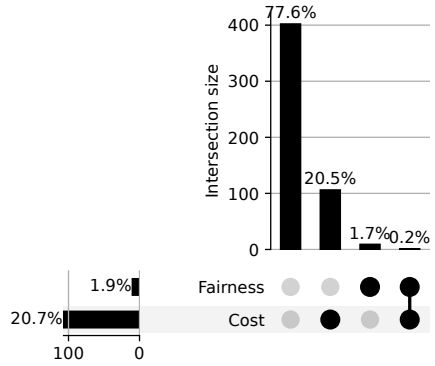
High-level pattern: No complete benchmark exists We show the overall flow of our criteria in Fig. 2. There are several things to note about the state of PSO-relevant agent benchmarks. The first and most important finding is that no single benchmark meets all of our criteria.

As the figure illustrates, the two main limiting factors are reporting all relevant metrics and being relevant to the public sector. Fig. 3b shows that only one paper [85] reports both cost and fairness — regardless of being public-sector relevant. 77.6% of the papers report neither fairness nor cost (see § 3.2). Of these, fairness is the least reported (1.9% of papers) with cost being reported by 20.7% (see also Fig. 3a). As argued in § 3.2, reporting these metrics is important for public-sector decision-makers to make informed decisions about the real-world usability of the benchmarks. We

[3] `https://github.com/jhrystrom/public-sector-agent-benchmarks`



(a) Unconditional distribution of the different criteria. The rarest criteria are public sector relevance and fairness.



(b) Distribution of reported metrics. The rarest metric is fairness, and 77.6% of papers report none of our metrics.

will discuss the specific cases of fairness and public-sector relevant benchmarks in the following sections.

Fairness is not integrated: As evident from Fig. 3b, only 1.9% of papers report fairness metrics. Most of these papers are exclusively focused on fairness evaluation. For instance, Tur et al. [79],Zheng et al. [95], and Hofman et al. [39] all explicitly evaluate safety and alignment, while Li et al. [50] evaluates algorithmic bias. Only Schmidgall et al. [71] and Shen et al. [76] integrate fairness metrics in non-fairness-focused benchmarks.

As previously argued, measuring fairness is an essential (but fraught) requirement for deploying agentic solutions in the public sector (see § 3.2). Unforeseen failures are the norm, not the exception. It is thus imperative that future benchmarks incorporate fairness measures regardless of their focus.

Few public-sector-specific benchmarks: Finally, a core limitation is that there are few public-sector relevant benchmarks. As seen in Fig. 2, there are only eight process-based benchmarks uniquely relevant for the public sector per our definitions (see § 3). These benchmarks focus on judicial administration [38, 76], climate finance [80], logistics [78, 47], and security [91].

While the above benchmarks cover a broad range of use cases, they are far from spanning the full breadth of administration tasks within public-sector organizations. Ideally,a plethora of public-sector relevant benchmarks should be available to inform implementation efforts. While the lack of existing benchmarks presents a challenge, it also provides an opportunity for researchers to develop these benchmarks.

This is not to say that benchmarks must be specific to the public sector to have informative value. Benchmarks like WorkArena [22] and FOFO [format following; 90] , which tests the ability of language models to produce text of specific formats, measure important abstract capabilities that are relevant in the public sector. Still, the stringent requirements and process-based flow in public-sector organisations might be particularly challenging for LLM-based agents; we will only know when we start measuring them.

6 Conclusion

The safe adoption of LLM-based agents in public-sector institutions is a sociotechnical challenge spanning technology and organisation [52]. A precondition for its success is robust, contextually valid technical evaluation of these systems against the public sector's unique requirements [72]. We derive here that candidate LLM agent benchmarks must therefore, at a minimum, live up to a set of criteria to be relevant for the public sector. We propose these criteria as a *signal-strength rubric*: the more criteria a benchmark meets, the more informative it is for public-sector adoption decisions, and gaps indicate where benchmark scores provide weak evidence about real-world deployment. Concretely, benchmarks should be *process-based* (measuring *interdependent subtasks* that reflect how work unfolds in practice), *realistic* (measuring *real tasks with realistic data*), focused on *public sector-*

specific tasks (i.e., tasks actually performed in public-sector organisations), and report public-sector relevant *metrics*, including *cost* and *fairness* alongside other context-dependent measures.

Using an LLM-assisted literature analysis technique, we survey more than 1,300 papers against this rubric. The results are sobering. No single benchmark meets all of the criteria. The largest gaps concern public-sector specificity and comprehensive metric reporting: only one benchmark reports both cost and fairness (with fairness rarely integrated), and there are few public-sector-specific benchmarks. These gaps risk encouraging overconfidence in agents on the basis of benchmark results that are not well-matched to public-sector constraints and trade-offs.

While our findings highlight vast public-sector gaps in the current benchmark literature, our findings also provide a call to action. We call on the academic community to: (1) develop public-sector-specific benchmarks, (2) integrate more relevant metrics, including fairness, and (3) prioritise ecological validity in benchmark design. In parallel, public-sector organisations should assess any benchmarks they rely on against these criteria, and treat missing dimensions as limits on applicability rather than as reassurance. If this call is heeded, public-sector organisations around the world will be in a better position to responsibly integrate LLM agents for the benefit of all citizens.

Limitations

Despite our extensive stratified annotation strategy, it is challenging to ensure coverage across all categories. Since few papers live up to, e.g., public-sector relevance, we might not cover all types of papers in our manual annotations. However, since we carefully match the rates and agreement, we are confident in our overall conclusions.

As noted in § 4, we have imperfect agreement between the LLM classifications and human annotations. However, since the annotators are generally more conservative than the LLM (see Table 1) and our core claims are about the non-existence of complete benchmarks, we are nevertheless confident in the directionality of our conclusions.

Acknowledgments and Disclosure of Funding

We thank the members of the Digital State Forum for their helpful discussion. JR was supported by the Engineering and Physical Sciences Research Council [Grant Number EP/W524311/1]. JB was supported by the Federal Ministry of Research, Technology, and Space of Germany [Grant Number 16DII131]. CS and JB acknowledge funding by the Hertie School's program "AI and Data Science for the Public Sector", funded by the Dieter Schwartz Foundation. KK was supported by the Clarendon Fund.

References

[1] Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings*. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

[2] Acemoglu, D. and Robinson, J. A. (2019). *The Narrow Corridor: States, Societies and the Fate of Liberty*. Viking, London.

[3] Alkhatib, A. and Bernstein, M. (2019). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk. ACM.

[4] Bachmann, M., Layday, Thomasryde, Kokkinou, G., Schreiner, H., Fihl-Pearson, J., Dheeraj, Vioshim, Pekkarr, Górny, M., Sherman, M., Le, C., Odidev, Glynn, H, T., Tr-Halfspace, Dotlambda, Renkamp, N., Tang, K., Moine, H. L., Rosin, G., Hess, D., Clauss, C., V, B., and Delfini (2025). Rapidfuzz/RapidFuzz: Release 3.13.0. Zenodo.

[5] Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., Kim, H., Kirk, H. R., Lin, F., Liu, G. K.-M.,

Luettgau, L., Magomere, J., Rystrøm, J., Sotnikova, A., Yang, Y., Zhao, Y., Bibi, A., Bosselut, A., Clark, R., Cohan, A., Foerster, J. N., Gal, Y., Hale, S. A., Raji, I. D., Summerfield, C., Torr, P., Ududec, C., Rocher, L., and Mahdi, A. (2025). Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[6] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models Be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

[7] Bertelli, A. M., Cannas, S., Sharova, K., and Uriarte, I. (2025). Anchoring institutional values in democracy: A systematic review, a constructive critique, and a research agenda. *Public Administration Review*, page puar.70036.

[8] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR.

[9] Bovens, M. and Zouridis, S. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review*, 62(2):174–184.

[10] Bowman, S. R. and Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

[11] Brailas, A. (2025). Artificial intelligence in qualitative research: Beyond outsourcing data analysis to the machine. *Psychology International*, 7(3):78.

[12] Bryson, J. M., Crosby, B. C., and Bloomberg, L. (2014). Public value governance: Moving beyond traditional public administration and the new public management. *Public Administration Review*, 74(4):445–456.

[13] Bullock, J., Young, M. M., and Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, 25(4):491–506.

[14] Buzhinsky, I., Nerinovsky, A., and Tripakis, S. (2023). Metrics and methods for robustness evaluation of neural networks with generative models. *Machine Learning*, 112(10):3977–4012.

[15] Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., and Anderljung, M. (2024). Visibility into AI agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, Rio de Janeiro Brazil. ACM.

[16] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 651–666, New York, NY, USA. Association for Computing Machinery.

[17] Chen, Y.-C., Hsu, P.-C., Hsu, C.-J., and Shiu, D.-s. (2025). Enhancing function-calling capabilities in LLMs: Strategies for prompt formats, data integration, and multilingual translation. In Chen, W., Yang, Y., Kachuee, M., and Fu, X.-Y., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 99–111, Albuquerque, New Mexico. Association for Computational Linguistics.

[18] Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., Dey, V., Xue, M., Baker, F. N., Burns, B., Adu-Ampratwum, D., Huang, X., Ning, X., Gao, S., Su, Y., and Sun, H. (2024). ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*.

[19] Daub, M., D'Emidio, T., Howard, Z., and Ungur, S. (2020). Automation in government: Harnessing technology to transform customer experience. Technical report, McKinsey & Company.

[20] Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality.

[21] Dierdorff, E. C. and Norton, J. J. (2011). Summary of procedures for O*NET task updating and new task generation. *Summary of Procedures for O*NET Task Updating and New Task Generation*.

[22] 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*, pages 11642–11662. PMLR.

[23] Dunleavy, P. (2006). *Digital Era Governance: IT Corporations, the State, and e-Government*. Oxford University Press, Oxford.

[24] Dunleavy, P. and Margetts, H. (2015). Design principles for essentially digital governance. In *American Political Science Association Annual Meeting*, number 111, USA.

[25] Elyounes, D. (2021). "computer says No!": The impact of automation on the discretionary power of public officers. *Vanderbilt Journal of Entertainment & Technology Law*, 23(3):451.

[26] Engstrom, D. F., Ho, D. E., Sharkey, C. M., and Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies.

[27] Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., and Fernandez-Llorca, D. (2025). Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation.

[28] Gerth, H., Weber, M., and Mills, C. W., editors (1997). *From Max Weber: Essays in Sociology*. Routledge Sociology Classics. Routledge, London, new ed., repr edition.

[29] Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-economics*, 33(5):587–606.

[30] Gioacchini, L., Delsanto, A., Drago, I., Mellia, M., Siracusano, G., and Bifulco, R. (2025). AutoPenBench: A vulnerability testing benchmark for generative agents. In Potdar, S., Rojas-Barahona, L., and Montella, S., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1615–1624, Suzhou (China). Association for Computational Linguistics.

[31] Gmyrek, P., Berg, J., and Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. *SSRN Electronic Journal*.

[32] 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.

[33] Guerdan, L., Barocas, S., Holstein, K., Wallach, H., Wu, Z. S., and Chouldechova, A. (2025). Validating LLM-as-a-judge systems in the absence of gold labels.

[34] 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*.

[35] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[36] Hardy, A., Reuel, A., Jafari Meimandi, K., Soder, L., Griffith, A., Asmar, D. M., Koyejo, S., Bernstein, M. S., and Kochenderfer, M. J. (2025). More than marketing? On the information value of AI benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1032–1047, Cagliari Italy. ACM.

[37] Hashem, Y., Bright, J., and Chakraborty, S. (2025). Mapping the potential: Generative AI and public sector work.

[38] He, Z., Cao, P., Wang, C., Jin, Z., Chen, Y., Xu, J., Li, H., Liu, K., and Zhao, J. (2024). AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.

[39] Hofman, O., Rachmil, O., Bose, S., Pahuja, V., Brokman, J., Shimizu, T., Starostina, T., Marchisio, K., Goldfarb-Tarrant, S., and Vainshtein, R. (2025). MAPS: A multilingual benchmark for global agent performance and security.

[40] Hsu, C.-C., Bransom, E., Sparks, J., Kuehl, B., Tan, C., Wadden, D., Wang, L., and Naik, A. (2024). CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 118–132, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

[41] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. (2023). SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

[42] Jørgensen, T. B. (2007). Public values, their nature, stability and change. the case of Denmark. *Public Administration Quarterly*, 30(4):365–398.

[43] Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. (2025). AI agents that matter. *Transactions on Machine Learning Research*.

[44] Kasirzadeh, A. and Gabriel, I. (2025). Characterizing AI Agents for Alignment and Governance.

[45] Khouja, J., Korgul, K., Hellsten, S., Yang, L., Neacsu, V., Mayne, H., Kearns, R., Bean, A., and Mahdi, A. (2025). LINGOLY-TOO: Disentangling reasoning from knowledge with templatised orthographic obfuscation.

[46] Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

[47] Lai, S., Ning, Y., Yuan, Z., Chen, Z., and Liu, H. (2025). USTBench: Benchmarking and dissecting spatiotemporal reasoning of LLMs as urban agents.

[48] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

[49] Laux, J., Wachter, S., and Mittelstadt, B. (2024). Trustworthy artificial intelligence and the european union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32.

[50] Li, H., Ma, M. D., Huang, J.-t., Weng, Z., Wang, W., and Zhao, J. (2025). BIASINSPECTOR: Detecting bias in structured data through LLM agents.

[51] Lipsky, M. (2010). *Street-Level Bureaucracy, 30th Ann. Ed.: Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

[52] Lowe, R., Edelman, J., Zhi-Xuan, T., Klingefjord, O., Hain, E., Wang, V., Sarkar, A., Bakker, M. A., Barez, F., Franklin, M., Haupt, A., Heitzig, J., Holliday, W. H., Jara-Ettinger, J., Kasirzadeh, A., Kearns, R. O., Kirkpatrick, J. R., Koh, A., Lehman, J., Levine, S., Revel, M., and Vendrov, I. (2025). Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

[53] McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., and Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*, pages 1–18.

[54] Meimandi, K. J., Aránguiz-Dias, G., Kim, G. R., Saadeddin, L., and Kochenderfer, M. J. (2025). The measurement imbalance in agentic AI evaluation undermines industry productivity claims.

[55] Miller, E. (2024). Adding error bars to evals: A statistical approach to language model evaluations.

[56] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

[57] Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507.

[58] Mittelstadt, B., Wachter, S., and Russell, C. (2024). The unfairness of fair machine learning: Leveling down and strict egalitarianism by default. *Michigan Technology Law Review*, 30(1).

[59] Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115.

[60] Musumeci, E., Brienza, M., Suriani, V., Nardi, D., and Bloisi, D. D. (2024). LLM based multi-agent generation of semi-structured documents from semantic templates in the public administration domain. In Degen, H. and Ntoa, S., editors, *Artificial Intelligence in HCI*, pages 98–117, Cham. Springer Nature Switzerland.

[61] 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.

[62] Oortwijn, Y., Ossenkoppele, T., and Betti, A. (2021). Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In Belz, A., Agarwal, S., Graham, Y., Reiter, E., and Shimorina, A., editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online. Association for Computational Linguistics.

[63] Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, 42(5):533–544.

[64] 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.

[65] Rainey, H. G. and Bozeman, B. (2000). Comparing public and private organizations: Empirical research and the power of the a priori. *Journal of Public Administration Research and Theory*, 10(2):447–470.

[66] Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. (2021). AI and the everything in the whole wide world benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

[67] Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. (2024). BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[68] Rogers, A. (2023). Closed AI models make bad baselines.

[69] Rogers, E. M., Singhal, A., and Quinlan, M. M. (2014). Diffusion of innovations. In *An Integrated Approach to Communication Theory and Research*, pages 432–448. Routledge.

[70] Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. (2023). From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9.

[71] Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., and Moor, M. (2025). AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments.

[72] Schmitz, C. and Bryson, J. (2025). A moral agency framework for legitimate integration of AI in bureaucracies.

[73] Schmitz, C., Rystrøm, 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., editors, *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pages 298–308, Vienna, Austria. Association for Computational Linguistics.

[74] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA. ACM.

[75] Selten, F., Robeer, M., and Grimmelikhuijsen, S. (2023). 'just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2):263–278.

[76] Shen, J., Xu, J., Hu, H., Lin, L., Zheng, F., Ma, G., Meng, F., Zhou, J., and Han, W. (2025). A law reasoning benchmark for LLM with tree-organized structures including factum probandum, evidence and experiences.

[77] 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.

[78] Sun, S., Zhao, L., Deng, M., and Fu, X. (2025). VTS-LLM: Domain-adaptive LLM agent for enhancing awareness in vessel traffic services through natural language.

[79] Tur, A. D., Meade, N., Lù, X. H., Zambrano, A., Patel, A., Durmus, E., Gella, S., Stanczak, K., and Reddy, S. (2025). SafeArena: Evaluating the safety of autonomous web agents. In *Forty-Second International Conference on Machine Learning*.

[80] Vaghefi, S. A., Hachcham, A., Grasso, V., Manicus, J., Msemo, N., Senni, C. C., and Leippold, M. (2025). AI for climate finance: Agentic retrieval and multi-step reasoning for early warning system investments.

[81] Van Der Wal, Z., De Graaf, G., and Lasthuizen, K. (2008). What's valued most? Similarities and differences between the organizational values of the public and private sector. *Public Administration*, 86(2):465–482.

[82] Van der Wal, Z., Nabatchi, T., and de Graaf, G. (2015). From galaxies to universe: A cross-disciplinary review and analysis of public values publications from 1969 to 2012. *The American Review of Public Administration*, 45(1):13–28.

[83] Wachter, S., Mittelstadt, B., and Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41:105567.

[84] Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N. J., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., and Jacobs, A. Z. (2025). Position: Evaluating generative AI systems is a social science measurement challenge. In *Forty-Second International Conference on Machine Learning Position Paper Track*.

[85] Wang, P., Tao, R., Chen, Q., Hu, M., and Qin, L. (2025). X-WebAgentBench: A multilingual interactive web benchmark for evaluating global agentic system. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19320–19335, Vienna, Austria. Association for Computational Linguistics.

[86] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

[87] Weidinger, L., Raji, D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Kapoor, S., Ganguli, D., Koyejo, S., and Isaac, W. (2025). Toward an evaluation science for generative AI systems.

[88] Wong, K., Paritosh, P., and Aroyo, L. (2021). Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

[89] Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. (2024). Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *The Thirteenth International Conference on Learning Representations*.

[90] Xia, C., Xing, C., Du, J., Yang, X., Feng, Y., Xu, R., Yin, W., and Xiong, C. (2024). FOFO: A benchmark to evaluate LLMs' format-following capability. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand. Association for Computational Linguistics.

[91] Xiang, Z., Zheng, L., Li, Y., Hong, J., Li, Q., Xie, H., Zhang, J., Xiong, Z., Xie, C., Yang, C., Song, D., and Li, B. (2025). GuardAgent: Safeguard LLM agents by a guard agent via knowledge-enabled reasoning.

[92] Young, M. M., Bullock, J. B., and Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance*, 2(4):301–313.

[93] 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.

[94] Zhao, Z., Yue, X., Xie, J., Fang, C., Shao, Z., and Guo, S. (2025). A dual-agent collaboration framework based on LLMs for nursing robots to perform bimanual coordination tasks. *IEEE Robotics and Automation Letters*, 10(3):2942–2949.

[95] Zheng, J., Wang, H., Zhang, A., Nguyen, T. D., Sun, J., and Chua, T.-S. (2024). ALI-agent: Assessing LLMs' alignment with human values via agent-based evaluation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[96] Zhu, S., Fang, C., Larson, D., Pochareddy, N. R., Rao, R., Zeng, S., Peng, Y., Summer, W., Goncalves, A., Pudota, A., and Robert, H. (2025). Compliance brain assistant: Conversational agentic AI for assisting compliance tasks in enterprise environments.