# Fair balancing? Evaluating LLM-based Privacy Policy Ethics Assessments

Vincent Freiberger[1,2,*], Erik Buchmann[1,2]

[1]*Leipzig University, Leipzig, Germany*
[2]*Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany*

## Abstract

At this moment, we observe a strong increase in use cases and organizations using machine learning and artificial intelligence (AI), particularly large language models (LLMs) like OpenAI's GPT series. This has many advantages for organizations. However, it becomes increasingly important to evaluate ethical and fairness-related aspects of using personal information as data for model training or as input for automated processes. Therefore, privacy policies are an important resource. Privacy policies aim to make data collection, sharing, and usage transparent to users. However, privacy policies are also known to be long and complex. This has raised issues like failing to understand such policies or even consent fatigue, i.e., users just accepting all tracking of their data. Potentially abusive or unfair privacy practices may remain unnoticed. An independent, automated assessment of ethics in privacy policies could help in fairly balancing out existing information asymmetries.

We explore using an LLM for an ethics assessment of data management practices documented in privacy policies. In particular, we develop and systematically evaluate a prompting template for ethics evaluation. By means of qualitative experiments with privacy policies from the Top-100 German web shops, we quantitatively investigate the robustness and quality of the LLM-based ethics assessment, and how varying roles, user interests, ethical frameworks, etc. in the LLM prompts affect the assessment. To the best of our knowledge, we are the first to investigate how LLMs can be used for ethics assessments of data management practices. Our results show that LLM-based ethics assessment, yet still limited in its specificity and consistency, shows promise for the future. The identified criteria are consistent with those from related work. We find that varying the role assigned to the LLM has the largest effect on the LLM's ethics assessment. An ethics assessment could allow end users to make more informed decisions. The moral judgment of LLMs regarding online privacy is not just relevant to policy assessment, but could be used to investigate specifications, regulations, norms, or legal documents.

### Keywords

Large Language Models, Privacy Policies, Ethics, Morality

## 1. Introduction

We observe a rapid proliferation in the use of machine learning and artificial intelligence (AI) in our everyday lives. A large share of the data from today's domestic devices, public transportation, electric vehicles, industrial control systems, and office equipment is used for training or analyzed by AI. Amongst the technology industry's biggest companies are some

---

*EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany*

[*] Corresponding author.

freiberger@cs.uni-leipzig.de (V. Freiberger); buchmann@informatik.uni-leipzig.de (E. Buchmann)
0009-0000-8549-9119 (V. Freiberger); 0009-0009-5874-4313 (E. Buchmann)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3908/paper_30.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

that collect excessive amounts of sometimes even personally identifying data [1]. User data is valuable to companies as it allows them, for instance, to track clients, improve services, target marketing, or sell data to third parties. The increasing use of large language models (LLMs) such as GPT-3.5 [2, 3], Gemini [4], or Opus [5] in complex IT ecosystems is likely to increase the impact on the user's privacy by an order of magnitude.

This makes it challenging for users to decide if a digital service meets their privacy preferences and moral standards. Organizations are required by General Data Protection Regulation (GDPR) [6] in the EU to publish a privacy policy. It must contain data management practices of the organization, as well as users' rights regarding their data. Existing approaches from data protection legislation assume an aware user. This user actively reads the privacy policy and boilerplate information of each service, to make an informed decision about using a service and consenting to data collection. Given hundreds of services every day and non-transparent multi-page privacy statements in fine print [7, 8, 9], it is unlikely that this approach ensures a fair balance between the interests of the service provider and the service user. We investigate this issue by using an LLM for an ethics assessment of privacy policies. In particular, we want to find out if an LLM can be used to automatically identify data management practices that conflict with existing social and ethical standards regarding fairness, accountability, etc. For example, a policy would be considered informationally unfair, if it extensively uses complex legal phrases that discriminate against people with dyslexia and non-native speakers. An example of an unethical policy would be one, that grants extensive rights to utilize the user's personal data.

An automated assessment of such aspects is challenging. Privacy policies need to justify, particularly for AI-based services, complex data collection and usage practices. That results in a large spectrum of ethical problems that may be encountered, depending on the context of the privacy policy and the assessment. Thus, we strive for insight into the shortcomings of LLMs in moral judgment regarding online privacy. Our research questions are as follows:

*RQ1. How capable are LLMs of assessing the ethicality of privacy policies?*

*RQ2. Do variations in the context of the LLM prompt influence its ethics assessment?*

To approach our research questions, we perform experiments on recent German privacy policies with the at the time of writing widely used LLM, GPT-3.5 turbo [2, 3]. Our experiments systematically vary the role assigned to the LLM for assessment, the interests of the user affected by the policy, the ethical framework assigned for assessment, and the scope and temporal span of the assessment in the prompt. Also, we assess the robustness of moral judgment to a different seed for generation, and to a paraphrased wording of the prompt. We evaluate the LLM outputs' quality on a broad set of metrics. In particular, we make the following contributions:

- We conduct an ethics assessment on 55 privacy policies of the Top-100 German web shops and evaluate their quality.
- We identify how changes in the prompt influence the ethics assessment of an LLM. In total, we run 9240 experiments with GPT-3.5 turbo.
- We reveal ethical shortcomings of recent privacy policies, based on 1116 distinct criteria returned by our experiments.

To the best of our knowledge, we are the first to systematically evaluate an LLM's ethics assessment of privacy policies. Such an assessment offers guidance to users, and provides insight into the moral judgment of LLMs regarding online privacy. When looking at the details, we find that an LLM-based assessment of privacy policies is still limited in its depth, specificity, and consistency. The identified criteria are consistent with related work on privacy policies, and particularly a change in role changes the perspective of an assessment. Thus, it is a step towards fairly balancing provider and user interests by empowering the user.

**Paper structure:** Section 2 reviews related work. Section 3 outlines our research method. The Sections 4 and 5 evaluate and discuss our results. Finally, Section 6 concludes.

## 2. Related work

This section provides a brief overview of ethics, privacy policies, large language models, and tools from computer linguistics needed to evaluate an LLM output.

### 2.1. Ethics

Privacy ethics addresses the degree of access others have to one's information as well as control one has over it and actions one can take concerning one's privacy [10, 11, 12, 13]. It discusses often complex privacy trade-offs and the balance of power between the data holder and data subject [14, 15]. Seeking privacy serves two major interests: Security interests (stay unharmed) and privacy per se [16]. Privacy per se is about influencing the way we present ourselves to others and, more broadly, our autonomy [17, 16]. The philosophical debate on privacy ethics distinguishes between privacy and the right to privacy. Depending on the context of the acquisition of information about oneself by another party and the underlying intention privacy may be violated, however, not the right to privacy [17]. Hence, focusing on the right to privacy leans towards a deontological perspective, whereas viewing privacy by itself leads to a consequentialist perspective.

In the realm of online privacy, ethical concerns regarding surveillance [16, 18], impact of choice [19], manipulation [20, 15], and power imbalance [15, 14] have been raised.

Codes of ethics, legal statutes, or international declarations could provide helpful input for an ethics assessment of online privacy [21]. They represent values ingrained in our society. The Universal Declaration of Human Rights (UDHR) [22] and the European Convention on Human Rights (ECHR) [23] provide a relevant minimum standard to consider in an ethics assessment. Aspects like freedom, equality, and dignity (Art. 1 UDHR), non-discrimination (Art. 2 UDHR, Art. 21 ECHR), freedom of expression (Art. 19 UDHR), protection of children, elderly and the disabled (Art. 24, 25, 26 ECHR), protection of intellectual property (Art. 27(2) UCHR), or consumer protection (Art. 30 ECHR) are stated. Additionally, the Digital Services Act [24] could be seen as a foundation of ethical principles. Amongst the principles it enforces are the legality of content, accountability, non-manipulative practices, transparency, and the protection of minors. Underlying virtues like inclusiveness, protection of the vulnerable or accountability could motivate an approach considering virtue ethics.

This leads us to define the following quality criteria for an assessment: It should consider all relevant perspectives and stakeholders with enough depth in its normative grounding [25]. Ethical aspects may come into effect unintended, sometimes as second-order consequences [25] which need to be considered. The ongoing debate about privacy ethics and its definition [12, 11, 16, 17] shows that an assessment must be thorough, consistent, structured, and comprehensible in its reasoning. Consistency of moral advice given by the model is important as it influences user judgment [26]. A good assessment should be concise and understandable for everyone as we want to provide informational fairness [9]. Addressing privacy trade-offs and power imbalances is essential for a meaningful assessment [14, 15].

LLMs can be investigated based on moral psychology [27]. This involves looking into the extent to which moral reasoning and moral judgments are represented in model outputs and into biases the model might have in that regard. For instance, ChatGPT has been found to be inconsistent in its moral advice when facing a moral dilemma [26]. The closest to our work is the ETHICS data set [28] used to investigate general ethical judgment of language models.

## 2.2. Privacy policies

Privacy policies aim to make data collection, sharing, and usage transparent to users [29]. A website owner is required by the General Data Protection Regulation (GDPR) [6] to publish a privacy policy. There is an information asymmetry between website owners' full knowledge of their privacy practices as well as their potential shortcomings and the average users' unawareness of them. A privacy policy addresses this information asymmetry between service provider and user and helps in creating trust. Specifically, privacy policies should inform users about their data protection rights and explain data management practices. This includes clearly stating retention periods as well as to which third parties what data is transferred and explaining the respective purpose. The GDPR establishes the principles (1) lawfulness, fairness and transparency, (2) purpose limitations, (3) data minimization, (4) accuracy, (5) storage limitations, (6) integrity and confidentiality, and (7) accountability.

The advance of generative AI-based applications, and services integrating them, introduces more complexity to data protection and respective privacy policies. Interests between the service provider and user are conflicting. Users may want their data to stay as private as possible. The service provider can leverage data, for instance, to improve services or target marketing, and wants to collect data from users. The GDPR handles this conflict of interest by enforcing a notion of fair balancing [30]. This means that the privacy risks faced by users should be balanced with the business interests of the service provider. We would like to note that this two-sided view is a simplification of reality that involves many more stakeholders with differing interests, e.g., politicians, consumer protection initiatives, independent ethicists, etc.

One common issue is transparency in privacy policies [31, 7, 8, 32]. Policies tend to be long, written in inaccessible language, and hard to understand. This informational unfairness [9] leads to issues like consent fatigue [33] and users potentially being exposed to unethical privacy practices without their awareness of the consequences. Privacy policies can hide and mitigate unethical data handling practices and deceive users into trusting the service using persuasive appeals [31, 34]. This was not yet in the context of the GDPR. Unfair representations are another common problem [9]. Privacy policies contain for instance gender bias. A third type of issue is lacking fair balancing, as seen in privacy policies that have claimed all rights over users' data [35]. Over time consistent ethical shortcomings can cause privacy fatigue which can have a stronger influence on online privacy behavior than privacy concerns [33].

The given issues, particularly regarding complexity and length of privacy policies, have motivated privacy assistants [36, 37, 38]. Emerging capabilities by scaling up LLMs [39] have given them a wide range of applicability [40]. This makes them interesting as a tool for assessing privacy policies. They have been utilized by prompting them with a privacy policy and automatic queries on concepts around privacy and compliance [41], for topic classification of sections of policies [42], or as an interactive question-answering assistant [43].

## 2.3. Large language models and prompting

LLMs generate text for a given prompt by iteratively predicting the next token based on probability. GPT-3.5 is an LLM based on GPT-3 [3] improved by reinforcement learning with human feedback [2]. It is, at the time of writing, the underlying model of the free version ChatGPT with millions of users [44]. The model version GPT-3.5 turbo supports up to 16385 tokens in context length and is accessible via OpenAI's API. GPT-3.5 is capable of deductive, abductive, and commonsense reasoning but struggles with inductive reasoning [45, 46]. Deductive reasoning goes from rather general concepts in the reasoning process down to specifics. Inductive reasoning generalizes from specifics. Abductive reasoning takes a set of observations and draws the most likely conclusion. Commonsense reasoning is "understanding and reasoning about everyday concepts and knowledge that most people are familiar with, to make judgments and predictions about new situations" [46].

LLMs have recently been shown to have limited capabilities in many different domains [47, 40]. Smart and complex prompting strategies [48, 49, 50, 3, 51, 52] have been used to address limitations. For a general review on prompting LLMs, we refer to the literature [53].

Particularly reasoning-related tasks benefit from breaking down the problem and solving it step-by-step with an LLM [48, 49]. Such prompting strategies are referred to as Chain-of-Thought prompting [50]. To improve the result, asking a model to reflect and improve its output has been shown to improve quality [54]. To guide a model in a specific direction for model outputs, giving examples for potential outputs is effective [3]. This is also referred to as few-shot prompting. Repeating particularly relevant aspects in the input prompts helps with the model considering them in the generated output [55]. Controlled slight variations of the prompt can have a big impact on the model output [27]. If reproducible outputs are important, the seed needs to be fixed and other parameters need to be unchanged over the experiments [56]. Robustness testing of LLM outputs can be implemented by paraphrasing the prompt [41, 57, 58].

## 2.4. Tools from computer linguistics

We borrow from linguistics to evaluate the output of LLMs. The readability of text can be measured with the Flesch Reading Ease (FRE) [59]. We refer to word count and FRE as surface-level metrics. Sentiment captures the positive or negative associations transferred by text. Sentiment analysis delivers the most accurate results when performed sentence-wise. For German language, german-sentiment [60] can be used as a tool. Embeddings are numeric vectors representing the semantic meaning of the words or text for which they are calculated [61]. OpenAI's small embedding model version 3 [62] calculates 1536-dimensional text embeddings sufficient to capture a high degree of semantic nuance. For aggregating text embeddings, max-pooling can be effectively utilized [63]. Max-pooling takes the maximum for every dimension of the embedding vectors over all embeddings that are pooled. Calculating cosine similarities is commonly done to compare semantic similarity between embeddings.

# 3. Our research method

This section explains how we create our prompting template, extract relevant information from the LLM outputs, and evaluate the LLM assessment.

## 3.1. LLM experiments

Our prompting template follows the LLM prompting techniques introduced in Section 2.3. It can be found on the project's GitHub repository[1]. Since our privacy policies are written in German, we prompt the LLM in the same language. The system prompt describes the task for the LLM. The user prompt gives the privacy policy as input for this task. For example, Figure 1 shows an English translation of a specific system prompt and a user prompt template.

**System prompt:** You are a critically scrutinizing, experienced expert on an independent ethics board. You have a nuanced and in-depth perspective on ethical issues. You already have a lot of experience in the precise ethical assessment of privacy statements. You will be given a privacy policy. Your task is to uncover aspects in privacy policies that are ethically questionable from your perspective. To do this, proceed step by step:

1. Identify relevant ethical test criteria for this privacy policy from your perspective as criteria for a later assessment. When naming the test criteria, stick to terms and concepts that are as standardized/common in the field of ethics as possible.
2. Based on this, check for ethical problems or ethically questionable circumstances in the privacy policy. Describe your analysis.
3. Only after you have completed step 2: Based on your analysis, rate the privacy policy against each of your criteria on a 5-point Likert-scale. Explain what this rating means. Explain what the ideal case with 5 points and the worst case with one point would look like. The output in this step should be formatted in bold as follows:
   [Insert rating criterion here]: [insert rating here]/5 [insert line break]
   [insert justification here]
4. Important: State precisely in keywords what assumptions you have made regarding your [role], the [user interests], the [scope of ethical implications] (local, global, etc.), the [long or short term nature of ethically relevant implications] and the [ethical frameworks used] (e.g. utilitarianism, virtue ethics, deontology,...). You must put these 6 assumptions in square brackets [].
5. Reflect on your assessment and check whether it is complete. Show how your result is anchored in ethical frameworks. If something is missing, add it! Important: Check for errors in your analysis and correct them if necessary.

You must present your approach clearly and follow the steps mentioned.
**User prompt:** The privacy policy: <Privacy policy text is inserted here>

**Figure 1:** English translation of a prompt template

Our prompt tells the LLM to (a) generate a set of assessment criteria for each policy, (b) evaluate on a five-point Likert-scale how well each policy meets each criterion, and (c) produce output with a given structure and delimiter marks, that allows extracting parts of the output via regular expressions. The prompt structures outputs into five steps: (1) *Step 1* finds suitable assessment criteria for the policy being assessed, i.e., each policy might be evaluated with different criteria. (2) *Step 2* performs a written ethics assessment based on the criteria from step 1. (3) *Step 3* assigns a suitable Likert-score based on the assessment in step 2 for each criterion. (4) *Step 4* outputs assumptions made by the model on the role it takes when doing the assessment, the user interests, the temporal span considered in the assessment, the scope of the assessment, and the ethical framework utilized for the assessment. (5) *Step 5* asks for reflection on the result

---

[1] https://github.com/Freiberger97/LLM-based-privacy-policy-ethics-assessment

**Table 1**
Overview of variables for prompt variation (more specifics in the GitHub)

| Variable | Variations | Explanation |
|---|---|---|
| Role | independent ethicist, politician, CEO, investor, average user, consumer protection, data protection officer | Assigns the LLM with a role and by that a specific perspective for the assessment. Mimics an assessment from different stakeholders' points of view. |
| User interest | trust, fairness, morality, min. risk, societal good, accountability, data ethics, unknown, average user interests | Gives the LLM information about preferences of the affected user. A user may have different preferences which may influence the outcome of an assessment regarding ethics of privacy policies. |
| Ethical framework | consequentialism, deontology, virtue ethics, other | Introduces an ethical school-of-thought to be used for assessment. A change can drastically change the moral conclusions drawn. |
| Scope of assessment | local effects, global effects, indifferent | Gives information on how far-reaching the scope of an assessment is. It may make a difference whether only effects on the user or effects globally are assessed. |
| Temporal span | short-term effects, mid-term effects, long-term effects, indifferent | Indicates the time horizon considered for assessment. Some consequences may only materialize over time. |

and for a justification of the ethical framework that has been utilized. To evaluate the ethical capabilities of the LLM, we experiment with different contexts.

In particular, we have the objective of:

- giving a different perspective to an ethics assessment. More nuanced results should be possible by combining results from individual variations.
- understanding assumptions an LLM defaults to if given incomplete information.
- understanding how sensitive the LLM is to specifications of important variables for an ethics assessment.

We vary the context in five dimensions, as shown in Table 1. Detailed explanations can be found in our GitHub repository. In our experiments, we vary one dimension after the other. As the second column in Table 1 shows, this is a total of 27 variations, plus a baseline without such information. In addition, we test the robustness of the LLM's outputs by using three different seeds held constant for all variations and policies. Finally, we use [64] to translate each prompt to English and back to German, to obtain a paraphrased version of the prompt. We test each paraphrased prompt with the same three seeds to find out, if variations in the wording have an impact on the results. In total, we test 168 different prompts against each of our privacy policies.

Our data set is based on the Top-100 most popular German web shops from 2020 to 2023. [7] explains how we have obtained, cleansed and preprocessed this data set. Further, we dropped poorly scraped documents (e.g. missing content) and enforced a suitable length for GPT-3.5 turbo's context length (16385 tokens). In the end, we have the most recent versions of 55 distinct privacy policies in our data set. With our 168 different prompts, this results in 9240 individual experiments. We decided to use GPT-3.5 turbo [2, 3] via the OpenAI API, because it is similar to the well-known, free version of ChatGPT. Except for the seed and the system prompt, we used the default settings. Figure 2 shows an output for the prompt in Figure 1. All outputs can be found in our GitHub repository.

1. **Identified ethical review criteria**: - Transparency: How transparent are the company's data practices towards users? - Consent: Has user consent been obtained for data use in a legally and ethically correct manner? - Data minimization: Is only the data necessary for the respective purpose collected and processed? - Data security: Are appropriate measures implemented to protect data from security risks? - Data disclosure: Is it openly communicated to whom personal data is disclosed and for what purpose?
2. **Analysis of the privacy policy**: The privacy policy contains some ethically questionable aspects: - The collection of IP addresses every time a website is accessed without the explicit consent of the user and without a clear indication of how the data is anonymized. - The transfer of personal data to external third parties without an explicit explanation of who these third parties are and for what purpose the data is being transferred. - The use of cookies and tracking technologies without clear information on the purpose, storage duration and the possibility of deactivation.
3. **Assessment of the privacy policy**: - Transparency: 2/5 The privacy policy is unclear regarding data transfer and use. - Consent: 3/5 Consent is obtained for the use of cookies and data processing methods, but could be more detailed. - Data minimization: 2/5 IP addresses and personal data are collected, the necessity of which is not sufficiently explained. - Data security: 3/5 Security measures are mentioned, but no details on specific implementation. - Data transfer: 2/5 The transfer of data to external third parties raises questions, as it is not clear who these third parties are and for what purpose data is transferred.
4. **Assumptions**: - [Role]: Independent ethics expert on the Ethics Council - [User interests]: Protection of personal data, transparency, compliance with data protection regulations - [Scope of ethical implications]: Local (data protection on the website) - [Long- or short-term nature of ethically relevant implications]: Long-term impact on user privacy and trust - [Ethical frameworks used]: Data protection laws, ethics guidelines, principles of data ethics
5. **Reflection and supplementation**: The analysis of the privacy policy focused mainly on transparency, consent, data minimization, data security and data sharing. It is important that the privacy policy provides clear and concise information so that users can make informed decisions about their data. A more detailed examination of data retention periods, user rights and the implementation of data protection measures could add further depth to the analysis. It would be important to consider implications for vulnerable groups and possible privacy implications.

**Figure 2:** Example output translated to English for this paper (all outputs in the GitHub)

## 3.2. Evaluating model outputs

We evaluate the quality of ethics assessments of the LLM (RQ1) by evaluating overall metrics, comparing between policies, and assessing robustness to changed seeds or paraphrasing. We evaluate the impact of variations in the prompt context (RQ2) by comparing metrics between variations. Our metrics are based on the quality criteria established in Subsection 2.1, and we use the tools outlined in Subsection 2.4. Table 2 gives an overview of our evaluation. Column 2 indicates if a metric is computed over a single criterion. Column 3 considers one of the five steps the model is instructed to produce in its output (e.g. assumptions as step 4), and Column 4 points out if a metric considers the LLM output as a whole.

We calculate and compare the **number of assessment criteria** produced by the LLM. This indicates how specific, nuanced and in-depth an assessment may be.

To find out whether deviations in the context of the prompt lead to a different assessment, we compute the **criteria occurrence ratio**, i.e., the number of the assessment criteria that appear in multiple experiments with the same privacy policy. This helps judge how specific an assessment is in addressing a policy.

**Table 2**

Overview of the LLM output evaluation

| Metrics | Criteria output | Step in output | Whole output | What is measured |
|---|---|---|---|---|
| Number of assessment criteria | ✓ | | | overall, between variations; between seeds/paraphrasing; between policies |
| Criteria occurrence ratio | ✓ | | | |
| Distinctiveness of assessment criteria | ✓ | | | Similarity of criteria definitions between outputs; Similarity of distinctly labeled criteria overall |
| Descriptive statistics assessment scores | ✓ | | | overall scores vs overall sentiment; between variations; between seeds/paraphrasing; between policies |
| Surface-level metrics | | | ✓ | overall, between variations, between seeds/paraphrasing |
| Sentiment | | ✓ | ✓ | |
| Document embedding | | | ✓ | similarity between seeds/paraphrasing, between variations, between policies |
| Assessing assumptions | | ✓ | | between variations |

We further measure **distinctiveness of assessment criteria** in model outputs, i.e., how similar the assessment criteria are. Therefore, we measure the cosine similarities between the criteria from different outputs. We also measure cosine similarities of criteria definitions for a criterion overall and for criteria within an output. Distinctive assessment criteria that do not deviate from one run to another for the same policy indicate a consistent assessment.

We also calculate **descriptive statistics of assessment scores**, like the mean or standard deviation for the criteria. All of these descriptive statistics can be compared for the different prompt variations or between policies. Descriptive statistics indicate the impacts prompt variations have and can be assessed regarding specificity and consistency.

We use the model output as a whole to capture **surface-level metrics** on how concise and readable an assessment is, like word count or the FRE. Our prompt tells the LLM to generate for each criterion both a textual assessment and a Likert-score. To check the LLM output for consistency, we compare the **sentiment** of the assessment in step 2 of the LLM output. As a further consistency check, we evaluate the sentiment of the entire output.

We compute cosine similarities of the **document embeddings** of the whole outputs. By using max-pooling, we compare both variations and policies. Different policies having different assessments indicate whether the LLM adjusts its assessment to the differences between policies, or not. We also find out if prompt variations have a large impact on the ethics assessment.

**Assessing assumptions** evaluates the variations (role, user interest, ethical framework, scope of assessment, temporal context). Again, we compute the cosine similarities of assumptions made (step 4 of the output) to compare between prompt variations using max-pooling. This allows us to evaluate their inter-dependencies towards the moral judgment of the LLM.

# 4. Results

In this section, we evaluate the LLM's ethics assessments (**RQ1**) regarding the ethics criteria generated by the LLM, and we investigate its robustness w.r.t. seeds/paraphrasing or the differences between policies. We also address how variations in the context of the prompt influences the ethics assessment (**RQ2**). Our evaluation follows the metrics in Table 2. Our metrics are weakly to moderately correlated.

**Number of assessment criteria:** In all outputs of the LLM, we find 44987 criteria, of which 1116 are distinct. Figure 3 shows the number of distinct criteria per policy over all respective outputs of the LLM. We find a large difference in the number of distinct criteria being considered between policies. Comparing between policies, the number of criteria considered in a single output ranges from 4.46 to 5.45 on average. This suggests that the LLM adjusts to the requirements of assessing a specific policy. An unexpected finding was that variations of the prompt do not lead to notable changes in the number of criteria being considered for assessment. The rather small number of criteria per output raises concerns about the nuance of assessments.

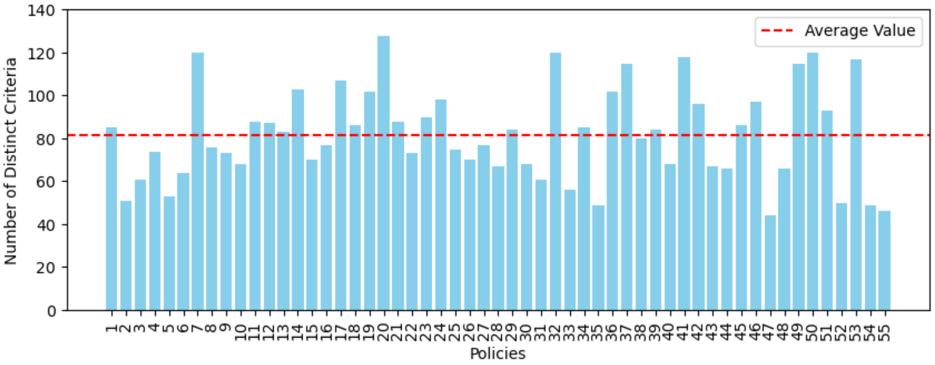

**Figure 3:** Number of distinct criteria over policies

**Criteria occurrence ratio:** The criteria 'Data minimization', 'Data security', 'Data sharing', 'Data sharing with third parties', 'Consent', 'Rights of data subjects', 'Security', 'Transparency', and 'Purpose limitation' appear for each policy in assessments (all criteria were translated to English). These criteria are closely related to requirements specified by the GDPR. They also reflect common issues regarding the fairness of privacy policies (see Section 2.2). Comparing the different prompt variations reveals that the three most occurring criteria (transparency, consent, and data security) are the same across all 28 prompt variations. The occurrence ratio over all outputs of the LLM can be found in Figure 4 for the 15 most frequently mentioned criteria, which are translated to English. Comparing robustness between normal and paraphrased prompt, we find that some criteria, like transparency, appear consistently more often with the paraphrased prompt, whereas others, like purpose binding, appear consistently less often. Comparing between policies we also find considerable differences in the criteria occurrence ratio. This is desirable as it shows that the assessment adapts to the specifics of a policy. Among less frequent criteria, we find many synonyms of more frequent criteria or criteria that combine two more frequent criteria into one criterion. This shows limitations in consistency.

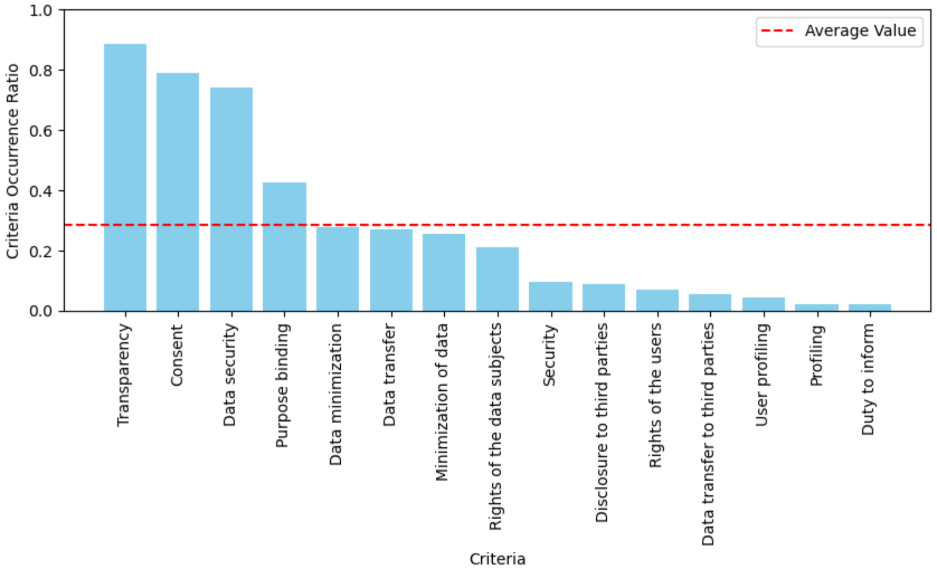

**Figure 4:** Occurrence ratio of the 15 most frequent criteria over all outputs of the LLM (translated)

**Distinctiveness of assessment criteria:** The average cosine similarity between all criteria is low (mean: 0.38). Hence, most criteria are distinct. A closer inspection of just the 50 most frequent criteria reveals that a few criteria are highly similar and hence possibly redundant. We go further by assessing the similarity of criteria definitions. We find that definitions for the same criterion assigned by the LLM are, on average (for the 30 most frequent criteria), not highly consistent (cosine similarity between 0.54 and 0.68). The similarity of definitions for distinct criteria in an individual output is rather high (mean cosine similarity: 0.59). The range for this between variations is slim (between 0.58 and 0.60). Hence, the consistency in the definition of criteria and their distinctiveness are not ideal.

**Descriptive statistics of assessment scores:** The differences in mean scoring averaged over all criteria per output are shown in Figure 5. We find differences between policies; however, average scores over all criteria of an output have large standard deviations (std: 0.59). The differences in mean scoring of the three most frequent criteria between prompt variations are shown in Figure 6. The mean score of the most frequent criteria is impacted by variations. Particularly, the variations regarding the role impact the scoring of criteria. Even though we have different scores for different policies and for different variations, the high standard deviations are problematic for getting consistent results with assessments. The scoring, particularly if combined with its reasoning, allows differentiating how problematic a policy is. Among the 30 most frequently occurring criteria, 'Data sharing with third parties' is rated with the overall lowest mean of 2.34, and 'Data security' is rated highest with an overall mean of 3.30. Shortcomings regarding data sharing are consistent with related work (cf. Sec. 2.2).

**Surface-level metrics:** The length of an LLM assessment is relatively short (mean: 366.64; std: 64.74) and on average not very much affected by a changed seed, paraphrased prompt or variations. Hence, assessments are rather short and consistent in their length. The mean

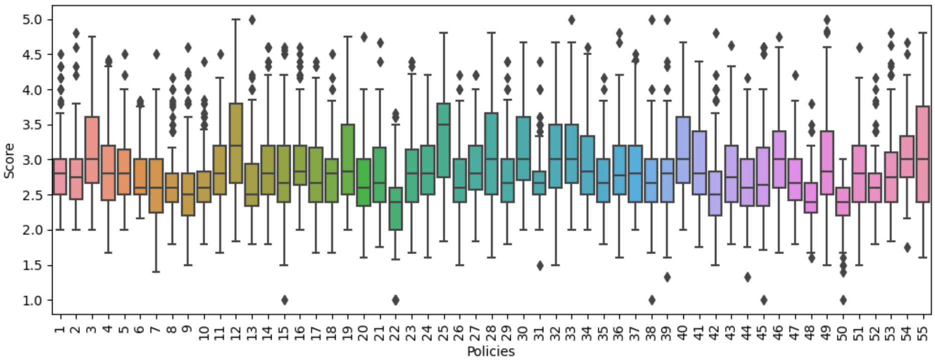

**Figure 5:** 5-point Likert-scores averaged over all criteria in an output aggregating over policies

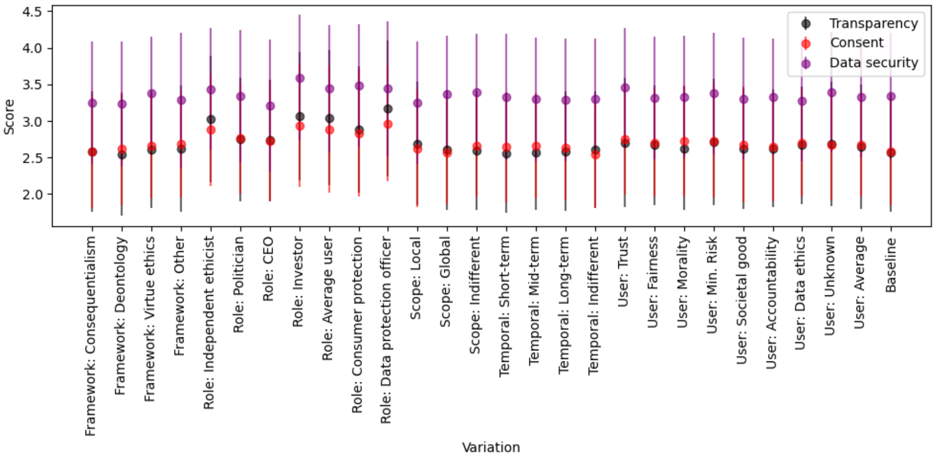

**Figure 6:** 5-point Likert-scores of the three most occurring criteria aggregating over prompt variations

FRE-score we find is 36.25. Scores vary notably (min: 15; max: 60; std: 5.59). Hence, readability and consistency in it could be improved upon.

**Sentiment:** Sentiment classifies how negative or positive associations transferred by text are on a scale of -1 (negative) to 1 (positive). Our assessments regarding sentiment show differences in mean sentiment between policies, as can be seen in Figure 7. We find little differences between variations and seeds/paraphrasing. This all applies to the sentiment assessed on the overall LLM output and sentiment for just the assessment step (step 2) of the LLM output. The overall mean of the sentiment of whole LLM outputs is slightly negative (mean: -0.11; std: 0.14), similar to sentiment of step 2 of the output (mean: -0.17; std: 0.20). These average values are consistent with the aggregated mean score for all criteria, which is 2.87 and the high aggregated standard deviation in scores, which is 0.86. This suggests that sentiment aligns with the scoring, gives foundation to the scoring, and shows consistency within the assessment process.

**Document embeddings:** The evaluation of document embeddings reveals high cosine similarity between policies as well as between variations. This means that assessments all

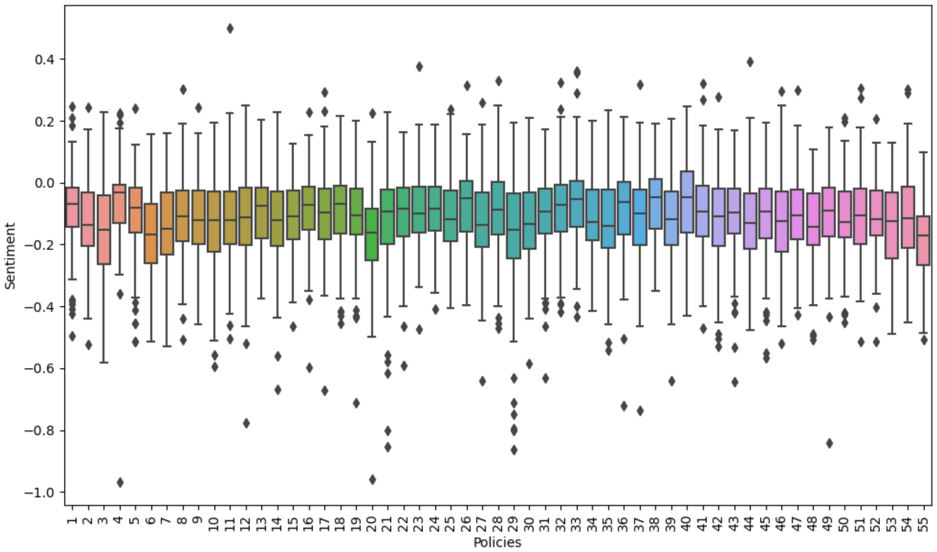

**Figure 7:** Sentiment scores over whole model outputs comparing between policies

go in a similar semantic direction. This may be a shortcoming regarding the specificity of an assessment. Regarding variations, the specified roles have the least similarity compared with the other variations, which are relatively homogeneous in their similarity. This shows that assigning roles as a variation can have an impact on assessments. Normal seed to normal seed or paraphrased to paraphrased are more similar than normal seed to paraphrased seed. Hence, there are limitations regarding robustness when paraphrasing the prompt.

**Assessing assumptions:** Cosine similarities of the assumptions' embeddings between variations can be grouped into the five assumptions that were requested:

- *Role:* Assigned roles in prompt variations are partially different (similarity approx. 0.25) compared with assumed roles when they are not given. The roles "average user" and "consumer protection officer" are relatively similar ($\sim$ 0.6) to other assumed roles. An exception to this is the baseline prompt and the variation considering the interests of an average user, which are less similar ($\sim$ 0.45).
- *User Interests:* Changed user interests are all assumed to be similar ($\sim$ 0.8) to all other variations, apart from the role of the "average user"($\sim$ 0.6). The latter leads to different assumptions being made about user interests.
- *Scope:* Scope is consistently not highly similar across all variations ($\sim$ 0.55).
- *Temporal Span:* All variations have similarities between 0.65 and 0.75 apart from the variation setting a short-term focus ($\sim$ 0.5).
- *Ethical Framework:* Assigning a specific framework in a variation leads to considerably lower similarity with other variations ($\sim$ 0.4) compared to similarities amongst other variations ($\sim$ 0.7). Noteworthy is also the role of the average user and the baseline, which are less similar (between 0.5 and 0.6) to all other variations not specifying an ethical framework.

We learn that the LLM defaults to a mid- to long-term assessment and is rather indifferent regarding user interests. It is inconsistent regarding scope. It typically employs what it calls data ethics or the legal frameworks as ethical frameworks if not specifically instructed. If not instructed otherwise, it assumes a role similar to that of a data protection officer, which is in line with it abiding by principles from the GDPR [6]. Variations between seeds occur irrespective of whether paraphrasing the prompt or not, and lead to cosine similarities of assumptions between 0.5 and 0.6. This means that assumptions generally vary considerably.

## 5. Discussion

**Addressing RQ1:** An indicator of the quality of our LLM-based assessment is that our findings are in line with related work (cf. Sec. 2.2). The criteria utilized in privacy assistants for evaluation are mostly contained in those identified by our approach to ethics assessment. Our low overall scores regarding data sharing reflect issues identified in prior research. In our results, the high occurrence rate of transparency as an assessment criterion underpins the relevance of informational fairness in privacy policies, as suggested by related work. The principles of the GDPR can also be seen as the basis for most of the frequently occurring criteria in LLM assessments. This in turn is also a limitation, as ethical reasoning should not, for the most part, only be grounded in legal compliance, but go beyond. This issue may be addressed by instructing the LLM not to focus on the legality of a policy. The scoring and the sentiment of the explanations in the assessment, even though they originally seemed to be consistent, are not strongly correlated. This means that the sentiment of the assessment cannot solely be taken as a foundation explaining the scoring. We find LLM outputs to be on average relatively robust to a changed seed, and to a lesser degree to a paraphrased prompt. However, we find that the consistency of model outputs still needs improvement. Between individual runs, we see too much variation in the scoring for it to be reliable. Even though we find assessment criteria and their scoring to change between policies, rather high similarities in their document embeddings lead to the conclusion that specificity to a policy could be improved upon. Overall, we find that an LLM-based assessment still needs some more refinement as well as improvements in the LLM itself to be viable. Our assessment of the quality is also just based on quantitative metrics, which need to be complemented by qualitative assessments to identify all shortcomings reliably.

**Addressing RQ2:** We find that variations in our prompting, which should have a strong impact on an assessment, had less impact on LLM outputs than expected. We found that only assigning roles is effective in changing LLM outputs. The effect of assigning roles can be particularly seen in the scoring of criteria. We find high cosine similarities in document embeddings across variations, which slightly drop when assigning a specific role. Apart from assumptions, the other metrics that we assessed are not considerably impacted by variations. Assumptions made by the model are, for the most part, highly similar and the model defaults to a generic assignment when not specifically instructed otherwise. This means that we mostly cannot reach an effective shift in perspective as set as an objective. The LLM tends to default to a legality-centric perspective. This is valuable knowledge about an LLM's representations regarding online privacy. As we would want more diverse perspectives for future assessments, a prompt may explicitly state that legality is assumed and this perspective should not be pursued.

**Limitations:** We deliberately chose to prompt the, at the time of writing, most frequently used model family with GPT-3.5. An assessment could improve in its quality by utilizing more capable models like GPT-4o [65], Claude Opus [5], or Google Gemini Ultra [4]. Also, investigating capable open-source models like Mixtral-8x22B [66], Llama 3 70B [67], or even models where more about the training data is known, is promising. Our investigated privacy policies are from web shops. Expanding the data set to a broader set of web services, particularly AI-based services, could be valuable for further testing of ethical depth and generalizing results. The utilized prompt leverages modern prompt engineering approaches. Highly complex prompting mechanisms and modifications to the prompt based on our findings (e.g. instruct that no legal aspects should be in outputs) may improve results. LLM chaining with iterative prompt feedback might be beneficial as well. Libraries like guidance [68] could enforce consistent structure of outputs. A future prompting may specify the use of plain language for more readable outputs.

The variations that can be made to the prompt are not limited to those we utilized. We chose a set of variables promising to strongly impact ethical assessment outcomes, as well as diverse and interesting variations for these variables. Assigning a role had the strongest impact on outcomes. Future research may investigate other variations.

The quality evaluation and comparison between the LLM's ethics assessments were handled automatically, utilizing quantitative metrics. Our metrics are overall moderately correlated and not redundant. An in-depth qualitative assessment of the LLM assessments by ethicists and data protection experts can give more detailed insights, but is beyond the scope of this paper. In future research, we aim to perform a thorough qualitative evaluation of our LLM ethics assessments. Moreover, investigating the potential for providers to hack such an LLM-based assessment with targeted variations of their policies without improving their privacy practices is an interesting prospect we want to pursue.

## 6. Conclusion

Investigating the quality of moral judgment of an LLM regarding online privacy is a relevant yet unexplored issue. We systematically utilize an LLM to assess ethics in privacy policies. We assess how giving the LLM a different context in the LLM prompt affects outputs. Furthermore, we evaluate the quality of LLM outputs based on a broad set of criteria.

Our results show that an LLM-based assessment of privacy policies is still limited in its consistency and specificity. However, the identified criteria are consistent with related work on fairness and ethics in privacy policies. We also find that only a change in role effectively changes the perspective of an assessment. Our other variations show little effect on outputs.

As a next step, we will conduct in-depth qualitative evaluations with ethicists, jurists, and data protection experts on LLM assessments to identify shortcomings and improve our ethics assessment approach. Our findings help guide the way toward automated privacy policy ethics assessment and, by doing so, toward fairly balancing provider and user interests by empowering the user. The moral judgment of LLMs regarding online privacy gains relevance as generative AI may be increasingly used in the creation of privacy policies.

## References

[1] A. Vigderman, G. Turner, The data big tech companies have on you, security.org (2024). URL: https://www.security.org/resources/data-tech-companies-have/, accessed Mar 2024.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] Gemini Team Google, Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[5] Anthropic, The Claude 3 model family: Opus, Sonnet, Haiku, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024. Accessed on Mar 2024.

[6] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union L119/1 (2016).

[7] B. Bartelt, E. Buchmann, Transparency in privacy policies, in: 12th International Conference on Building and Exploring Web Based Environments, 2024.

[8] S. I. Becher, U. Benoliel, Law in books and law in action: The readability of privacy policies and the GDPR, in: Consumer law and economics, Springer, 2021, pp. 179–204.

[9] V. Freiberger, E. Buchmann, Legally binding but unfair? Towards assessing fairness of privacy policies, arXiv preprint arXiv:2403.08115 (2024).

[10] J. W. DeCew, The scope of privacy in law and ethics, Law and Philosophy (1986) 145–173.

[11] B. Lundgren, A dilemma for privacy as control, The Journal of Ethics 24 (2020) 165–175.

[12] J. Mainz, An indirect argument for the access theory of privacy, Res Publica 27 (2021) 309–328.

[13] M. Blaauw, The epistemic account of privacy, Episteme 10 (2013) 167–177.

[14] A. Acquisti, L. Brandimarte, G. Loewenstein, Privacy and human behavior in the age of information, Science 347 (2015) 509–514.

[15] L. Wilsdon, Carissa Véliz, Privacy Is Power: Why and How You Should Take Back Control of Your Data, International Data Privacy Law 12 (2022) 255–257.

[16] D. Elliott, E. Soifer, Ai technologies, privacy, and security, Frontiers in Artificial Intelligence 5 (2022) 826737.

[17] A. Marmor, What is the right to privacy?, Philosophy and Public Affairs 43 (2015) 3.

[18] C. Benn, S. Lazar, What's wrong with automated influence, Canadian Journal of Philosophy 52 (2022) 125–148.

[19] J. P. Choi, D.-S. Jeon, B.-C. Kim, Privacy and personal data collection with information externalities, Journal of Public Economics 173 (2019) 113–124.

[20] K. Martin, Manipulation, privacy, and choice, North Carolina Journal of Law & Technology 23 (2022) 452.

[21] I. Harris, R. C. Jennings, D. Pullinger, S. Rogerson, P. Duquenoy, Ethical assessment of new

technologies: a meta-methodology, Journal of Information, Communication and Ethics in Society 9 (2011) 49–64.

[22] UN, Universal declaration of human rights, General Assembly resolution 217 A (1948). URL: https://www.un.org/sites/un2.un.org/files/2021/03/udhr.pdf.

[23] Council of Europe, European convention for the protection of human rights and fundamental freedoms (1950). URL: https://www.refworld.org/docid/3ae6b3b04.html.

[24] European Union, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market For digital services and amending Directive 2000/31/EC (Digital Services Act), Official Journal of the European Union L277/1 (2022).

[25] J. Gogoll, N. Zuber, S. Kacianka, T. Greger, A. Pretschner, J. Nida-Rümelin, Ethics in the software development process: From codes of conduct to ethical deliberation, Philosophy & Technology 34 (2021) 1085–1108.

[26] S. Krügel, A. Ostermaier, M. Uhl, ChatGPT's inconsistent moral advice influences users' judgment, Scientific Reports 13 (2023) 4569.

[27] T. Hagendorff, Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods, arXiv preprint arXiv:2303.13988 (2023).

[28] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI with shared human values, arXiv preprint arXiv:2008.02275 (2020).

[29] R. N. Zaeem, K. S. Barber, The effect of the GDPR on privacy policies: Recent progress and future promise, ACM Trans. Manage. Inf. Syst. 12 (2020).

[30] G. Malgieri, The concept of fairness in the GDPR: A linguistic and contextual interpretation, in: Proceedings of the 2020 Conference on fairness, accountability, and transparency, 2020, pp. 154–166.

[31] V. Belcheva, T. Ermakova, B. Fabian, Understanding website privacy policies—a longitudinal analysis using natural language processing, Information 14 (2023) 622.

[32] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, R. Ramanath, et al., Disagreeable privacy policies: Mismatches between meaning and users' understanding, Berkeley Tech. LJ 30 (2015) 39.

[33] H. Choi, J. Park, Y. Jung, The role of privacy fatigue in online privacy behavior, Computers in Human Behavior 81 (2018) 42–51.

[34] I. Pollach, A typology of communicative strategies in online privacy policies: Ethics, power and informed consent, Journal of Business Ethics 62 (2005) 221–235.

[35] J. Koetsier, Viral app faceapp now owns access to more than 150 million people's faces and names, 2019. URL: https://www.forbes.com/sites/johnkoetsier/2019/07/17/viral-app-faceapp-now-owns-access-to-more-than-150-million-peoples-faces-and-names/, accessed on 31 January 2023.

[36] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, J. Serna, Privacyguide: Towards an implementation of the eu GDPR on internet privacy policy evaluation, in: Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, 2018, pp. 15–21.

[37] R. Nokhbeh Zaeem, S. Anya, A. Issa, J. Nimergood, I. Rogers, V. Shah, A. Srivastava, K. S. Barber, Privacycheck v2: A tool that recaps privacy policies for you, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020,

pp. 3441–3444.

[38] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, L. C. Briand, Ai-enabled automation for completeness checking of privacy policies, IEEE Transactions on Software Engineering 48 (2021) 4647–4674.

[39] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[40] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection, arXiv preprint arXiv:2301.07597 (2023).

[41] A. Hamid, H. R. Samidi, T. Finin, P. Pappachan, R. Yus, Genaipabench: A benchmark for generative AI-based privacy assistants, arXiv preprint arXiv:2309.05138 (2023).

[42] C. Tang, Z. Liu, C. Ma, Z. Wu, Y. Li, W. Liu, D. Zhu, Q. Li, X. Li, T. Liu, et al., Policygpt: Automated analysis of privacy policies with large language models, arXiv preprint arXiv:2309.10238 (2023).

[43] P. Pałka, M. Lippi, F. Lagioia, R. Liepiņa, G. Sartor, No more trade-offs. gpt and fully informative privacy policies, arXiv preprint arXiv:2402.00013 (2023).

[44] K. Hu, ChatGPT sets record fastest growing user base, analyst note, Reuters (2023). URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, accessed Feb 2024.

[45] J. L. Espejel, E. H. Ettifouri, M. S. Y. Alassan, E. M. Chouham, W. Dahhane, GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts, Natural Language Processing Journal 5 (2023) 100032.

[46] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, arXiv preprint arXiv:2302.04023 (2023).

[47] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv preprint arXiv:2303.12712 (2023).

[48] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al., Scaling language models: Methods, analysis & insights from training gopher, arXiv preprint arXiv:2112.11446 (2021).

[49] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[50] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837.

[51] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in Neural Information Processing Systems 36 (2024).

[52] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, arXiv preprint arXiv:2211.01910 (2022).

[53] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A

systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[54] G. Kim, P. Baldi, S. McAleer, Language models can solve computer tasks, Advances in Neural Information Processing Systems 36 (2024).

[55] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, Q. Yang, Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–21.

[56] E. Lee, Control openai model behavior with seed: Step-by-step with code, https://drlee.io/control-openai-model-behavior-with-seed-step-by-step-with-code-9bba4e137a63, 2024. Accessed Feb 2024.

[57] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438.

[58] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, Advances in Neural Information Processing Systems 34 (2021) 27263–27277.

[59] R. Flesch, A new readability yardstick., Journal of applied psychology 32 (1948) 221.

[60] O. Guhr, A.-K. Schumann, F. Bahrmann, H. J. Böhme, Training a broad-coverage German sentiment classification model for dialog systems, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1620–1625.

[61] D. D. Otter, J. Medina, J. Kalita, A survey of the usages of deep learning for natural language processing, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 604–624.

[62] OpenAI, New embedding models and API updates, https://openai.com/blog/new-embedding-models-and-api-updates, 2024. Accessed on Mar 2024.

[63] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, L. Carin, Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 440–450.

[64] N. Baccouri, deep-translator: A python library for language translation, https://deep-translator.readthedocs.io/en/latest/README.html, 2020. Accessed Feb 2024.

[65] OpenAI, GPT-4 turbo, 2024. URL: https://openai.com/index/hello-gpt-4o/, accessed May 2024.

[66] Mistral AI, Mixtral 8x22b, 2024. URL: https://mistral.ai/news/mixtral-8x22b/, accessed May 2024.

[67] Meta, Llama 3, 2024. URL: https://llama.meta.com/llama3/, accessed May 2024.

[68] S. Lundberg, M. T. Ribeiro, H. Nori, Guidance: A guidance language for controlling large language models, Online, 2023. GitHub repository: https://github.com/guidance-ai/guidance.