Fairness Perceptions of Large Language Models

Benjamin Cookson, Soroush Ebadian, Nisarg Shah

University of Toronto
{bcookson,soroush,nisarg}@cs.toronto.edu

Abstract

Large language models (LLMs) are increasingly used for decision-making tasks where fairness is an essential desideratum. But what does fairness even mean to an LLM? To investigate this, we conduct a comprehensive evaluation of how LLMs perceive fairness in the context of resource allocation, using both synthetic and real-world data.

We find that several state-of-the-art LLMs, when instructed to be fair, tend to prioritize improving collective welfare rather than distributing benefits equally. Their perception of fairness is somewhat sensitive to how user preferences are represented, but less so to the real-world context of the decision-making task. Finally, we show that the best strategy for aligning an LLM's perception of fairness to a specific criterion is to provide it as a mathematical objective, without referencing "fairness", as this prevents the LLM from mixing the criterion with its own prior notions of fairness. Our results provide practical insights for understanding and shaping how LLMs interpret fairness in resource allocation problems.

Code —
https://github.com/bcookson/FairnessPerceptionsofLLMs

Extended version —
https://www.cs.toronto.edu/%7Enisarg/papers/llm-fairness.pdf

1 Introduction

The concept of fairness has captivated human thought for centuries, shaping the foundations of our core institutions, such as democracy, law, and healthcare. But what does fairness truly entail? While universally appealing, fairness is far from universally defined, and its interpretation often depends on the lens through which it is examined.

Fairness is a quintessential sociotechnical concept, explored extensively across disciplines. Philosophy deliberates the underlying principles of fairness, comparing Rawls' (1971) egalitarianism to Harsanyi's (1975) utilitarianism, and examining concepts such as desert, the right to a minimum, and fair equality of opportunity. Meanwhile, the machine learning literature takes a mathematical perspective on fairness, and often narrows its focus to deal with the most practically relevant issues such as mitigating race- or gender-based discrimination (Mehrabi et al. 2021). The fair division literature, at the intersection of economics and computer science, also takes a mathematical perspective, but formalizes individual and group fairness principles in an abstract resource allocation context devoid of specific attributes such as race or gender (Amanatidis et al. 2022; Shah 2023). Finally, studies on human perceptions of fairness provide a descriptive counterpart to these normative approaches to fairness (Grgic-Hlaca et al. 2018; Srivastava, Heidari, and Krause 2019; Saxena et al. 2019).

Recently, researchers have begun bridging these disciplinary silos by, e.g., applying the fairness criteria from fair division to machine learning (Balcan et al. 2019; Hossain, Mladenovic, and Shah 2020; Chen et al. 2019; Micha and Shah 2020; Kellerhals and Peters 2024; Caragiannis, Micha, and Shah 2024), or connecting fairness definitions in machine learning to those from moral and political philosophy (Binns 2018). However, a complete integration of these diverse perspectives has remained elusive, partly due to disciplinary boundaries and methodological divides.

Enter large language models (LLMs)! The advent of LLMs has been one of the most profound technological disruptions in recent years. These models are increasingly driving decision-making by sitting at the core of powerful AI agents that can autonomously act in the real world (OpenAI 2025). They exhibit social understanding gleaned from their pretraining on vast repositories of human-generated data, ethical considerations learned from academic research and post-training techniques such as reinforcement learning from human feedback (RLHF), and mathematical reasoning abilities. This unique blend of sociotechnical abilities has enabled breakthrough performance across domains such as healthcare, education, finance, engineering, and programming (Hadi et al. 2023). This makes LLMs particularly intriguing for exploring the multifaceted nature of fairness.

In this work, we investigate the perceptions of fairness exhibited by LLMs using fair division — specifically, fair allocation of indivisible goods to a set of agents — as our example domain. We choose fair division because there are several reasons that make LLMs aptly suited for adoption in real-world fair division applications. They are widely popular, easy to use, and often freely available. Further, their unique ability to understand contextual nuance can give them an edge over traditional algorithms (see Section 7 for further discussion). Our objectives are threefold:

1. *What is fair in the eyes of LLMs?* When LLMs are asked to be "fair", what metrics do they prioritize?
2. *What influences fairness perception?* How does an LLM's understanding of fairness depend on factors such as the nature of agents and goods involved, and the framing of the agents' preferences?
3. *To what extent can we steer LLMs?* Do LLMs have the reasoning abilities to optimize user-specified fairness criteria?

Under the first two objectives, our goal is to identify patterns that are common across different LLMs. These patterns may reflect perceptions of fairness encoded in the (largely common) pretraining datasets that the LLMs are trained with and, therefore, are likely to persist even as more capable LLMs are deployed in the future. Under the third objective, on the other hand, we seek to conduct an evaluation of the capabilities of the current state-of-the-art (SOTA) LLMs. While these models may soon be superseded, this portion of our work contributes a framework that can be used for continuous monitoring of the fairness capabilities of LLMs; thus, it contributes to the quickly growing literature in AI on conducting LLM evaluations on various dimensions such as safety, trustworthiness, and inclination to hallucinate (Guo et al. 2023; Chang et al. 2024; Chu, Wang, and Zhang 2024).

Our results. We evaluate fairness perceptions of three state-of-the-art families of LLMs — `Claude` (by Anthropic) (Anthropic 2024), `Gemini` (by Google) (Team et al. 2023), and `GPT` (by OpenAI) (Achiam et al. 2023) — using both synthetic data and real data from Spliddit.org. Using carefully designed prompts, we ask the LLMs to allocate a set of goods fairly to a set of agents based on (additive) valuations provided as part of the prompt, and compare their behavior to that of traditional algorithms based on (multiplicative) approximations to popular fairness and efficiency criteria, such as envy-freeness up to one good (EF1) and social welfare, with the goal of analyzing the fairness-efficiency tradeoff exhibited by LLM-generated allocations.

Our main takeaway is that when asked for fairness, LLMs value high social welfare, seemingly at the expense of envy-based notions of fairness. This can be seen visually in Figure 1. Although the different models vary in the exact approximations they achieve of the criteria we examine, all three models largely follow the same trends. Namely, in instances where it is impossible to achieve high approximations of EF1 and social welfare simultaneously, the LLMs opt for high social welfare.

To better understand what goes into the LLMs' allocation process, we investigate three variations in prompt design:

- **Context variation.** Whether the task is to allocate objects to people, heirlooms to siblings after a parent's death, or machines to teams in a corporate setting, the context appears to make little difference in how LLMs perform the allocation, at least when given only a brief description of the context.
- **Preference framing.** When agent preferences are provided grouped by goods (with each line specifying all agents' values for a given good), as opposed to grouped by agents (with each line specifying a given agent's values for all the goods), all models become a bit more efficient, with `Claude` and `Gemini` also becoming a bit fairer while `GPT` becomes a bit less fair. The effect size, however, is small.
- **Goal framing.** When LLMs are prompted to explicitly seek EF1, as opposed to simply maximizing "fairness", their tradeoff between fairness and efficiency changes slightly. Specifically, they tend to achieve higher EF1 approximations on average, although the overall trend of EF1 approximation degrading as it becomes harder to achieve EF1 and social welfare simultaneously still remains. We also prompt the LLMs using a purely combinatorial definition of EF1, dropping the language of "fairness" and "allocations of goods" entirely. Here, `GPT` and `Gemini` both do not see the same drop off in EF1 approximation as previously, while `Claude` still appears to prioritize efficiency over fairness in this setting.

In our analysis of the real-world data from Spliddit.org, we find that the LLMs do better at achieving good EF1 approximations than in comparable synthetic instances. While we primarily focus on analysis of the synthetic instances in this paper, we provide a detailed look at how all our tests performed on the Spliddit.org instances in the full version.

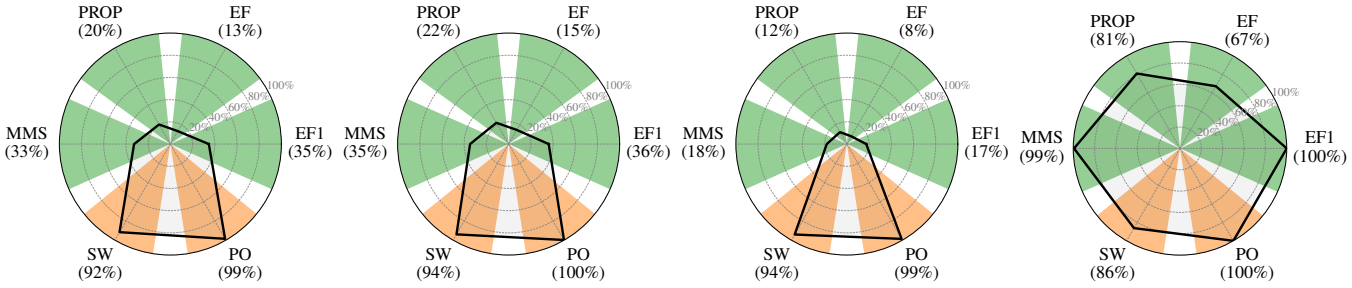
Although the main goal of our work is to dissect the interplay between EF1 and social welfare in the LLMs' perception of fairness, we also include a detailed summary of the aggregate performance of LLMs under a variety of fairness and efficiency metrics. These summaries, shown in Figure 1, give a high-level overview of exactly what the LLMs are prioritizing in their allocations, with the key takeaway again being that they seem to value efficiency more than fairness.

Related work. To the best of our knowledge, ours is the first work to explore the use of LLMs in fair division, with the exception of the simultaneous and independent recent work of Hosseini and Khanna (2025).

Hosseini and Khanna also investigate fairness perceptions of LLMs in the fair division context, but using a different approach. They primarily use 10 hand-crafted instances borrowed from the work of Herreiner and Puppe (2007), along with several newly developed instances. They evaluate how LLMs select allocations using five primary metrics: envy-freeness, equitability, egalitarian welfare, Pareto optimality, and social welfare. In contrast, our study uses tens of thousands of instances generated in a randomized fashion, with large variations in the number of agents and goods involved. This allows us to evaluate the fairness perceptions of LLMs at a large scale under a broader class of instances.

More broadly, our work is tangentially related to three lines of work.

LLM $\rightarrow$ social choice. Use of LLMs in the adjacent world of voting has been explored recently. When the candidates to be voted on are (policy) statements, LLMs have the remarkable potential of finding consensus candidates that are widely agreeable out of the vast space of possible statements. Bakker et al. (2022) design a system in which a fine-tuned set of LLMs generate statements that would be agreeable to large groups of humans and a traditional vot-



(a) `GPT-4o` (b) `Claude-3.5-Sonnet` (c) `Gemini-1.5-Pro` (d) `MNW`

Figure 1: Radar charts showing average approximation performance of LLMs and the MNW baseline across fairness (green) and efficiency (orange) metrics. Each axis corresponds to a criterion, with higher values (closer to the outer edge) indicating better approximation to that metric.

ing rule picks a single winning statement ("winner selection"), showing that such a system can outperform humans. Fish et al. (2024) develop this into *generative social choice*, which can design a representative slate of statements ("committee selection"); they use *generative queries*, which ask LLMs to find statements that would be agreeable to a specified target group of users. Small et al. (2023) discuss broader opportunities and risks of LLMs in deliberative platforms like Pol.is. Our work suggests extending LLM use to social choice more broadly, possibly to other problems such as matching and coalition formation.

Social choice $\rightarrow$ LLM. In the opposite direction, researchers have recently explored applying social choice concepts to the design of LLMs. For example, Zhong et al. (2024); Williams (2024) use the Nash social welfare in the RLHF stage of LLM training in order to get LLMs to proportionally represent the preferences of human annotators. Chakraborty et al. (2024) similarly use the egalitarian welfare to guide RLHF. It remains to be seen whether other social choice principles, such as envy-freeness or harm ratio (Ebadian, Freeman, and Shah 2024), can be applied to designing LLMs.

LLM evaluations. A growing literature evaluates LLMs on safety, trustworthiness, hallucination, reasoning, and more; see surveys by Guo et al. (2023), Chang et al. (2024), and Chu, Wang, and Zhang (2024). Several studies focus specifically on *fairness*, either broadly (Li et al. 2023) or in particular domains such as recommendations (Zhang et al. 2023) and ranking (Wang et al. 2024). The most common approach defines "fairness" as a quantification of bias exhibited toward predefined protected groups. Such bias can be mitigated through fine-tuning (Chung et al. 2024) or prompt engineering (Tamkin et al. 2023). We are primarily interested in how LLMs choose to *trade off* multiple desirable fairness and efficiency criteria, not whether they satisfy a given fairness metric. This examination of an LLM's "perception" of fairness is conceptually aligned with the works of Ji et al. (2025), Scherrer et al. (2023), and Dickerson et al. (2025), who also evaluate the moral reasoning of LLMs using subjective ethical questions and scenarios, thereby learning the moral frameworks embodied by LLMs and comparing them to those observed in humans.

2 Experimental Setup

In this section, we describe the fair division model at the heart of our experiments, the data and LLMs we use, our experimental setup, and our evaluation criteria.

Fair division model. For any $t \in \mathbb{N}$, let $[t] = \{1, 2, \ldots, t\}$. A fair division instance consists of a set of n agents $N = [n]$ and a set of m indivisible goods $M = [m]$. Each agent $i \in N$ has a valuation function $v_i : 2^M \rightarrow \mathbb{R}_{\geqslant 0}$, which represents the utility of agent i for each subset of goods. We focus on *additive* valuation functions, meaning $v_i(S) = \sum_{g \in S} v_i(\{g\})$ for all $S \subseteq M$ and $v_i(\emptyset) = 0$. With slight abuse of notation, we write $v_i(g) \coloneqq v_i(\{g\})$ for a single good $g \in M$. An allocation $A = (A_1, \ldots, A_n)$ is a partition of the set of goods M into n disjoint bundles, where $A_i \subseteq M$ is the bundle allocated to agent i, $A_i \cap A_j = \emptyset$ for all $i, j \in N$ with $i \neq j$, and $\cup_{i \in N} A_i = M$.

Synthetic data. For our synthetic data experiments, we build on the setup of Ebadian, Freeman, and Shah (2024). They draw agent utilities from the Dirichlet-multinomial distribution, defined as follows. First, a vector $\vec{p}$ is drawn uniformly from the $(m-1)$-simplex (i.e., from the Dirichlet distribution), where p_g represents the "market value" of good g. Then, for each agent i, a utility vector $(v_i(\{g\}) : g \in M)$ is independently drawn from the multinomial distribution with parameters T_i and $\vec{p}$, ensuring that $\mathbb{E}[v_i(\{g\})] = p_g \cdot T_i$ for each $g \in M$ and $\sum_{g \in M} v_i(\{g\}) = T_i$. They choose this distribution to induce a sharper tradeoff between fairness and efficiency than simply drawing all utilities i.i.d. We sample a different total utility T_i for each agent i independently from the uniform distribution over the set of integers $\{(50 - \lambda) \cdot m, \ldots, (50 + \lambda) \cdot m\}$. When $\lambda = 0$, our sampling process coincides with theirs. As λ increases, the total utility varies more across agents, thereby intensifying the tension between fairness (equal distribution of goods) and efficiency (allocating more to higher-utility agents).

We vary the number of agents $n \in \{2, 3, \ldots, 10\}$ (default $n = 5$), the number of goods $m \in \{n, 2n, \ldots, 5n\}$ (default $m = 3n$), and the total utility variation parameter $\lambda \in \{0, 5, \ldots, 40\}$ (default $\lambda = 20$). When varying a parameter, we fix the remaining two parameters to their default values, sample 200 instances, and plot the averages along with 95% confidence intervals.

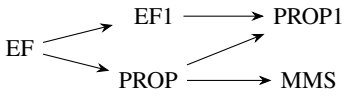


Figure 2: Relationships between fairness notions.

Spliddit data. We utilize real-world goods division instances from Spliddit.org. In these instances, the total utility of each agent for all goods is always 1000. Out of the 5295, we focus on the 4835 instances in which a positive Nash welfare is attainable (see Footnote 1), and show results averaged over these instances. These instances involve between 2 to 15 agents and 2 to 96 goods, with more than 99% of the instances involving at most 5 agents and at most 15 goods.

Evaluation: fairness criteria. The cornerstone notion of fairness in the fair division literature is *envy-freeness* (Gamow and Stern 1958; Foley 1967), which demands that no agent prefer the bundle allocated to another agent over their own bundle, i.e., $v_i(A_i) \geqslant v_i(A_j)$ for all $i, j \in N$. For indivisible goods, this is not always attainable. Hence, we measure its multiplicative approximation, and multiplicative approximations of its four widely studied relaxations: envy-freeness up to one good (EF1) (Budish 2011), proportionality (PROP) (Steinhaus 1948), proportionality up to one good (PROP1) (Conitzer, Freeman, and Shah 2017), and maximin share (MMS) (Budish 2011).

Figure 2 depicts the logical relationships between these criteria. In the interest of space, we define and present results for only EF1 approximation, deferring the definitions of and results for the rest to the full version.

- *EF1 approximation:* For an allocation A, this is the largest value $\alpha \in [0, 1]$ such that, for all $i, j \in N$ with $A_j \neq \emptyset$, there exists a good $g \in A_j$ such that $v_i(A_i) \geqslant \alpha \cdot v_i(A_j \setminus \{g\})$.

EF1 allocations are guaranteed to exist, and the maximum Nash welfare (MNW) algorithm (Caragiannis et al. 2019), which provably satisfies EF1, serves as our primary baseline. MMS allocations need not exist (Kurokawa, Procaccia, and Wang 2018), but a $\frac{3}{4} + \frac{3}{3836}$ approximation is achievable (Akrami and Garg 2024).

While the MMS approximations are quantitatively similar to EF1, the PROP1 approximations are quite different. This is due to subtleties about how our synthetic instances were generated, which we also explain in the full version. We emphasize that our results are pessimistic for fairness of LLMs, and our use of the weaker EF1 criterion instead of the stronger EF criterion only makes them stronger.

Evaluation: efficiency criteria. We use two prominent efficiency criteria from the literature: (utilitarian) social welfare (SW) and Pareto optimality (PO). Since maximizing SW implies PO, and PO approximation is at least as high as SW, we focus on SW here and defer the definition and similar results for PO to the full version.

- *SW approximation*: The (utilitarian) social welfare of an allocation A is the sum of agent utilities, i.e., $\mathrm{SW}(A) = \sum_{i \in N} v_i(A_i)$, and its SW approximation is its social welfare as a fraction of the highest possible social welfare, i.e., $\frac{\mathrm{SW}(A)}{\max_B \mathrm{SW}(B)}$.

Baseline algorithms. We compare the behavior of LLMs to that of three popular fair division algorithms:

- *Maximum Nash welfare* (`MNW`) (Caragiannis et al. 2019) returns an allocation that maximizes the Nash welfare, i.e., $\prod_{i \in N} v_i(A_i)$. This provably achieves EF1 and PO (Caragiannis et al. 2019), and is the state-of-the-art algorithm deployed to Spliddit.org due to its combination of fairness and efficiency guarantees.[1]
- *Round Robin* (`RR`) is an iterative algorithm that guarantees EF1 but not necessarily PO. Agents pick goods one by one in a cyclic fashion; specifically, in each round $k \in [m]$, agent $(k-1) \bmod n + 1$ is allocated her most preferred good among the ones remaining.
- *Maximum social welfare* (`MSW`) returns an allocation with the highest utilitarian social welfare. Under additive valuations, this simply allocates each good to an agent with the highest value for it. This is PO but does not guarantee any positive EF1 approximation.

Our primary focus is to investigate how LLMs behave when asked to be fair, and not to compare them with traditional algorithms. Hence, for clarity, we show only the MNW rule in the plots in the main body. In the full version, we compare LLMs to the other two baselines.

Large language models. We use three state-of-the-art commercial LLMs: `gpt-4o` (in short, `GPT`) from OpenAI, `claude-3.5-sonnet-20241022` (in short, `Claude`) from Anthropic, and `gemini-1.5-pro` (in short, `Gemini`) from Google.

In the full version, we report input/output token sizes, provide rough estimates of LLM costs for fair division, and show how costs scale with instance size.

Experiments and prompts. Each datum in our experiments is generated by sending a prompt to an LLM, which fully describes the fair division problem at hand, and asking the model to return an allocation. At a high level, all prompts have the same structure involving four components, whose designs we experiment with. We provide a summary below; details are available in the full version.

1) Context. First, the prompt describes the contextual scenario including the nature of agents and goods, which may affect LLMs' perceptions of fairness. We test three contexts:

- *Person/Object* (default): An abstract scenario with "objects" (goods) to be allocated to "people" (agents).
- *Sibling/Heirloom*: A "subjective" inheritance division scenario with "heirlooms" (goods) to be allocated to "siblings" (agents) following the passing of their parent.
- *Team/Machine*: An "objective" corporate scenario with "machines" (goods) to be allocated to "teams" (agents).

2) Goal. Next, the prompt describes the goal we want the LLM to achieve in the allocation it returns.

[1]The algorithm is more subtle in edge cases where all allocations yield zero Nash social welfare, but our experiments focus on instances that admit allocations with strictly positive utility for all agents (and thus positive Nash social welfare).

- *"Fair"* (default): The model is asked to allocate goods "fairly," without an explicit definition of fairness.
- *EF1 fair*: The model is instructed to find an EF1 allocation, with EF1 introduced as a fairness criterion and defined mathematically.
- *EF1 combinatorial*: Same as the EF1 fair prompt, but framed as a purely combinatorial problem — without reference to "fairness" or the context of allocating goods.

3) Preference framing. Next, we provide agents' valuations in one of two formats:

- *Person/Object* (default): For each agent, we provide a separate line listing their values for the m goods as integers, where the k-th value corresponds to good k:

```
Person 1: [1, 0, ...]   // m values
Person 2: [5, 8, ...]   // m values
```

- *Object/Person*: For each good, we provide a separate line listing the values of all n agents for that good as integers, where the i-th value corresponds to agent i:

```
Object 1: [1, 5, ...]   // n values
Object 2: [0, 8, ...]   // n values
```

4) Output format. We instruct the model to return a JSON object,[2] mapping each good to the index of its assigned agent. We explicitly instruct the model not to include any additional text or reasoning.

```
{ Object 1: 3,   // index (from 1 to n)
  Object 2: 2, ... }.
```

In Section 3, we compare all models and baselines using the default settings for the first three components. Then, in Sections 4 to 6, we vary each component individually while keeping the others at their default.

3 LLMs for Fair Division

The plots in Figure 3 highlight how the LLMs behave when prompted to simply find a "fair" allocation, with no further instruction on the problem context, or what "fairness" should entail. From these results, it is clear that all models generally prioritize efficiency (measured by approximation to SW) over fairness (measured by approximation to EF1). As a baseline, we first examine the performance of maximum Nash welfare (MNW), which is known to always return an EF1 allocation. This explains why, in figures (c) and (f), as λ, the utility variation parameter, increases, the SW approximation of the MNW allocations decreases sharply. When one agent has a much higher utility for all goods compared to another agent, achieving high social welfare requires allocating all goods to that agent, which goes against fairness. In contrast to MNW, we observe that as λ increases, the EF1 approximation of all three LLMs declines rapidly, while their SW approximation remains high.

[2]For `GPT` and `Gemini`, we use a built-in feature to restrict their output to the JSON schema. For `Claude` (and one Spliddit instance with 5 agents and 96 goods for which `Gemini` rejected the schema for being too long), we simply requested the models to follow the schema as part of the prompt, which they do very well.

Takeaways. In plots (a) and (d), and (b) and (e), we can also see how the EF1 and SW approximations of the models change as we vary n and m respectively. These represent increasing the complexity of the instances. As n increases, we can again see that MNW becomes worse at approximating SW. Intuitively, this is because having more agents raises the probability that one agent gets a much lower utility sum than some other agent, making it so that some goods must be inefficiently allocated to ensure fairness. Here we see that this worsening tradeoff causes the same behavior in the LLMs, which get worse at fairness in order to maintain efficiency.

In contrast, when m increases, we can see that MNW's SW approximation does not see significant change. It can be seen that when $m = 5$, the models all perform much better at fairness than when m is higher. Between $m = 5$ and $m = 10$, we see a steep drop-off in the level of fairness the models achieve, and an increase in efficiency. For all $m \geqslant 10$, the fairness and efficiency levels stay much more constant, with only small changes. In all our experiments, it appeared that when LLMs are provided with the same number of goods as there are agents $n = m$, their behavior was much different than when $m > n$, with the models being more likely to provide a *balanced* allocation, where all agents received the same number of items, even if that led to inefficiencies. This behavior is what explains the steep drop-off.

We also show the LLMs' performance against real-world instances from Spliddit.org in Figure 6. Interestingly, the LLMs perform significantly better on fairness for these real-world instances than for the synthetic ones, even when only looking at synthetic instances that are normalized ($\lambda = 0$), as the Spliddit.org instances are. We provide a detailed analysis of the Spliddit.org results in the full version.

4 Does the Allocation Context Matter?

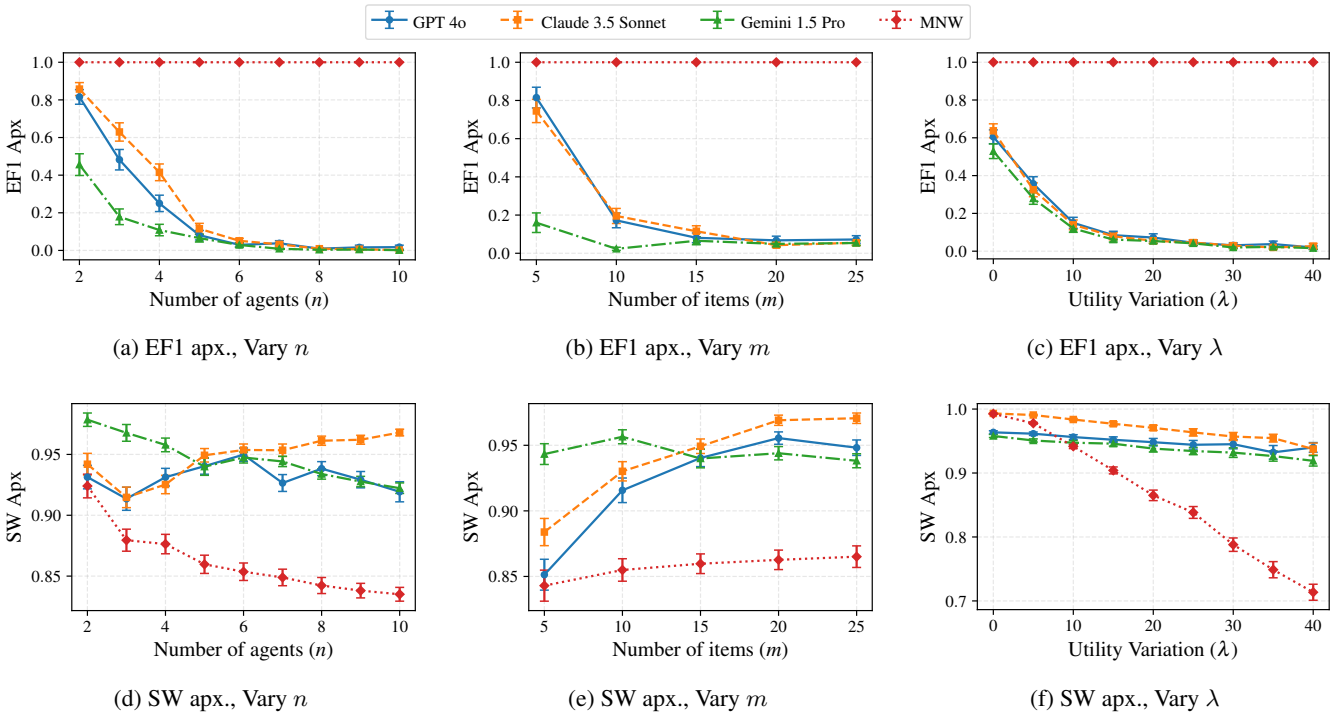
In this section, we examine whether the context of the allocation — be it abstract objects allocated to people, heirlooms divided among siblings following a parent's death, or machines distributed among corporate teams — affects how LLMs chart the fairness-efficiency tradeoff.

Takeaways. The results in Figure 4 show that contextual changes have little effect on the LLMs' fairness and efficiency behavior. In both the Siblings/Heirlooms and Teams/Machines scenarios, the models' approximations closely mirror those of the default setting, suggesting that small contextual shifts do not alter the tradeoffs these models make.

5 Does the Preference Framing Matter?

In this section, we test providing the preferences one agent at a time (Person/Object) versus one good at a time (Object/Person). This simply transposes the valuation matrix, which does not affect traditional algorithms' ability to access the values, but it may affect how an LLM interprets the preference data (just as it might affect a human too, at least in larger instances).

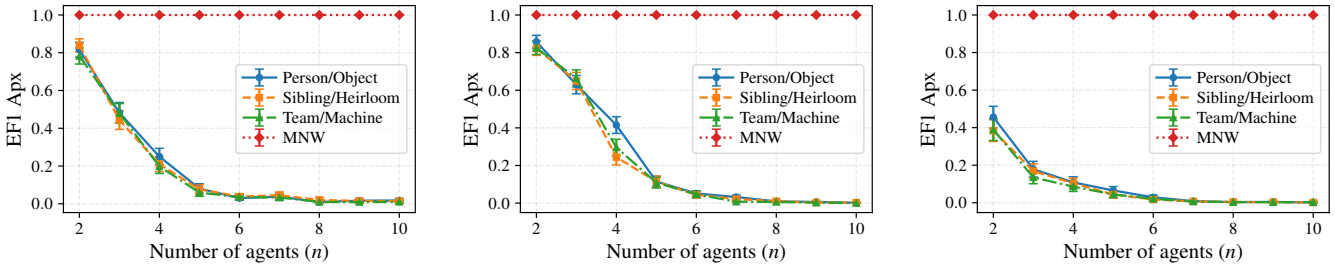
Takeaways. Figure 5 shows that how preferences are framed does affect 2 out of 3 models. For `Claude` and `Gemini`, the Object/Person framing leads to lower EF1 approximations but higher social welfare, suggesting a shift



(a) EF1 apx., Vary n (b) EF1 apx., Vary m (c) EF1 apx., Vary λ

(d) SW apx., Vary n (e) SW apx., Vary m (f) SW apx., Vary λ

Figure 3: Comparison of models for the default prompt by varying n, m, or λ.



(a) EF1 apx., `GPT` (b) EF1 apx., `Claude` (c) EF1 apx., `Gemini`

Figure 4: Comparison of models based on varying context with $m = 3n$ and $\lambda = 20$.

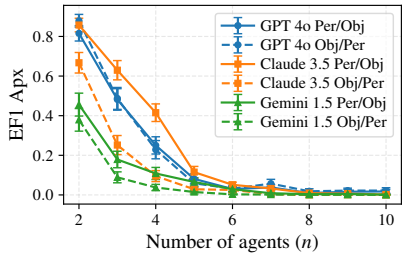
toward efficiency at the expense of fairness. One possible explanation is that presenting all agents' valuations for each object in a single list makes it easier for the LLM to compare utilities across agents and assign each object to the agent who values it most. This raises an important question: when LLMs fail to find a maximum social welfare allocation, is it due to a preference for fairness, or simply an inability to identify the optimal outcome? Interestingly, `GPT` appears largely unaffected by preference framing, with near-identical scores across both settings.

6 Steer LLMs or Let Them Be Free?

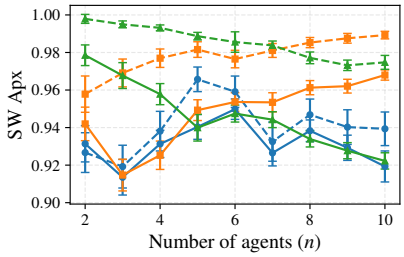
In this section, we evaluate how LLMs perform when specifically asked to aim for fairness, both by asking them directly to find an allocation that is EF1, and by providing them the instance as a purely combinatorial problem, and asking them to find an allocation with a property equivalent to EF1.

Takeaways. Figure 7 varies λ to control how difficult it is to satisfy fairness and efficiency simultaneously. For two of the three models (`GPT` and `Gemini`), we observe a very interesting difference between the "Fair" and "Combinatorial" versions of the EF1 prompt. Across all models, allocations from the fair prompt are consistently fairer than those from the default prompt. However, the EF1 approximations from these allocations still decline as λ increases. This seems to suggest that the LLMs still place some amount of priority on efficiency to the detriment of fairness, even when specifically instructed to prioritize EF1.

In contrast, for `GPT` and `Gemini`, the combinatorial prompt produces allocations whose fairness remains stable as λ increases. This suggests that when the task is framed as explicitly satisfying EF1 in a combinatorial setting, without the usual allocation context, LLMs deprioritize efficiency and focus more narrowly on the specified goal. When the

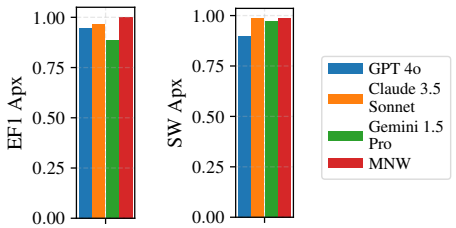


(a) EF1 apx.



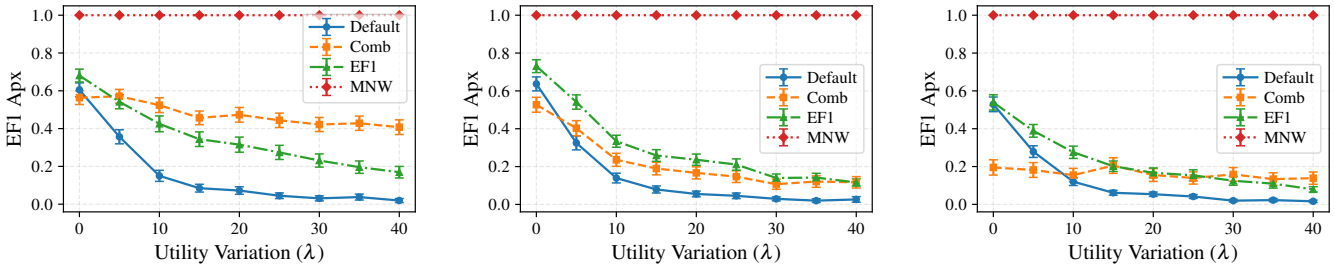
(b) SW apx.

Figure 5: Comparison of models under different input valuation framings with $m = 3n$ and $\lambda = 20$.



(a) EF1 apx. (b) SW apx.

Figure 6: Comparison of models on Spliddit.org.



(a) EF1 apx., `GPT` (b) EF1 apx., `Claude` (c) EF1 apx., `Gemini`

Figure 7: Comparison of models based on varying goals with $n = 5$ and $m = 15$.

allocation context is present, however, even explicit instructions to satisfy EF1 may be overridden by implicit reasoning about tradeoffs. Interestingly, `Claude` does not follow this pattern — it appears to favor efficiency over fairness even when the prompt strips away allocation context.

In the full version, we again observe that all prompt types degrade similarly as n increases, likely due to the increasing complexity of achieving fair and efficient allocations.

7 Discussion

While our work charts a rather large experimental landscape, it represents merely the tip of the iceberg in the exploration of LLM applications in fair division, let alone in the comprehensive evaluation of their fairness. There are many directions in which one can deepen our investigation.

Prompt engineering. While we experimented with variations of our base prompt, the possibilities of prompt engineering are vast, ranging from a mere reordering of the components to testing entirely novel task and goal descriptions.

Task generalization. We focused on the allocation of indivisible goods under additive valuations. Do our observations generalize to other fair division tasks, such as allocation of divisible goods, chore division, allocation under feasibility constraints, or allocating to agents with non-additive valuations? These tasks are notably more difficult, even for traditional algorithms, but that is precisely what may allow LLMs to be more competitive with traditional algorithms.

Better fairness evaluation. Our use of approximations to EF1, SW, and other fairness and efficiency notions are only proxy criteria; after all, if that is all that we care about, traditional algorithms already offer appealing trade-offs. The true power of LLMs lies in their unique sociotechnical understanding of fairness, so their efficacy must also be evaluated by human subjects (or, perhaps, other LLMs).

Leveraging contextual understanding. In Section 4, we found that a mere one-line description of the context does not significantly alter LLMs' behavior, but this may change if more context is provided. For example, an LLM performing inheritance division may lean towards optimizing fairness if there is a history of rivalry between the siblings, but optimizing efficiency if their relationships are largely harmonious. One can also follow the "generative social choice" approach (Fish et al. 2024; Bakker et al. 2022), whereby an LLM's contextual understanding is used to shape the problem instance (e.g., by detecting likely substitutes and complements among the goods based on their descriptions or likely cases of human error in providing valuations), but a traditional algorithm is used thereafter to hammer out the allocation, thereby achieving the best of both worlds.

Ethical considerations. Our work investigates the capabilities of existing models rather than introducing new ones, which limits the ethical risks involved. However, there remains a potential risk that our methodology could be used to "validate" a model as fair even when it exhibits significant unfairness along dimensions not captured in our analysis. We stress that our evaluation focuses on *specific* fairness properties in how LLMs allocate goods, and should not be interpreted as a comprehensive audit of fairness.

Acknowledgments

This research was partially supported by an NSERC Discovery grant and an NSERC-CSE Research Communities Grant. Researchers funded through the NSERC-CSE Research Communities Grants do not represent the Communications Security Establishment Canada or the Government of Canada. Any research, opinions or positions they produce as part of this initiative do not represent the official views of the Government of Canada.

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*.

Akrami, H.; and Garg, J. 2024. Breaking the 3/4 barrier for approximate maximin share. In *Proceedings of the 35th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 74–91.

Amanatidis, G.; Birmpas, G.; Filos-Ratsikas, A.; and Voudouris, A. A. 2022. Fair Division of Indivisible Goods: A Survey. In *Proceedings of the 31st European Conference on Artificial Intelligence (ECAI)*, 5385–5393.

Anthropic. 2024. Introducing Claude 3.

Bakker, M. A.; Chadwick, M. J.; Sheahan, H. R.; Tessler, M. H.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M. M.; and Summerfield, C. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 38176–38189.

Balcan, M.-F.; Dick, T.; Noothigattu, R.; and Procaccia, A. D. 2019. Envy-Free Classification. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 1238–1248.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, 149–159.

Budish, E. 2011. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy*, 119(6): 1061–1103.

Caragiannis, I.; Kurokawa, D.; Moulin, H.; Procaccia, A. D.; Shah, N.; and Wang, J. 2019. The Unreasonable Fairness of Maximum Nash Welfare. *ACM Transactions on Economics and Computation*, 7(3): Article 12.

Caragiannis, I.; Micha, E.; and Shah, N. 2024. Proportional fairness in non-centroid clustering. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 19139–19166.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Manocha, D.; Huang, F.; Bedi, A. S.; and Wang, M. 2024. MaxMin-RLHF: alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, 6116–6135.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Chen, X.; Fain, B.; Lyu, L.; and Munagala, K. 2019. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1032–1041.

Chu, Z.; Wang, Z.; and Zhang, W. 2024. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1): 34–48.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Conitzer, V.; Freeman, R.; and Shah, N. 2017. Fair Public Decision Making. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*, 629–646.

Dickerson, J. P.; Hosseini, H.; Khanna, S.; and Pierce, L. 2025. Who Gets the Kidney? Human-AI Alignment, Indecision, and Moral Values. *arXiv preprint arXiv:2506.00079*.

Ebadian, S.; Freeman, R.; and Shah, N. 2024. Harm ratio: A novel and versatile fairness criterion. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–14.

Fish, S.; Gölz, P.; Parkes, D. C.; Procaccia, A. D.; Rusak, G.; Shapira, I.; and Wüthrich, M. 2024. Generative Social Choice. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*, 985.

Foley, D. K. 1967. Resource Allocation and the Public Sector. *Yale Economics Essays*, 7: 45–98.

Gamow, G.; and Stern, M. 1958. *Puzzle-Math*. Viking.

Grgic-Hlaca, N.; Redmiles, E. M.; Gummadi, K. P.; and Weller, A. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 International World Wide Web Conference (TheWebConf)*, 903–912.

Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; Xiong, D.; et al. 2023. Evaluating large language models: A comprehensive survey. arXiv:2310.19736.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints.

Harsanyi, J. C. 1975. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American political science review*, 69(2): 594–606.

Herreiner, D. K.; and Puppe, C. 2007. Distributing indivisible goods fairly: Evidence from a questionnaire study. *Analyse & Kritik*, 29(2): 235–258.

Hossain, S.; Mladenovic, A.; and Shah, N. 2020. Designing Fairly Fair Classifiers Via Economic Fairness Notions. In *Proceedings of the International World Wide Web Conference (TheWebConf)*, 1559–1569.

Hosseini, H.; and Khanna, S. 2025. Distributive Fairness in Large Language Models: Evaluating Alignment with Human Values. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. Forthcoming.

Ji, J.; Chen, Y.; Jin, M.; Xu, W.; Hua, W.; and Zhang, Y. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1): 62–71.

Kellerhals, L.; and Peters, J. 2024. Proportional fairness in clustering: A social choice perspective. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 111299–111317.

Kurokawa, D.; Procaccia, A. D.; and Wang, J. 2018. Fair Enough: Guaranteeing Approximate Maximin Shares. *Journal of the ACM*, 64(2): article 8.

Li, Y.; Du, M.; Song, R.; Wang, X.; and Wang, Y. 2023. A survey on fairness in large language models. arXiv:2308.10149.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Micha, E.; and Shah, N. 2020. Proportionally Fair Clustering Revisited. In *Proceedings of the 47th International Colloquium on Automata, Languages and Programming (ICALP)*, 85:1–85:16.

OpenAI. 2025. Introducing Operator. https://openai.com/index/introducing-operator/. Accessed: 2025-12-15.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2nd AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 99–106.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36: 51778–51809.

Shah, N. 2023. Pushing the Limits of Fairness in Algorithmic Decision-Making. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 7051–7056. Early Career Spotlight.

Small, C. T.; Vendrov, I.; Durmus, E.; Homaei, H.; Barry, E.; Cornebise, J.; Suzman, T.; Ganguli, D.; and Megill, C. 2023. Opportunities and risks of LLMs for scalable deliberation with Polis. arXiv:2306.11932.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining (KDD)*, 2459–2468.

Steinhaus, H. 1948. The Problem of Fair Division. *Econometrica*, 16: 101–104.

Tamkin, A.; Askell, A.; Lovitt, L.; Durmus, E.; Joseph, N.; Kravec, S.; Nguyen, K.; Kaplan, J.; and Ganguli, D. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, Y.; Wu, X.; Wu, H.-T.; Tao, Z.; and Fang, Y. 2024. Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers. arXiv:2404.03192.

Williams, M. 2024. Multi-objective Reinforcement learning from AI Feedback. arXiv:2406.07295.

Zhang, J.; Bao, K.; Zhang, Y.; Wang, W.; Feng, F.; and He, X. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 993–999.

Zhong, H.; Deng, Z.; Su, W. J.; Wu, Z. S.; and Zhang, L. 2024. Provable multi-party reinforcement learning with diverse human feedback. arXiv:2403.05006.