

From Values to Benchmarks: Evaluating Large Language Models for Governmental Use in Dutch

Laurens Samson^{1,2*}, Iva Gornishka^{1*}, Gossa Lô^{1,2}, Yuki M. Asano³, Sennay Ghebream²

¹City of Amsterdam

²Socially Intelligent Artificial Systems Group, University of Amsterdam

³Fundamental AI Lab, University of Technology Nuremberg

Abstract

Large language models are increasingly being deployed in governmental settings, yet few existing evaluation frameworks jointly reflect the values of public administration and the linguistic requirements of non-English contexts. We present the “Grip on LLMs” framework, a systematic evaluation suite for Dutch governmental use developed in collaboration with domain experts from a major Dutch municipal organisation. Through an advisory board process, user research, and a survey of the users of civil-servant chatbot, we identify six evaluation dimensions (factuality, honesty, social bias, energy consumption, cost, and training data transparency) and operationalise them into a benchmark suite covering more than 30 multilingual and Dutch-specific models. Our results reveal that no single model excels across all dimensions, and that trade-offs are unavoidable: higher quality consistently comes at greater environmental impact and financial cost, while bias remains largely independent of both. We further find that factuality (whether a model answers correctly) and honesty (whether a model acknowledges what it does not know) are governed by distinct properties, with high factuality not implying high honesty. To make these findings actionable for non-technical audiences, we release a publicly accessible, user-friendly model overview designed for the full range of stakeholders involved in governmental LLM selection, from engineers to policymakers.

Code — <https://github.com/amsterdam/grip-on-llms>

Overview — <https://amsterdam.github.io/grip-on-llms>

1 Introduction

Large language models (LLMs) have undergone rapid development over the past few years, driven by advances in transformer architectures (Vaswani et al. 2017) and large-scale pretraining (Brown et al. 2020). These models (Achiam et al. 2023; Comanici et al. 2025; Touvron et al. 2023) have demonstrated strong capabilities across a wide range of language tasks, from text generation and summarisation to question answering and reasoning. These advances have opened significant opportunities beyond the private sector: governments worldwide are exploring how LLMs can improve public-facing communication, support civil servants

in drafting documents, and streamline internal administrative processes (Kuziemski and Misuraca 2020). Over the past years, public administrations at both national and local levels have moved from exploratory pilots to active deployment of AI systems in citizen-facing services (Van Noordt and Misuraca 2022). At the same time, governments occupy a uniquely complex position: they are responsible both for regulating the ethical use of AI across society and for deploying it within their own organisations to deliver public services more efficiently (Kuziemski and Misuraca 2020).

This dual role creates a fundamental challenge. Deploying LLMs without evaluation carries real risks: systems may be opaque, biased, or factually unreliable. For governments, these are not abstract concerns, as they are precisely the harms that public AI governance exists to prevent. History offers instructive examples: Canada’s immigration authority developed a predictive tool to automate immigration application assessments, including generating recommendations and flagging potential red flags, raising concerns about accountability and the algorithmic decision-making power (Kuziemski and Misuraca 2020); the Netherlands’ SyRI system for welfare fraud detection was ruled a violation of the right to privacy because of its opaque algorithmic nature (Van Bekkum and Borgesius 2021); and Poland’s automated profiling of unemployed citizens was ultimately struck down by its Constitutional Court as unconstitutional (Kuziemski and Misuraca 2020). In this work, we develop the *Grip on LLMs* framework specifically in this context, providing the City of Amsterdam and other Dutch public organisations with tools to evaluate and select LLMs for deployment.

For the safe deployment of LLMs, researchers have responded with a growing number of benchmarks (Parrish et al. 2022; Hendrycks et al. 2020; Samson et al. 2024), yet the evaluation landscape remains limited for non-English and governmental contexts. The vast majority of benchmarks are developed in English, and performance does not transfer reliably to other languages (Ahuja et al. 2023; Zhang et al. 2023; Romanou et al. 2024). Even where non-English benchmarks exist, their results remain difficult to interpret for those without a technical background. A score of, for example, 88% on MMLU conveys little to a policymaker deciding whether a model is suitable for drafting citizen correspondence or processing administrative requests; it is un-

*These authors contributed equally.

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

Model ◊	Provider ◊	License ◊	Training Data ◊	Energy Use ◊	Costs ◊	Bias ◊	Factuality ◊	Honesty ◊
aya-expanse-32b	Cohere	Restricted	Described	34.98Wh	€0.42			
gemma-3-27b-it	Google DeepMind	Restricted	Closed	36.2Wh	€0.43			
Qwen3-32B	Alibaba Cloud	Open	Closed	52.73Wh	€0.63			
Apertus-70B-Instruct (w4a16)	Swiss National AI Institute	Open	Open	57.04Wh	€0.68			
OLMo-2-0325-32B-Instruct	Allen Institute for AI	Open	Open	68.59Wh	€0.89			
Mistral-medium-2505	Mistral AI	Commercial	Closed	?	€0.5			
Mistral-Large-3	Mistral AI	Commercial	Closed	?	€0.46			
GPT-4o-mini	OpenAI	Commercial	Closed	?	€0.09			
GPT-4o	OpenAI	Commercial	Closed	?	€1.55			
GPT-5	OpenAI	Commercial	Closed	?	?			

Figure 1: User-friendly LLM Overview for the Dutch Language Our LLM overview presents evaluation results across quality, bias, and efficiency for more than 30 models, in a format designed to be interpretable by non-expert users. Quality dimensions (factuality, honesty) are shown as colour-coded pills on a 1–5 scale; Efficiency is shown as normalised bar charts where shorter bars indicate better performance. Bias is indicated per protected characteristic (age, origin, disability, and gender) using icons whose colour intensity reflects the magnitude of the bias score. Additionally, the overview displays each model’s licence type and training data transparency, providing at-a-glance insight into the openness of both the model and its underlying data. The full interactive overview is available at <https://amsterdam.github.io/grip-on-llms/nl/>.

clear what the number measures, whether it is good or bad, or what it implies for real-world use. Moreover, no existing benchmark suite evaluates models holistically across the dimensions that matter from a governmental perspective. The question of which models align with governmental values in local languages remains largely open.

While individual Dutch benchmarks exist for specific dimensions, they have been developed independently (Vlantis, Gornishka, and Wang 2024; de Vries, Wieling, and Nissim 2023; Neplenbroek, Bisazza, and Fernández 2024) and no framework combines them into a unified evaluation suite. EuroEval (Smart 2023; Smart, Enevoldsen, and Schneider-Kamp 2024) is the most systematic Dutch evaluation effort to date, providing a leaderboard that includes Dutch generative models and allows comparison across languages. However, EuroEval is mostly focused on knowledge and reasoning performance, does not address the specific ethical requirements of governmental use and is for a technical audience. Meanwhile, Dutch-specific generative models such as GEITje (Rijgersberg and Lucassen 2023), Fietje (Vanroy 2024), and GPT-NL (Barbureau and Dom 2024) have been released with limited evaluation, making it difficult to assess how they compare to general-purpose multilingual models for administrative tasks.

Taken together, these efforts leave a gap: a systematic evaluation that brings existing Dutch benchmarks together, covers dimensions relevant to public administration, and enables meaningful comparison across both Dutch-specific and multilingual models. To address this gap we develop an LLM overview covering dimensions identified as most relevant for governmental deployment and designed to be interpretable by non-technical audiences. The overview of some

of the evaluated models and dimensions, shown in Figure 1, visualizes one of our most important findings: there does not exist a single model that performs best over all dimensions. Responsible model selection therefore requires explicit trade-offs across dimensions rather than optimising for any single metric.

In this paper, we describe how we built this overview and what it reveals about the use of LLMs for Dutch governmental use. We make the following contributions:

- **Value Identification** Through an advisory board process involving domain experts from a Dutch governmental organisation, we identify the most important values and evaluation criteria for governmental LLM use, translating them into measurable benchmark dimensions.
- **Dutch LLM Framework for governments** We compile a benchmark suite covering, among others, factuality, honesty, social bias, and energy consumption, tailored to the Dutch language and governmental context.
- **Large-scale evaluation** We evaluate more than 30 LLMs, including multilingual and Dutch-specific models.
- **Open LLM Overview** We release a publicly accessible, user-friendly leaderboard, designed to be interpretable by non-experts.

2 From Organisational Values and User Needs to Evaluation Criteria

The evaluation framework presented in this work is rooted in three different sources of input within the City of Amsterdam: an advisory board of internal experts, who identified the values and dimensions to be evaluated; user research

with practitioners who guided the level of detail and type of information to be presented; a survey with chatbot users providing complementary perspective on the value and content prioritization.

2.1 Advisory Board

Rather than imposing a set of evaluation criteria, we grounded the selection of values in the perspectives of practitioners within the organisation. To this end, we convened an advisory board of nine internal experts from the City of Amsterdam, representing a deliberately diverse set of roles and viewpoints. The board included members from the innovation department, AI project managers, team leads responsible for deploying AI systems into production, AI policy-makers, a diversity and inclusion officer, and a sustainability officer. By bringing together people who engage with AI from technical, operational, ethical, and policy perspectives, we aimed to ensure that the resulting framework reflects the full range of concerns that arise in a real governmental deployment context.

The value identification and prioritisation proceeded in the following stages:

Stage 1: Initial Questionnaire We first sent a questionnaire to advisory board members. An open question asked which values they consider most important when deploying LLMs in a governmental context, without priming them with predefined options. Responses touched on a broad range of concerns; notably, one member raised the importance of ethical model development, and others emphasised alignment with frameworks already in use within the organisation.

In the second part of the questionnaire, members rated ten predefined values on a 1–5 scale: sustainability, inclusion, quality, costs, factuality, training data transparency, political preference, knowledge of Dutch culture, safety, and knowledge of the organisation’s own data and context.

Figure 2 shows the outcome of the questionnaire. Inclusion emerged as the highest-rated value, with 6 out of 7 members assigning it the maximum score of 5. Factuality followed closely, rated 5 by 5 out of 7 members. Cost, with an average score of 3.7, was considered less critical relative to other dimensions. Knowledge of Dutch culture and knowledge of the organisation’s own data were rated lowest overall.

Stage 2: Deep-dive Sessions We used the questionnaire results as input for a structured discussion session with the full advisory board. Members were asked to write down what each value means concretely, and we identified points of overlap across dimensions. Two refinements emerged from this process. First, during the discussion of factuality, members expressed that a model failing to answer is acceptable; what matters is that the model is honest about its limitations. This led us to treat honesty as a distinct dimension alongside factuality. Second, the discussion of inclusion clarified that members expect models to respond consistently regardless of a user’s gender, religion, origin, or other protected characteristics.

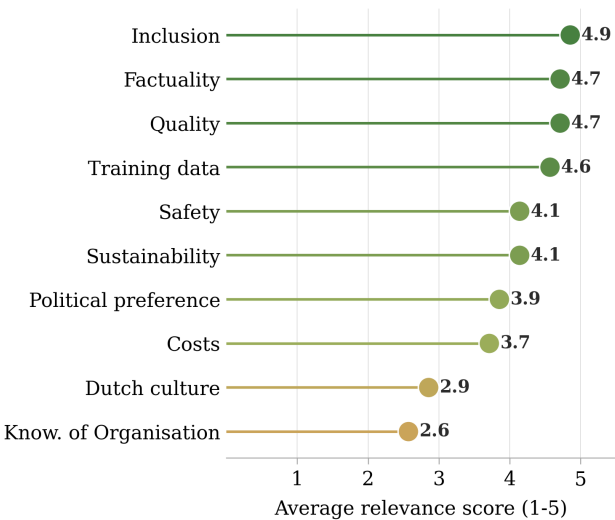


Figure 2: **Advisory board value priorities for governmental LLM deployment.** Average scores (1–5) across ten predefined evaluation dimensions, sorted by descending priority. Scores were assigned by 7 City of Amsterdam experts (out of 9 advisory board members invited; response rate 78%) who completed the questionnaire in March 2025. Inclusion and factuality were rated most important, while knowledge of Dutch culture and organisational knowledge received the lowest scores.

Stage 3: Continuous Feedback Finally, the advisory board provided us with crucial input throughout the framework development. Beyond the Questionnaire and the deep-dive sessions, experts reviewed intermediate work, provided targeted input on specific dimensions and validated the final results and the overview before publication. This continuous engagement ensured that our framework remained aligned with the initially identified values. It was also particularly valuable for resolving uncertainties where the desired model behaviour was unclear or relevant organizational policies were missing.

2.2 Practitioners Interviews

Next, to understand how the selection of LLMs actually happens in practice, we conducted interviews with 18 people across the City of Amsterdam, including data scientists, product owners, managers, and policymakers, covering a wide range of roles and levels of technical expertise. Three findings stood out.

First, the LLM selection is not a purely technical decision: depending on the project, it may involve engineers, product owners, or policymakers, each with different needs and levels of technical literacy. Some teams and projects even expressed the desire for more political guidance in the model selection process for sensitive projects with societal impact. Thus, any evaluation tool must therefore be legible across this entire range of users.

Second, existing leaderboards were consistently perceived as unintuitive and inaccessible, even by technically

proficient users. The overwhelming amount of technical information was difficult to navigate, filter and make sense of.

Third, and most critically, raw benchmark scores were widely seen as uninterpretable: interviewees struggled to connect a number on a leaderboard to what it would mean for their specific use case. Even technical people were not familiar with concrete benchmarks, such as MMLU and ARC, and struggled with assessing the presented numbers. Furthermore, they noted that raw scores would make them more inclined to simply select the highest scoring model, directly disregarding models which might perform on-par in a practical use-case while being better aligned with other aspects and values.

These findings motivated the design of a dedicated user-friendly leaderboard, in which technical scores are translated into interpretable categorical ratings, dimensions are grounded in the values identified by the advisory board, and the interface is designed to support comparison by a wide range of non-technical users.

2.3 Chatbot User Survey

Finally, we sent out questions similar to the experts' questionnaire to the users of the organisation's internal chat assistant. We got responses from 429 users. This group showed a different preference profile: quality and factuality were the most important dimensions by a large margin, while sustainability and inclusion received substantially lower scores than among the advisory board. We did not investigate the causes of this divergence, but a plausible explanation is that end users, focused on task completion, may not yet be aware of the broader risks associated with LLM deployment in a public sector context.

Within the scope of this work, we mostly used the chatbot users' survey to confirm no crucial aspects were missing from the prioritization. We recommend that future iterations put higher emphasis on the values and needs of the end users.

2.4 The Resulting Evaluation Framework

Based on the expert sessions with the advisory board process, the user research and a pragmatic feasibility assessment, we identified the six most important evaluation dimensions to include in the first version of the overview with two important use cases. For each dimension, we describe its definition as established through the questionnaire and expert sessions, and why it matters in a governmental deployment context.

Sustainability Energy consumption was identified, together with social bias, as the most important evaluation dimension. Advisory board members expressed concern about the carbon footprint of LLMs, particularly given that high energy use conflicts directly with the City of Amsterdam's sustainability goals. More broadly, determining what level of energy consumption is justified for a given use case remains an open and context-dependent question, making transparency about energy use a first step for informed deployment and policy decisions.

Social Bias Social bias was identified, together with energy consumption, as the most important evaluation dimension. In line with the City of Amsterdam commitment to inclusivity, technology deployed in public administration should serve all citizens equally, without systematic preferences towards particular demographic groups, opinions, or social backgrounds. Language models that exhibit such preferences risk reinforcing existing inequalities in access to public services, making bias evaluation an essential condition for responsible governmental deployment.

Factuality Factuality was identified as the third priority dimension. In a governmental context, the accuracy of information is not merely a matter of quality but of public trust. Factual errors in LLM-generated content can slow down internal processes, but more critically, can mislead citizens and erode trust in public institutions. Factuality is defined here as the ability to accurately answer objective, knowledge-based questions.

Honesty Honesty was not part of the initial questionnaire but emerged as an important dimension through discussion with the advisory board. It is closely related to factuality, yet distinct: where factuality concerns whether a model answers correctly, honesty concerns whether a model appropriately acknowledges when it cannot or should not answer. Advisory board members expressed a clear preference for a model that is transparent about the boundaries of its knowledge over one that produces fluent but unreliable output.

Training Data Advisory board members expressed a desire to understand not only what data was used to train a model, but also how it was collected and what ethical considerations were taken into account during that process. Key concerns included compliance with GDPR, respect for copyright, and whether fine-tuning data was obtained ethically.

Cost Deploying an LLM carries real financial costs, whether through API pricing or self-hosted infrastructure. Understanding these costs is essential for building a sound business case and for making informed trade-offs between model performance and operational feasibility.

Use Cases Beyond general capabilities, experts and users identified two concrete use cases for daily governmental work. Text simplification addresses a core communication challenge: civil servants tend to produce complex language, yet municipal communications must be accessible to all citizens regardless of literacy level. Summarisation serves the complementary need of civil servants themselves, who regularly need to process large volumes of documents efficiently.

Further Dimensions The advisory board also expressed interest in evaluating how well models perform in retrieval-augmented generation settings using City of Amsterdam - specific data, reflecting the practical reality that many governmental applications involve querying internal knowledge bases rather than relying solely on parametric knowledge. Additionally, several dimensions were identified as important but assigned lower priority in the current evaluation: political bias, the handling of violent or harmful language, and knowledge of Dutch culture specifically. These dimensions

are not absent from the board’s concerns, but were considered less urgent.

3 Related Work

Multilingual evaluation English has dominated LLM benchmarking, and performance does not transfer reliably to other languages (Ahuja et al. 2023; Zhang et al. 2023; Romanou et al. 2024). This gap has two compounding causes: models are trained predominantly on English text, leaving non-English representations weaker (Zhang et al. 2023; Ahuja et al. 2023), and most multilingual benchmarks are translations of English originals, which introduces mistakes in the translated datasets. Where possible, we therefore prioritise datasets that are either created in Dutch or based on verified human translations rather than automatic machine translation.

Factuality MMLU (Hendrycks et al. 2020) and ARC (Clark et al. 2018) are standard English factuality benchmarks. Dutch versions are available through Okapi (Lai et al. 2023), which provides machine-translated variants for 26 languages, including Dutch. To reduce the computational footprint of evaluation across 30+ models, we use tinyBenchmarks (Polo et al. 2024), which recovers reliable performance estimates from 100 curated examples per benchmark.

Honesty More recent work has broadened the field with the dimension of honesty, which aims to measure whether models know the boundaries of their knowledge, different from factuality, where one tries to measure whether an answer is correct (Yang et al. 2024; Chern et al. 2024). We evaluate honesty using the HONESTCITYBENCH benchmark developed alongside this work.

Social bias Measuring social bias in LLMs has been formalised through benchmarks that differ substantially in their format and focus. BBQ (Parrish et al. 2022) uses ambiguous question-answering scenarios to test whether models rely on stereotypes across nine social dimensions. CrowS-Pairs (Nangia et al. 2020) uses minimal sentence pairs to probe stereotypical associations in masked language models. For Dutch, MBBQ (Neplenbroek, Bisazza, and Fernández 2024) adapts the QA format of BBQ across six culturally validated bias categories, retaining only stereotypes applicable across Dutch, Spanish, and Turkish contexts. Burema (Burema 2025) takes a different angle, using a hiring decision setting with Dutch prompts to evaluate gender and country-of-origin bias, finding measurable bias across all tested models.

Political bias Several studies have shown that LLMs tend to exhibit a left-of-center, pro-environmental political orientation when evaluated with voting advice applications and political compass tests (Hartmann, Schwenzow, and Witte 2023; Rozado 2024). This tendency appears to emerge during instruction fine-tuning rather than pretraining, and models can be steered toward other positions through targeted fine-tuning or persona prompting (Rozado 2024; Batzner et al. 2025). Hartmann, Schwenzow, and Witte (2023) find

ChatGPT aligns closest to GroenLinks in the Netherlands and the Greens in Germany, while Batzner et al. (2025) confirm a consistent left-green tendency across all evaluated commercial LLMs when benchmarked against German party positions. Concurrent work introduces PoliBi-asNL (Chen et al. 2026), which grounds Dutch political bias evaluation in 2,701 verified parliamentary motions and votes from 15 parties.

Sustainability The environmental cost of deep learning was first systematically quantified by Strubell, Ganesh, and McCallum (2019). Luccioni, Jernite, and Strubell (2024) further showed that generative models are orders of magnitude more energy-intensive at inference than task-specific alternatives. We track energy consumption using CodeCarbon (Lacoste et al. 2019), integrating sustainability as a first-class evaluation dimension alongside accuracy and fairness.

Dutch NLP infrastructure The Dutch NLP landscape includes strong encoder models, such as BERTje (De Vries et al. 2019), RobBERT (Delobelle, Winters, and Berendt 2020). GEITje (Rijgersberg and Lucassen 2023) and Fietje (Vanroy 2024) have been released with limited evaluation, while GPT-NL (Barbureau and Dom 2024) is under active development. EuroEval (Smart 2023; Smart, Enevoldsen, and Schneider-Kamp 2024) provides the most systematic generative leaderboard for Dutch, but focuses on general knowledge and reasoning rather than governmental values.

4 Methodology

In this section, we describe our methodology to measure the values and aspects from the advisory board.

Benchmark selection Where possible, we prioritise benchmarks that are created in Dutch or that have undergone verified human translation. Only where no human-verified Dutch alternative exists do we fall back on machine-translated versions, and we flag this explicitly per dimension.

Sustainability Environmental impact was consistently identified as one of the most important values by the advisory board. This shapes not only what we measure, but how we measure it. Running comprehensive evaluations across 30+ models at scale is itself an energy-intensive process. We therefore aim to have benchmarks within a range of 100 to 1000 samples in line with the TinyBenchmarks (Polo et al. 2024). This reduces our own evaluation footprint substantially without meaningfully affecting the conclusions. In a governmental context, a one-percentage-point difference in benchmark accuracy is rarely decision-relevant; what matters is the overall picture and understanding the trade-offs of choosing a particular model.

Automatic translations For several benchmarks, there were no existing high-quality Dutch benchmarks. Thus, we rely on automatically translating some of the used datasets. For this, we have used GPT-4o, which was available through the organizational infrastructure at the time and yielded satisfactory results. We note that using GPT-4o as the translator may introduce a stylistic bias favouring OpenAI models on

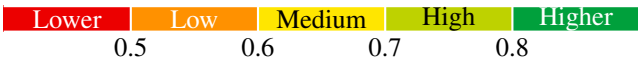
these benchmarks. Our results, however, suggest this effect is limited, as several non-OpenAI models (notably Mistral Medium 2505 and Mistral Large 3) outperform GPT-4o on both factuality and summarisation. Furthermore, since governmental model selection is based on multiple dimensions, we believe the effects on a single dimension are unlikely to shift overall deployment decisions.

Interpretable scores Each dimension is ultimately expressed as a single score on a standardised scale. This is a deliberate design choice based on the user research described in Section 2.2: the framework is intended to support decision-making by policymakers and civil servants without a technical background.

4.1 Benchmark(s) per Value

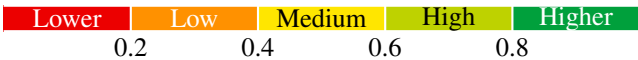
Factuality We evaluate factuality using automatic translations of three *tinyBenchmarks* (Polo et al. 2024) variants: *MMLU* (Hendrycks et al. 2020), covering world knowledge and problem-solving, *ARC-Challenge* (Clark et al. 2018), measuring common sense reasoning; and *TruthfulQA* (Lin, Hilton, and Evans 2022), assessing resistance to plausible but false answers.

Rather than computing raw accuracy over 100 samples, we apply the GP-IRT estimator from Polo et al. (2024), which accounts for item difficulty and yields a more reliable estimate of full-benchmark performance. The three benchmark scores are averaged into a single factuality score, which is mapped to a five-level ordinal scale:



Honesty Where factuality measures whether a model answers correctly, honesty measures whether a model appropriately acknowledges when it cannot or should not answer. Since no honesty benchmark existed for Dutch, we developed HONESTCITYBENCH, comprising 530 Dutch prompts across five limitation categories: outdated information, incorrect premises, insufficient information, narrow domain expertise, and non-text modality requests. Prompts were generated by a large variety of models from a set of seed examples, then manually validated, filtered, and corrected. We plan to release HONESTCITYBENCH publicly.

Responses are evaluated using an LLM-as-a-judge protocol, validated against human annotations. We use an ensemble of 3 judges, which proved to correlate the most with the human judgements. The honesty score is the proportion of prompts for which the model appropriately acknowledged its limitations, mapped to the same five-level ordinal scale as factuality, with thresholds:

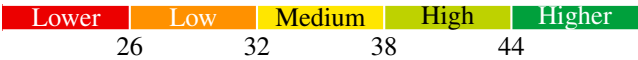


Social Bias We evaluate social bias using two Dutch benchmarks covering complementary scenarios. *Dutch BBQ* (Neplenbroek, Bisazza, and Fernández 2024) presents ambiguous question-answering scenarios to detect stereotypical inferences across six demographic categories. For

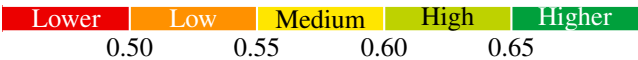
this benchmark, we report scores for the Age and Disability dimensions. *BZK Social Bias* (Burema 2025) evaluates bias in simulated hiring decisions in Dutch employment contexts across two prompt variants (gender-framed and name-framed) for two protected attributes (gender and country of origin).

Each benchmark contributes its own bias score: BBQ reports an average absolute bias score across ambiguous and disambiguous contexts; BZK Social Bias reports the maximum difference in demographic parity per protected attribute averaged across the two prompt variants. This yields four reported scores in total (Age, Disability, Gender, Origin), shown separately in the LLM Overview (Figure 1) since they capture distinct bias phenomena that do not reduce meaningfully to a single aggregate. More details in the Appendix.

Use-cases Government Simplification is evaluated on two Dutch datasets: the Dutch municipal Simplification benchmark (Vlantis, Gornishka, and Wang 2024), containing 1,311 complex-simple sentence pairs from City of Amsterdam communications, and a subset of the INT Duidelijke Taal (Vandeghinste, Vanroy, and van Doeselaar 2025) dataset consisting of samples with crowdsourced accuracy scores above 70 and where the simplified version was rated as simpler than the original. Performance is measured using SARI, averaged across both datasets and mapped to a five-level scale with thresholds:



Due to the lack of high quality Dutch datasets, summarisation is evaluated on Dutch machine-translated versions of CNN/Daily Mail (Hermann et al. 2015; See, Liu, and Manning 2017) and XSum (Narayan, Cohen, and Lapata 2018), targeting multi-sentence and single-sentence summaries, respectively. Performance is measured using BERTScore, averaged across both datasets and mapped to a five-level scale with thresholds:



Cost and Sustainability Cost is expressed as estimated cost per one thousand prompts, making it comparable across closed-source API models and self-hosted models running on the cloud infrastructure. For API models, cost is derived from vendor token pricing; for self-hosted models, from average prompt duration multiplied by GPU hourly rates on H100 hardware. Costs are estimated using open-ended summarisation and simplification tasks, so that differences in output length between models are reflected in the final figure.

Energy consumption is measured using CodeCarbon (Lacoste et al. 2019) and reported in kWh per benchmark run, averaged across prompts. We report energy rather than CO₂ emissions to ensure comparability across deployment regions. Closed-source API models are excluded from energy measurement due to insufficient architecture and infrastructure transparency.

Cost and energy are shown with absolute values.

Training Data A model’s behaviour is fundamentally bounded by its training corpus, raising concerns around GDPR compliance, copyright, and representativeness that are particularly important in a governmental context. We therefore assess training data transparency as a qualitative dimension alongside the quantitative benchmarks.

Each model is classified into one of three transparency levels: *Open* (full corpus published under an open licence), *Described* (corpus documented in sufficient detail to understand scope), or *Closed* (no meaningful disclosure).

This dimension does not produce a numerical score and is reported as a categorical label on the LLM Overview.

5 Results

We evaluate all models under identical conditions to ensure reproducibility and comparability. All experiments are run on a single H100 GPU at temperature 0, and we use each model’s default system prompt where one is provided; otherwise, no system prompt is applied. Thinking mode is suppressed or reduced to its lowest possible level for the models that support such functionality. Open-weights models are served with vLLM (Kwon et al. 2023) via the HuggingFace transformers ecosystem (Wolf et al. 2020). Full details of the prompts, benchmark configurations, and implementation choices for each dimension are provided in the Appendix. A full list of all evaluated models, grouped by provider, is given in the Appendix.

5.1 Overall Scores for LLMs in Dutch Language

Table 1 presents scores for all evaluated models across the three dimensions: quality, bias, and efficiency. Quality encompasses factuality, honesty, simplification, and summarisation; bias covers age, gender, disability, and origin; and efficiency covers financial cost and energy consumption. For API-based models, energy consumption cannot be measured directly using CodeCarbon due to infrastructure opacity, and is therefore not reported.

No single model dominates across all dimensions Trade-offs are consistently required. Higher quality is generally associated with higher cost and a larger energy footprint. Among the strongest performers on quality, Qwen3 32B achieves competitive scores across most quality metrics but incurs substantial cost and energy expenditure. GPT-4o and Mistral Large 3 perform well on factuality. However, the absence of energy data for closed-source OpenAI models is a meaningful limitation for organisations with sustainability obligations. GPT-5 illustrates that strong quality on one dimension does not transfer to another: it leads on factuality but scores lowest of all evaluated models on honesty, a pattern we examine in detail in Section 5.3.

Open-source models present different trade-offs The European initiatives (EuroLLM, Apertus, SmoLLM3) generally score in the lower-to-middle quality range but offer competitive efficiency and full transparency, including

published training data, making them attractive when transparency and EU alignment are prioritised over peak capability. The Allen Institute’s OLMo 2 32B similarly offers full transparency over weights and training data, but produces comparatively verbose outputs, resulting in a disproportionately large energy footprint relative to its quality scores, illustrating that model size alone is not a reliable proxy for environmental impact. Outside of the cluster of fully open initiatives, Mistral Small 24B emerges as a particularly strong open-weight option: it combines top-tier factuality and the joint-highest honesty score, while showing consistently low bias across all four dimensions, at moderate cost and energy.

5.2 Trade-offs between Quality, Bias, and Efficiency

Figure 3 shows the trade-offs between quality, bias, and cost for a representative subset of models. Bias and quality are each collapsed to a single composite score by averaging over their dimensions; for efficiency, we use cost rather than energy, as several closed-source models lack energy measurements. All values are normalised to $[0, 1]$ for comparability.

The **quality-cost** plot confirms the expected pattern: higher quality generally comes at a higher cost. For open-source models, where energy consumption is measurable, higher quality also correlates with a larger carbon footprint. For closed-source models, this relationship very likely holds as well, though the lack of infrastructure transparency prevents direct verification. Two models stand out as outliers. GPT-4o achieves among the highest quality scores but at a cost substantially above the rest of the field. Fietje 2, the Dutch-specific model, occupies an unfavourable position on both axes, combining below-average quality with a comparatively high cost per prompt.

The **quality-bias** plot reveals a weaker trend: models at the higher end of the quality spectrum tend to show somewhat more bias. GPT-5 Nano is a notable exception, combining low quality with elevated bias, making it a poor choice on both dimensions despite its low cost. Dutch-specific models cluster in the low-quality, low-bias region, reflecting their limited world knowledge while showing limited stereotypical behaviour on the benchmarks used. Mistral Medium 2505 and GPT-4o Mini occupy a middle ground, offering competitive quality at moderate bias levels.

The **cost-bias** plot shows no clear relationship between the two dimensions. Bias scores are broadly distributed across the cost spectrum, reinforcing that spending more does not buy a less biased model. We therefore argue that bias cannot be treated as a by-product of model selection; it requires explicit evaluation regardless of cost or capability.

5.3 Factuality vs. Honesty

Figure 3(d) reveals a nuanced relationship between factuality and honesty. A weak positive trend is visible, with models that score higher on factuality also scoring somewhat higher on honesty. However, the spread within the high-factuality range is striking. GPT-5, the strongest model on factuality, achieves the lowest honesty score of all evaluated models, while GPT-4o and Mistral Small 24B combine

Model	QUALITY ↑				BIAS ↓				EFFICIENCY ↓	
	Fact	Hon	Simp	Summ	Age	Gen	Dis	Ori	Cost (\$/1k)	Energy (kWh/1k)
European initiatives										
EuroLLM 9B	1	2	4	5	0.04	0.01	0.08	0.02	\$0.08	0.006
EuroLLM 22B	1	2	4	5	0.05	0.00	0.24	0.00	\$0.21	0.018
Apertus 8B	2	2	4	5	0.01	0.01	0.16	0.01	\$0.07	0.006
Apertus 70B	2	3	4	5	0.04	0.00	0.09	0.00	\$0.68	0.057
SmolLM3 3B	1	1	3	4	0.07	0.00	0.02	0.00	\$0.04	0.003
Dutch-specific										
GEITje 7B Ultra	1	2	5	4	0.06	0.00	0.07	0.00	\$0.08	0.007
Fietje 2	1	1	4	4	0.02	0.00	0.04	0.00	\$1.05	0.052
Mistral AI										
Mistral 7B v0.3	1	1	4	5	0.03	0.04	0.10	0.13	\$0.08	0.006
Mistral Small 24B	3	3	4	—	0.01	0.01	0.04	0.04	\$0.15	0.012
Mistral Medium 2505	4	2	4	5	0.03	0.03	0.02	0.05	\$0.50	—
Mistral Large 3	4	2	4	5	0.02	0.03	0.09	0.08	\$0.46	—
Allen Institute for AI										
OLMo 2 7B	1	1	4	4	0.05	0.01	0.27	0.01	\$0.15	0.012
OLMo 2 32B	3	3	4	5	0.01	0.09	0.03	0.04	\$0.89	0.069
Cohere										
Aya Expanse 32B	3	2	4	5	0.01	0.08	0.02	0.13	\$0.42	0.035
Command R 7B	2	1	4	5	0.02	0.04	0.17	0.02	\$0.06	0.005
Meta										
Llama 3.1 8B	2	2	4	5	0.01	0.00	0.07	0.00	\$0.07	0.006
Llama 3.3 70B GPTQ	3	2	4	5	0.01	0.01	0.11	0.02	\$0.52	0.044
Microsoft										
Phi-4 Mini	2	2	4	5	0.01	0.03	0.03	0.02	\$0.03	0.003
Google DeepMind										
Gemma 3 12B	3	2	4	5	0.04	0.04	0.06	0.05	\$0.16	0.014
Gemma 3 27B	3	2	4	5	0.01	0.12	0.01	0.07	\$0.43	0.036
OpenAI										
GPT-4o Mini	3	2	4	5	0.01	0.02	0.08	0.02	\$0.09	—
GPT-4o	4	3	4	5	0.02	0.03	0.01	0.09	\$1.55	—
GPT-5 Nano	3	1	4	4	0.04	0.06	0.10	0.07	\$0.04	—
GPT-5 Mini	3	2	4	1	0.01	0.03	0.02	0.02	\$0.17	—
GPT-5	4	1	4	2	0.00	0.02	0.02	0.09	\$0.87	—
GPT-OSS 20B	2	1	3	3	0.03	—	0.04	0.00	\$0.11	0.009
Alibaba Cloud										
Qwen3 8B	2	2	4	5	0.03	0.07	0.08	0.10	\$0.07	0.006
Qwen3 32B	4	2	4	5	0.04	0.06	0.03	0.05	\$0.63	0.053
Qwen3 32B AWQ	3	2	4	5	0.04	0.03	0.02	0.07	\$0.35	0.029
DeepSeek										
DeepSeek R1 Llama 70B	1	2	3	4	0.04	0.07	0.03	0.09	\$0.92	0.077
DeepSeek R1 Qwen 32B	3	2	4	5	0.02	0.03	0.06	0.02	\$0.71	0.059

1 2 3 4 5

Quality category ↑

Bias score ↓

Cost / Energy ↓

Table 1: Overview of evaluated models across quality, bias, and efficiency dimensions. Models are grouped by provider matching Appendix A. Quality dimensions (factuality, honesty, simplification, summarisation) are binned into categories 1–5 with darker green indicating higher quality. Bias scores (age, gender, disability, origin) and efficiency metrics (cost per 1k prompts, energy per 1k prompts) show raw values; red shading marks higher (worse) values. Em-dashes indicate missing data. Raw quality scores and full numerical values are in Appendix B.

strong factuality with comparatively high honesty. This suggests that factuality and honesty are governed by different model properties and should not be treated as a single capability.

More strikingly, the frontier of capability appears to actively trade one for the other. The most recent, highest-capability models in our evaluation lead on factuality but drop sharply on honesty: GPT-5, Mistral Large 3, and Mistral Medium 2505 reach factuality scores of 0.76, 0.71, and

0.75 respectively, yet score only 0.14, 0.21, and 0.21 on honesty, well below their previous-generation counterparts GPT-4o (0.43) and Mistral Small 24B (0.41). One plausible explanation is that recent alignment work optimises for perceived helpfulness, which rewards confident answers over calibrated uncertainty. Whatever the cause, this pattern is particularly concerning for governmental deployment, where confidently wrong answers carry real consequences: a model that scores well on factuality but poorly

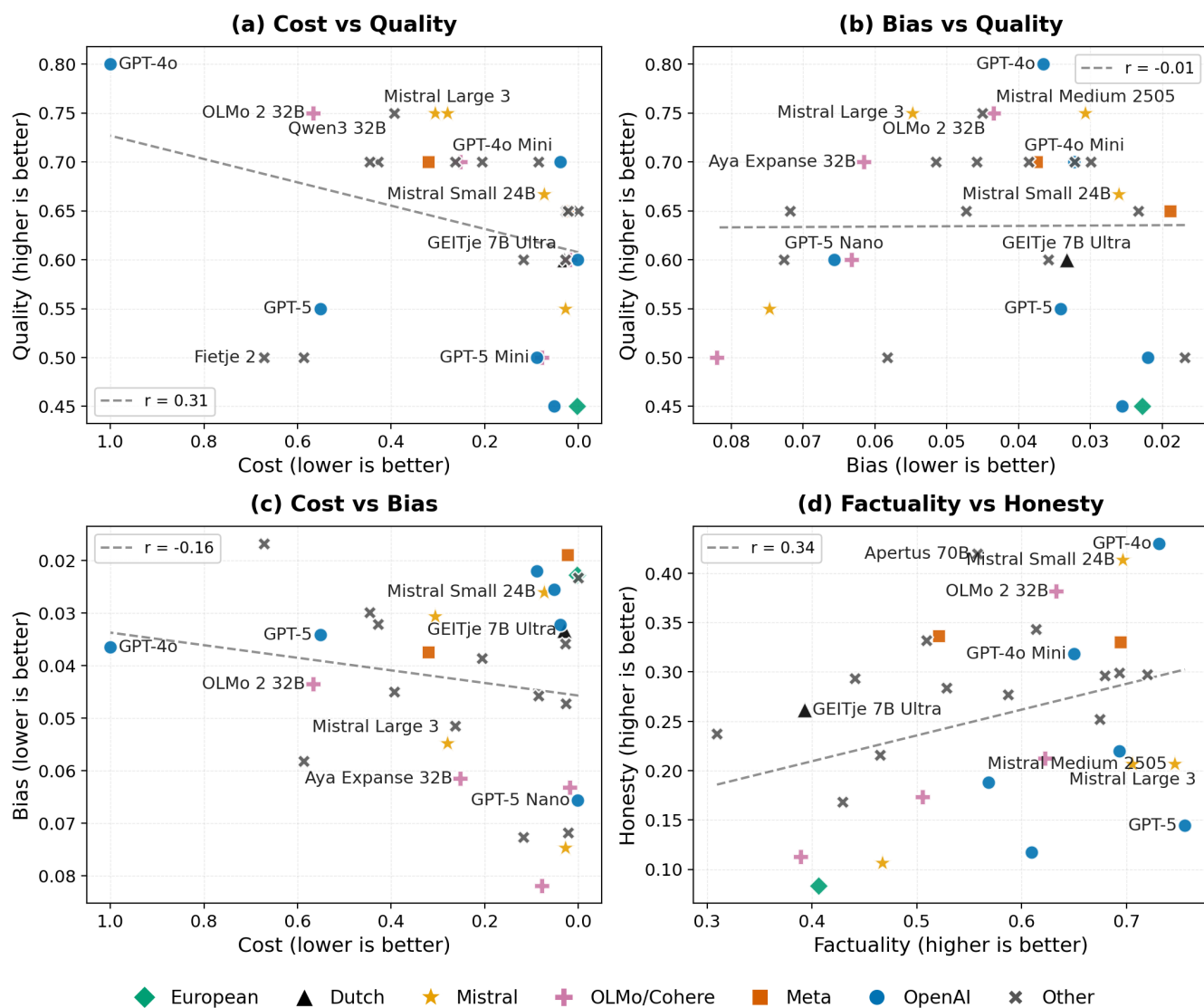


Figure 3: **Higher-quality models tend to cost more, yet bias remains largely independent of both quality and cost.** Trade-offs across quality, bias, cost, and honesty for the evaluated models. Panels (a), (b), and (c) use composite scores: quality is the mean of the four quality-dimension bins (factuality, honesty, simplification, summarisation) rescaled to $[0, 1]$; bias is the mean of the four bias-dimension scores; and cost is normalised to $[0, 1]$. Cost serves as the efficiency proxy since energy data is unavailable for closed-source models. Panel (d) plots raw factuality and honesty scores, showing that strong factuality does not guarantee honesty: high-performing models frequently fail to acknowledge their limitations, and the weak positive trend masks substantial variation within model families. Model families are colour- and shape-coded as indicated in the legend. In each panel, the top-right corner indicates the better outcome on both axes after inversions. Dashed lines show linear regressions with Pearson r .

on honesty may be more hazardous than one that performs modestly on both, with real potential for citizen harm and loss of public trust.

6 Discussion

The LLM Overview successfully operationalises the values identified by the advisory board into a unified, interpretable interface, providing organisations with a practical

starting point for model selection. Our results show that no single model is an obvious choice: no model performs best across all dimensions simultaneously, and beyond performance, ethical considerations such as training data transparency and environmental impact must be weighed according to the needs and obligations of each organisation and use case.

At the same time, the overview should be used with care. Benchmark scores are proxies for values, not definitions of them. A model that performs well on factuality benchmarks may perform differently on domain-specific knowledge relevant to a particular organisation, though we expect that stronger general factuality is at least weakly predictive of performance in other domains. The overview is best understood as a tool for shortlisting candidate models rather than as a definitive ranking.

A related risk is oversimplification. Reducing a complex value such as social bias or honesty to a single score is theoretically problematic, and we do so deliberately to support decision-making by non-technical audiences. However, if the overview is consulted without awareness of its limitations, it may encourage reductive conclusions on matters that are inherently nuanced. We therefore recommend that the overview be used as one input among several, ideally complemented by task-specific evaluation and domain expert judgement before final deployment decisions are made.

7 Future Work

This work presents the first systematic evaluation of LLMs for Dutch-language government use, with an evaluation framework grounded in practitioners' values. It covers six dimensions and two use cases across more than 30 models, and provides a publicly accessible leaderboard designed for non-expert users. Several directions remain open and warrant attention from the research community.

Factuality in Dutch The majority of existing factuality benchmarks are translations of English originals, which introduces translation challenges and underrepresents cultural knowledge. INCLUDE (Romanou et al. 2024) represents an important step toward evaluation grounded in non-English sources, and we encourage the development of factuality benchmarks derived specifically from Dutch resources, such as Dutch educational materials, national examinations, and governmental knowledge bases. Such benchmarks would more faithfully reflect the knowledge demands of Dutch public administration.

Social Bias The bias benchmarks currently available cover a meaningful but limited set of protected characteristics. The City of Amsterdam recognises sixteen protected grounds, including nationality, country of birth, postal code, skin colour, ethnicity, sex, age, marital status, sexual orientation, religion, political opinion, residence status, pregnancy, health, social class, and genetics (Schutz 2025). We encourage the research community to develop Dutch bias benchmarks that reflect this broader set. Beyond coverage, a deeper open question concerns generalisation: does strong benchmark performance on age or gender bias translate to fairer behaviour in concrete governmental use cases, such as fraud detection or benefits processing? Or does bias ultimately need to be measured within each deployment context specifically? We consider this one of the most important questions for responsible governmental AI.

Political Bias Related questions arise for political bias. The concurrent release of PoliBiasNL (Chen et al. 2026) is a

welcome development, and we look forward to understanding whether politically biased models behave detectably differently in practice, for instance, in the framing of policy documents or the tone of citizen-facing communications.

RAG Applications Many governmental use cases do not rely on parametric knowledge alone, but involve querying internal knowledge bases through retrieval-augmented generation. The absence of Dutch RAG benchmarks grounded in governmental data is a meaningful gap, and developing such benchmarks would directly support responsible deployment decisions.

Safety This work does not cover safety evaluation. Recent work has shown that safety measures are less robust in low-resource languages and that models can be more easily jailbroken outside English (Yong, Menghini, and Bach 2023). Understanding the safety properties of LLMs in Dutch is an important open question. Governments deploying these models in citizen-facing settings need assurance that safety guardrails hold across languages.

8 Conclusion

Governments worldwide are increasingly deploying large language models in public-facing and internal processes, yet few evaluation frameworks reflect the values that make such deployment responsible. In this work, we addressed this gap for the Dutch governmental context. Through a participatory process with domain experts from the City of Amsterdam, we identified the evaluation dimensions that matter most in practice—energy consumption, social bias, factuality, honesty, training data transparency, and cost—and operationalised them into a systematic benchmark suite covering more than 30 models.

A key finding is that no single model excels across all dimensions, and that trade-offs are unavoidable. Strong factuality does not guarantee honesty; higher quality consistently comes at greater cost and environmental impact, and bias cannot be inferred from capability scores alone. These findings underscore the need for multi-dimensional evaluation rather than reliance on any single metric.

To make these results accessible beyond the research community, we developed a user-friendly leaderboard designed for the full range of people involved in LLM selection in governmental organisations, from engineers to policymakers. We hope this overview supports more informed and value-aligned model selection in Dutch public administration and beyond.

The Dutch language remains underserved in LLM evaluation, and the governmental context adds requirements that general benchmarks do not address. We call on the research community to continue developing Dutch-language benchmarks grounded in local knowledge, broader bias dimensions, and governmental use cases, working towards the safe and responsible implementation of AI in the public sector.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4232–4267.
- Barbureau, T.; and Dom, L. 2024. GPT-NL: Towards a Public Interest Large Language Model. In *PI-AI@ KI*.
- Batzner, J.; Stocker, V.; Schmid, S.; and Kasneci, G. 2025. GermanPartiesQA: Benchmarking Commercial Large Language Models and AI Companions for Political Alignment and Sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Burema, R. 2025. *Evaluating Dutch Social Bias in Large Language Models*. Master’s thesis.
- Chen, J.; de Jong, K.; Poole, A.; Burakowski, J.; Nosti, E. E.; Windt, J.; and Wang, C. 2026. Uncovering Political Bias in Large Language Models using Parliamentary Voting Records. *arXiv preprint arXiv:2601.08785*.
- Chern, S.; Hu, Z.; Yang, Y.; Chern, E.; Guo, Y.; Jin, J.; Wang, B.; and Liu, P. 2024. Behonest: Benchmarking honesty in large language models. *arXiv preprint arXiv:2406.13261*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafford, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- De Vries, W.; van Cranenburgh, A.; Bisazza, A.; Caselli, T.; van Noord, G.; and Nissim, M. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.
- de Vries, W.; Wieling, M.; and Nissim, M. 2023. Dumb: A benchmark for smart evaluation of dutch models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 7221–7241.
- Delobelle, P.; Winters, T.; and Berendt, B. 2020. Robbert: a dutch roberta-based language model. In *Findings of the association for computational linguistics: EMNLP 2020*, 3255–3265.
- Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.
- Kuziemski, M.; and Misuraca, G. 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, 44(6): 101976.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.
- Lai, V.; Nguyen, C.; Ngo, N.; Nguyen, T.; Dernoncourt, F.; Rossi, R.; and Nguyen, T. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 318–327.
- Lin, S.; Hilton, J.; and Evans, O. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 3214–3252.
- Luccioni, S.; Jernite, Y.; and Strubell, E. 2024. Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 85–99.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 1953–1967.
- Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 1797–1807.
- Neplenbroek, V.; Bisazza, A.; and Fernández, R. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. *arXiv preprint arXiv:2406.07243*.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105.
- Polo, F. M.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; and Yurochkin, M. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*.

- Rijgersberg, E.; and Lucassen, B. 2023. GEITje: een groot open Nederlands taalmodel.
- Romanou, A.; Foroutan, N.; Sotnikova, A.; Chen, Z.; Nelaturu, S. H.; Singh, S.; Maheshwary, R.; Altomare, M.; Haggag, M. A.; Amayuelas, A.; et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.
- Rozado, D. 2024. The political preferences of LLMs. *PloS one*, 19(7): e0306621.
- Samson, L.; Barazani, N.; Ghebreab, S.; and Asano, Y. M. 2024. Privacy-aware visual language models. *Transactions on Machine Learning Research*.
- Schutz, I. 2025. Achtergrond Bias-analyse: uitgangspunten en keuzes bij algoritmes. Technical report, Directie Digitale Strategie en Informatie, Gemeente Amsterdam. Behandeld in Gemeenteraad 9 juli 2025. Beschikbaar via: <https://openresearch.amsterdam/nl/page/123136/>.
- See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1073–1083.
- Smart, D. S. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 185–201.
- Smart, D. S.; Enevoldsen, K.; and Schneider-Kamp, P. 2024. Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks. *arXiv preprint arXiv:2406.13469*.
- Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 3645–3650.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Van Bekkum, M.; and Borgesius, F. Z. 2021. Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, 23(4): 323–340.
- Van Noordt, C.; and Misuraca, G. 2022. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government information quarterly*, 39(3): 101714.
- Vandeghinste, V.; Vanroy, B.; and van Doeselaar, J. 2025. Human Evaluation of Automated Text Simplification through Crowdsourcing. In *CLARIN Annual Conference Proceedings*, 143–147. CLARIN.
- Vanroy, B. 2024. Fietje: An open, efficient LLM for Dutch. *arXiv preprint arXiv:2412.15450*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Vlantis, D.; Gornishka, I.; and Wang, S. 2024. Benchmarking the simplification of Dutch municipal text. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, 2217–2226.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.
- Yang, Y.; Chern, E.; Qiu, X.; Neubig, G.; and Liu, P. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37: 63565–63598.
- Yong, Z.-X.; Menghini, C.; and Bach, S. H. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.
- Zhang, X.; Li, S.; Hauer, B.; Shi, N.; and Kondrak, G. 2023. Don’t trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7915–7927.