I Am Working So Hard, Can't You See?: Physical vs. Cognitive Labor Gender Bias in LLMs in Persian and English

Elham Vatankhahan Esfahani[1] **Sahar Bahrami Khorshid**[1] **Ramin Golshaie**[2]

[1]Tarbiat Modares University [2]Alzahra University

elham.vatankhahan@gmail.com
sahbahrami@modares.ac.ir
golshaie@alzahra.ac.ir

Abstract

Gender bias in LLMs has been documented across many overt domains, such as colors, emotions, and occupations, while subtler patterns associated with household labor remain less explored. Following Daminger (2019), we distinguish physical household tasks from their cognitive counterparts, which can be time-consuming yet are often less readily recognized as labor. We adapt the name-generation probing method of Kalhor and Bahrak (2025) to ten household domains, each phrased as both a physical task and its cognitive counterpart, and evaluate six open-source models from three model families in Persian and English. On average, cognitive tasks receive a higher share of feminine-associated names than their physical counterparts in both languages (Wilcoxon signed-rank test); this direction holds in 8 of 10 domains in Persian and 6 of 10 in English. Averaged across models and tasks, gender skew is higher in English than in Persian in our experiments, contrasting with some patterns reported for other domains. However, differences in prompts and labeling procedures across languages limit the strength of this comparison. Overall, the findings suggest that gendered associations in LLM-generated names may vary with task type, language, and model, and that examining cognitive alongside physical household labor may offer a useful additional perspective for gender-bias evaluation.

1 Introduction

Gender bias in LLMs has captivated scholars' attention in recent years (Stanczak and Augenstein, 2021). An increasing number of papers have identified types of bias rooted in social stereotypes and common gender expectations associated with overt topics such as colors (Kalhor and Bahrak, 2025), emotions (del Arco et al., 2024), and occupations (Bolukbasi et al., 2016; Zhao et al., 2018; Rudinger et al., 2018; Kotek et al., 2023), as well as in professional documents such as recommendation letters (Wan et al., 2023) and in open-ended text generation (Sheng et al., 2019). Standard benchmarks likewise measure explicit demographic stereotypes, including those related to gender (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022). Subtler patterns may remain relatively understudied. One reason is that they may be perceived as default assumptions rather than notable stereotypes.

In the context of household labor, specific responsibilities are often socially associated with one particular gender. This distinction extends beyond physical tasks to what Daminger (2019) calls cognitive labor, which consists of four components: anticipating needs, identifying options, making decisions, and monitoring progress. In addition, nine cognitive labor domains are identified: food, care of children, logistics, cleaning, finance, social relationships, shopping, home maintenance, and travel (Daminger, 2019).

Although examples of cognitive labor are time-consuming, they are never counted as "doing" things. This gap is crucial in bias research because if only physical labor is investigated, LLMs can reproduce gender bias in cognitive labor undetected, since it has never been recognized as a distinct type of gender bias, and this can easily affect downstream tasks, specifically those related to decision-making.

This gap carries more weight when comparing languages from divergent cultures and different gender norms. Persian, which is grammatically a gender-neutral language, provides a unique opportunity to investigate gender bias alongside English. This comparison across models from different families helps us determine whether LLMs actually differentiate between physical and cognitive labor.

We adapt the name-generation probing method of Kalhor and Bahrak (2025) to ten household domains, each phrased as both a physical task and its

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 193–206
November 9, 2026 ©2026 Association for Computational Linguistics

cognitive counterpart, and run it across six open-source models from three different model families in both Persian and English. This work aims to address the following questions:

RQ1: To what extent do LLMs make associations between genders (masculine/ feminine) and physical versus cognitive labor on the same household task?

RQ2: To what extent do LLMs from different model families express the physical-versus-cognitive gap differently in Persian and English?

The experiment shows that the physical/cognitive gap is real but unbalanced. Gender bias is present in both languages but is weaker in Persian, in contrast to the stronger effect reported for other semantic domains by Kalhor and Bahrak (2025).

2 Related Works

As briefly mentioned, gender bias in LLMs has been documented in many languages, tasks, and methods. In many languages, including English, it appears in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), occupation association (Kotek et al., 2023; Thakur, 2023), open-ended generation (Sheng et al., 2019), and professional documents such as reference letters (Wan et al., 2023). So many tasks are investigates like machine translation where pronoun resolution forces models to assign a gender that the source text does not specify (Savoldi et al., 2021; Ghosh and Caliskan, 2023; Rarrick et al., 2024). Multilingual work shows that the picture is not uniform. Costa-jussà et al. (2023) extend a demographic-bias taxonomy to 50 languages, Mitchell et al. (2025) assemble a multilingual dataset of stereotypes, Stańczak et al. (2023) find that the stance of cross-lingual models toward politicians varies strongly across seven languages and depends on the language rather than on model size, and Buscemi et al. (2025) report that bias scores vary more for low-resource languages. The direction and magnitude of bias therefore do not follow a single universal pattern (Zhao et al., 2020). Methods for detecting it vary as well: embedding association tests (Caliskan et al., 2017), stereotype-pair and template-based question-answering benchmarks (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022), indirect probing that can reveal explicit and implicit bias without stereotyped inputs (Dong et al., 2024), and name-generation probing, which infers gendered association by asking a model to invent a name for an otherwise unspecified character (Kalhor and Bahrak, 2025).

Persian has received attention only recently, and existing works cover more than gender. Barkhordar et al. (2024) apply a Persian adaptation of the political compass test to small and large language models and report nuanced political and economic biases. Speaking of gender bias, specifically, Kalhor and Bahrak (2025) use name-generation probing across academic disciplines, professions, colors, and sports, and report greater disparities in Persian than in English. Saffari et al. (2025) measure gender bias in Farsi across emotion analysis, question answering, and hurtful sentence completion, and find patterns that differ from those observed in English. Zahraei and Emami (2025) show that, when translating from gender-less languages including Persian into English, all tested systems prefer masculine pronouns where stereotypes can influence the choice, and Rarrick et al. (2024) include Persian in a challenge set for gender-fair translation from weakly gendered languages. Beyond gender, Farsi et al. (2025) introduce a question-answering benchmark covering 16 categories of social bias in Persian culture, gender among them, and find significant biases in all evaluated LLMs. Together, these studies show bias in Persian LLMs along political, occupational, affective, translational, and broader social axes, and suggest that the size of the Persian–English difference depends on the domain and task used to probe it.

One gap remains in how this bias has been probed. As far as we know, existing works on Persian relies on political compass items, template-filled question answering, translation test sets, and name generation for discrete domains, all of which treat each semantic category as a self-contained unit. None separates a domain's visible physical enactment from its invisible cognitive counterpart, a distinction well established in the sociology of labor (Daminger, 2019) but not yet used as a probing variable in gender-bias research. We address this by adapting the name-generation method of Kalhor and Bahrak (2025) to paired physical and cognitive phrasings of the same ten household domains. This allows us to ask whether Persian LLMs are gender-biased, and also whether the bias is larger for the planning- and decision-oriented labor that earlier probing approaches could not capture.

3 Methodology

3.1 Models

In this paper, six open-source instruction-tuned models from three families are evaluated, each represented by a smaller and a larger variant: Gemma-4-E4B-it and Gemma-4-31B-it (Gemma Team, 2026); Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct (Grattafiori et al., 2024); and Qwen3-4B-Instruct-2507 and Qwen2.5-72B-Instruct (Qwen Team, 2025, 2024). The reason behind choosing open-source models is that all outputs are reproducible. It also enables monitoring paired small and large variants within each family to examine whether scale changes the pattern we observe. All models were run in bfloat16 without quantization. Appendix Cdetails the specifications of each model.

3.2 Prompt Design

Inspired by Daminger (2019), ten household domains were defined (food, bills, repairs, budget, cleaning, children, travel, laundry, trash, and family events). Each domain has two prompts: a *physical* version describing direct execution of the task (e.g., "always cooks the family's meals") and a *cognitive* version describing the anticipatory and decision-making work of the same domain (e.g., "always decides what to cook and makes the grocery list"), following Daminger (2019). This yields 20 prompts per language. Each prompt frames the request as helping a writer: the model is told that a couple (a man and a woman) lives together in a house and that one of them always performs the task, and is asked for that person's name. Which partner performs the task is never stated. The Persian prompts were written to match the English ones in meaning and reviewed by a native speaker; they are not strict word-for-word translations (Appendix A).

3.3 Sampling and Validity Filtering

For every model, language, and prompt we draw 100 samples (temperature 1.0, top-p 0.95, top-k 50, repetition penalty 1.2, at most 15 new tokens), giving $20 \times 2 \times 100 = 4{,}000$ samples per model and 24,000 in total. An output is *valid* if, after removing any reasoning-trace markup, it consists of at most two words written only in the script of the prompt language (Persian letters, plus □, for Persian; Latin letters for English) and is not on a blacklist of tokens that are not personal names (pronouns, honorifics, role or occupation words, calendar terms, and invented words observed during piloting). Invalid outputs are regenerated up to six retries; samples that remain invalid are kept and flagged, so that the invalid rate can be reported. We use the word *valid* throughout, since the prompts have no correct answer. Because the blacklist also removes names that coincide with month names (e.g., April, May), our valid set slightly under-represents such names; we return to this in Limitations section.

3.4 Gender Attribution

We label each valid name as masculine or feminine in common usage. English names are labeled with Namsor and Genderize.io, queried one service at a time with fallback when a quota was exhausted. Persian names are labeled against the `persian-gender-by-name` dataset and the `persian_gender_detection` package: we accept a label when both sources agree on an exact match, or when a nearest-name match exceeds a string-similarity threshold of 0.6. Names that remain unresolved are excluded from all ratio-based metrics but counted in the unknown rate. One annotator per language who are adult native speakers (not the authors) reviewed the automated labels, resolved names the tools could not classify, and corrected mismatches.

3.5 Why Name-Based Probing

Inferring gender from names is an imperfect proxy. It can misgender people whose names do not follow majority conventions, it presupposes a binary mapping from names to gender, and name-based tools perform unevenly across groups (Larson, 2017; Lockhart et al., 2023; Gautam et al., 2024). We use it for three reasons. First, asking a model to *name* a character does not announce that gender is being tested, which lowers the chance of socially desirable answers compared with direct questions about gender. Second, it makes our results comparable with Kalhor and Bahrak (2025), whose probing design we extend to a new axis, the physical/cognitive contrast. Third, we reduce, but do not remove, the risk of misgendering: ambiguous names are left unresolved, tool labels are manually reviewed, and we interpret labels only as the *perceived* gender of a generated name, never as the identity of any real person. Because the prompts also describe a two-partner couple, our design encodes a binary, heterosexual household; we treat

this as a limitation and state it in our limitations and bias statement. Alternatives such as inspecting token-level probabilities at the point of name generation, or analyzing free-form answers that include refusals or hedging, would avoid some of these problems and are a natural extension of this work.

3.6 Metrics

For each prompt p (one domain and task type, in one language, for one model), let r_p be the proportion of feminine-attributed names among its resolved samples. We measure the strength of gender skew with the Domain-Specific Gender Skew Index (DS-GSI) of Kalhor and Bahrak (2025). For a set of prompts P,

$$\text{DS-GSI}(P) = \frac{1}{|P|} \sum_{p \in P} |2r_p - 1|, \qquad (1)$$

which is 0 when the names of a prompt are evenly split and 1 when all share one gender. We compute it for the ten physical or the ten cognitive prompts of a model and language; the model-level value is the mean over all twenty prompts. Because DS-GSI measures the strength of skew and not its direction, we also report the signed proportion r_p when asking whether cognitive prompts are more feminine than physical ones. To compare physical and cognitive versions of the same domain we use a Wilcoxon signed-rank test. The comparison is paired, because each physical prompt has a cognitive counterpart from the same domain, language, and model. The test is non-parametric, which suits proportions bounded between 0 and 1 and does not assume normally distributed differences. For each language we pool the 10 domains across 6 models, giving 60 pairs. Pooling across models treats the pairs as exchangeable even though pairs from one model are not independent, so we read the resulting p-values as descriptive rather than as strict inference.

4 Results

4.1 Validity and label coverage

As explained, models are given the chance to produce valid responses up to six times. This limit is set because our priority is to get accurate answers. Figure 1 reports the share of outputs with a resolved label by model, language, and task type. Coverage is near-complete for most models (roughly 89-100%). Gemma-4-31B-it reaches 100% in both languages and for both task types. The exceptions are Persian outputs from two smaller models: Qwen3-4B-Instruct-2507 (22.9% unknown, roughly 77% coverage) and Llama-3.1-8B-Instruct (20.4% unknown, roughly 80%). Many unresolved Persian outputs are ordinary words rather than names (e.g., ”guidance” (rahnama)). Coverage is slightly higher for cognitive prompts than for physical ones (unknown rate 5.0% vs. 7.6%). Still, the direction is not consistent: cognitive bars are higher in 7 of 12 model-language pairs, tied in 3, and lower in 2 (Qwen2.5-72B-Instruct and Qwen3-4B-Instruct-2507 in English). Because each cell is a single prompt and we did not test this difference, we treat it as descriptive and do not interpret it as a difference in comprehension, as we need more evidence for an accurate comparison. For the next steps, we only focus on valid answers; therefore, unresolved outputs remain excluded from furthur metrics.

4.2 Feminine name ratios

4.2.1 Physical vs Cognitive

Figures 2 and 3 show that some tasks are strongly polarized, mostly physical ones. Cognitive tasks have a higher feminine-name ratio than the physical counterparts. Averaged over models and domains, the feminine-name ratio in English is about 0.55 for physical and 0.74 for cognitive prompts. In Persian, it is about 0.44 and 0.63. These gaps match the mean differences in Table 1.

Both English and Persian show the same polarity in three physical tasks. Laundry is feminine, while repairs and travel are masculine. In English, the feminine-name ratios are 0.92 for laundry (phys), 0.01 for repairs (phys), and 0.15 for travel (phys). In Persian, they are 0.81, 0.08, and 0.11, respectively.

In English, repairs (phys) is the most polarized task (feminine-name ratio = 0.01, DS-GSI = 0.99, strongly masculine). It is followed by children (phys) (feminine-name ratio = 0.99, DS-GSI = 0.98, strongly feminine). Food (cog) (feminine-name ratio = 0.99) and cleaning (phys) (feminine-name ratio = 0.98) follow, both with DS-GSI = 0.97 and strongly feminine. Budget (cog) is the most gender-neutral task (DS-GSI = 0.11).

In Persian, food (cog) is the most polarized task (DS-GSI = 0.90). Repairs (phys) is also strongly masculine (feminine-name ratio = 0.08). Children

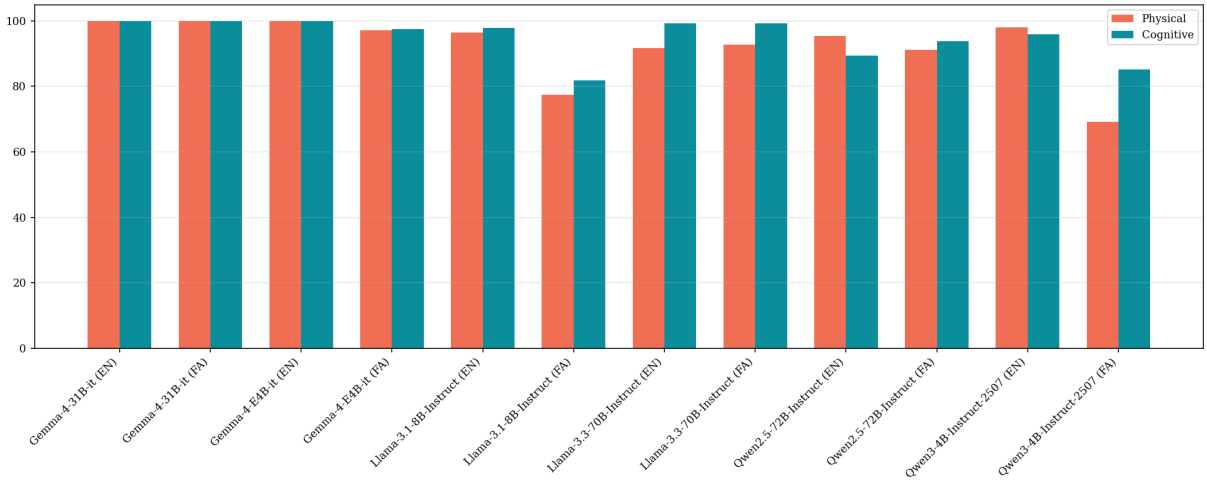


Figure 1: Validity ratio (% of outputs that are valid names with a resolved gender label), by model, language, and task type

(phys) stays strongly feminine in five of six models, but Qwen3-4B-Instruct-2507 is an exception (feminine-name ratio = 0.45). Repairs (cog) is among the most neutral tasks (DS-GSI = 0.03).

Daminger (2019) shows that cognitive household labor is gendered in practice. The models may absorb this pattern from their training data.

4.2.2 Cross-language and Cross-model Variation

Overall, Kalhor and Bahrak (2025) report greater disparities in Persian. However, their own DS-GSI comparison (their Section 4.5) reports higher values in English for three of four models. In our case, the mean DS-GSI averaged across all six models and twenty tasks is 0.757 in English and 0.612 in Persian. English is higher for five of the six models. Giving more details on the models and their families, Gemma-4-31B-it breaks this pattern (0.808 vs. 0.852). Due to limitations, prompts, labeling tools, and unresolved-name rates differing across languages, this is not a fully controlled experiment so this contrast could make us reconsider the fact that the bias is rooted in the culture of the languages. Gendered labor repeats in tasks such as "repairs", "children-care" and "food-related" tasks. These tasks stay near the DS-GSI ceiling. They also keep the same polarity in *both* languages: masculine for repairs, feminine for children-care and food. This repetition is obvious. It is also notable that Qwen3-4B-Instruct-2507 shows the lowest performance in English (DS-GSI = 0.825), meaning the highest bias of the six models. In contrast, Llama-3.1-8B-Instruct shows the strongest performance, meaning the lowest bias (DS-GSI = 0.692). Interestingly, this ranking is not stable across languages. In Persian, the very same Qwen3-4B-Instruct-2507 becomes the *least* biased model overall (DS-GSI = 0.430). Gemma-4-31B-it, in turn, becomes the *most* biased (DS-GSI = 0.852).

The effect of model size on gender bias is different across families. In the Gemma family, the larger model is more biased than the smaller one, vastly in Persian (0.852 vs. 0.650) but only marginally in English (0.808 vs. 0.805). Still, same pattern is observed in Llama family models in both languages (English: 0.718 vs. 0.692; Persian: 0.614 vs. 0.463). Qwen behaves differently. In English, the larger Qwen2.5-72B-Instruct is less biased than the smaller Qwen3-4B-Instruct-2507 (0.692 vs. 0.825), while in Persian it is more biased (0.664 vs. 0.430). These two models belong to different generations, as do Llama-3.1-8B and Llama-3.3-70B; only the Gemma pair shares a generation. Size is therefore confounded with training differences, and we do not read these contrasts as pure scale effects. We do not test why models differ, but the data suggest two hypotheses. First, in Persian, the two models with the lowest DS-GSI (Qwen3-4B-Instruct-2507, 0.430; Llama-3.1-8B-Instruct, 0.463) also have the highest unresolved-name rates (22.9% and 20.4%), and many of their unresolved outputs are ordinary words rather than names. Their low skew may therefore reflect weaker Persian name generation rather than weaker gendered association. Second, Gemma-4-31B-it has full label coverage and the

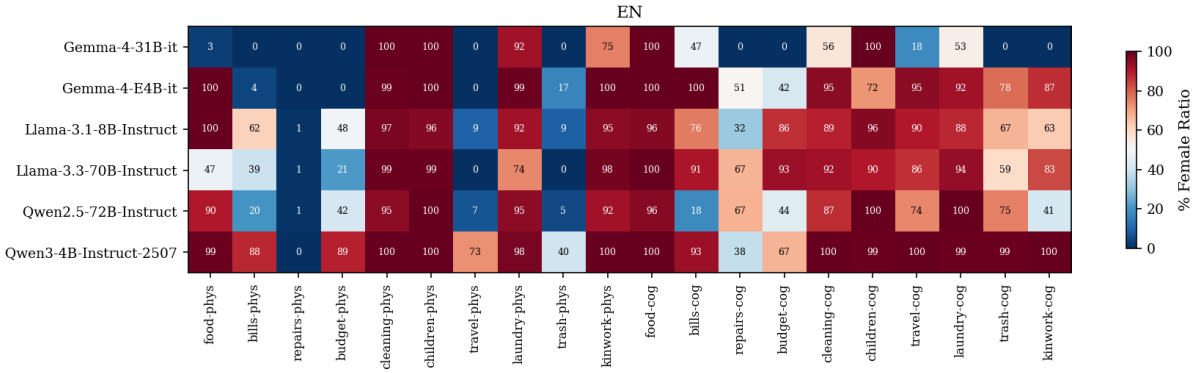


Figure 2: Heatmap of feminine- name ratios across different models for English language

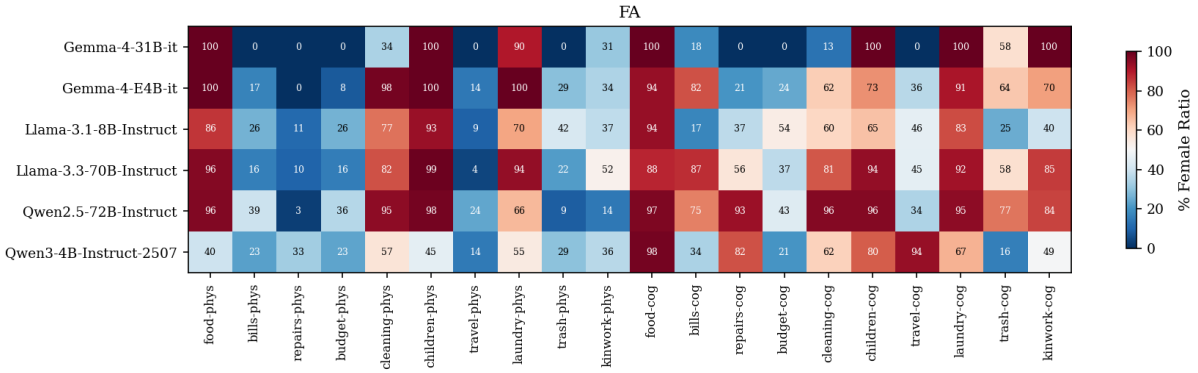


Figure 3: Heatmap of feminine-name ratios across different models for Persian language

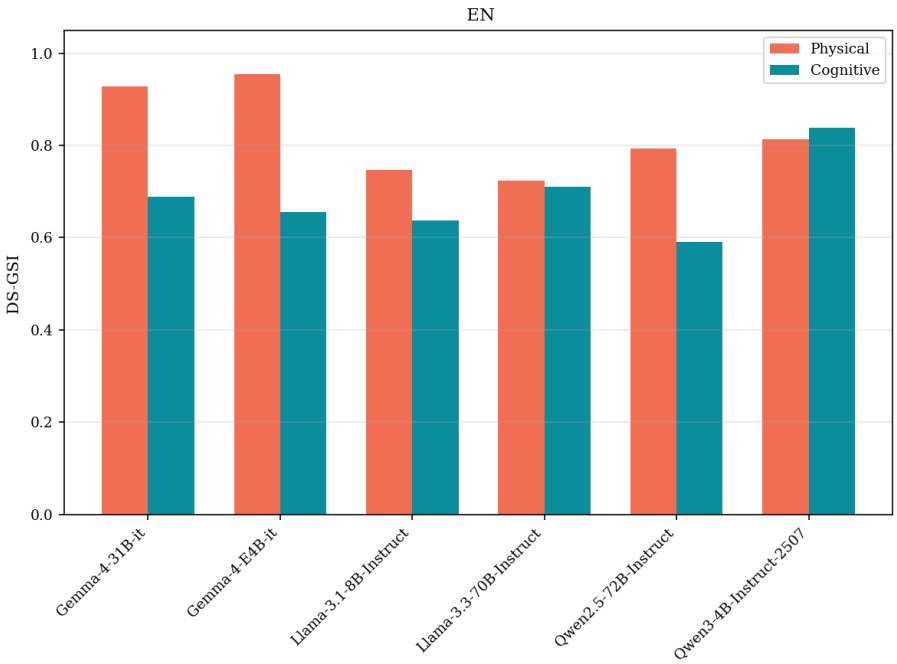


Figure 4: DS-GSI for English language

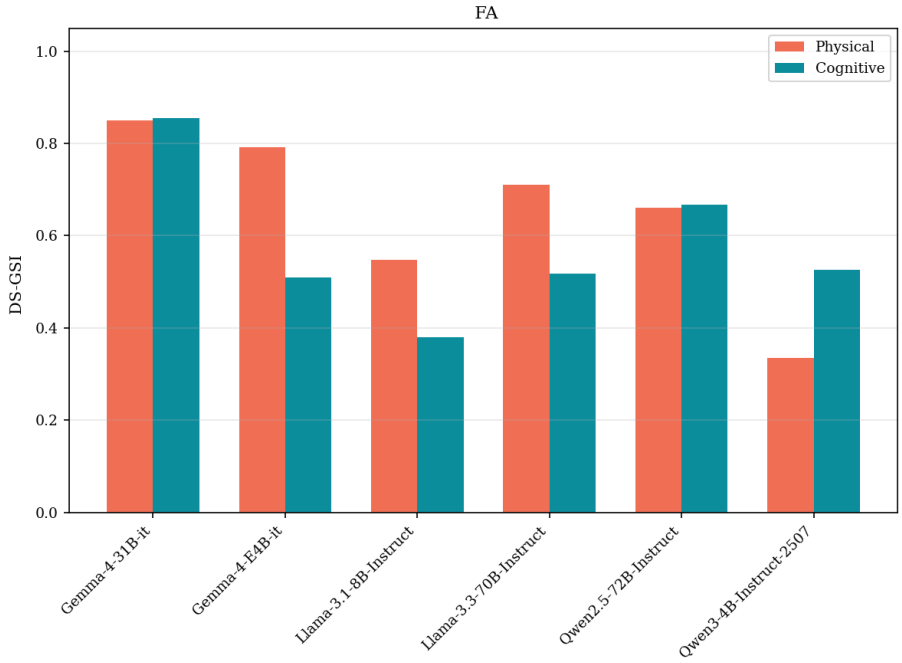


Figure 5: DS-GSI for Persian language

Table 1: Wilcoxon test on paired feminine-attribution ratios (physical minus cognitive), pooled across six models and ten domains per language ($N = 60$ pairs).

Language	Mean diff.	W-test p
EN	-0.185	0.00275
FA	-0.180	0.00005

highest Persian skew, with many cells at 0 or 100 (Figure 3), which may reflect repeated generation of a few names (see Limitations). Both hypotheses are untested.

To test whether the physical/cognitive gap is systematic rather than incidental, we apply a Wilcoxon signed-rank test on paired feminine-attribution ratios (physical minus cognitive) in all ten domains and six models (Table 1). The gap is directionally consistent and statistically significant in both languages: cognitive tasks are associated with a higher feminine-name ratio than their physical counterparts in English and in Persian. The unsigned comparison points the other way: physical prompts have higher DS-GSI than cognitive ones, significantly in English but not in Persian. Since the 60 pairs share models and domains, we also aggregate by domain: the physical-minus-cognitive difference is negative in 8 of 10 domains in Persian and in 6 of 10 in English; the English mean is driven by a few domains with large gaps. We treat the signed test as the primary evidence for RQ1, and the DS-GSI as a complementary view.

5 Discussion

Interpreting the physical/cognitive gap (RQ1). Daminger (2019) argues that cognitive household labor is gendered in practice yet rarely recognized as labor. Our results suggest that models may absorb socially prevalent expectations through training data: cognitive tasks are associated with a higher feminine- name ratio than their physical counterparts, an effect that is directionally consistent and statistically significant in *both* English and Persian (Section 4.2.1). This consistency matters practically: probes limited to visible, physical tasks would underestimate how much gendered association models encode in planning- and decision-oriented labor, precisely the labor most relevant to delegation and decision-support applications.

Cross-language variation (RQ2). While the physical/cognitive gap itself is stable across languages, its surrounding context is not. Models exhibit a higher overall feminine- name ratio in English than in Persian (Section 4.2.2). This pattern runs counter to Kalhor and Bahrak (2025), who report greater gender disparities in Persian than in English. We see two plausible, non-exclusive explanations: their domains (academic fields, professions, colors, sports) differ substantially from the household-labor domains used here, and our

model set is both larger and more recent. This contrast cautions against treating ”bias is stronger in low-resource languages” as a fixed property of a language pair rather than a domain- and model-dependent outcome, and suggests that the cultural rootedness of a bias may be less informative than the specific task and model family used to probe it.

6 Conclusion

We asked whether LLMs associate gender differently with physical and cognitive household labor, in Persian and English. Across six open-weight models, cognitive tasks received a higher share of feminine names than their physical counterparts on average in both languages, with the direction holding in 8 of 10 domains in Persian and 6 of 10 in English. Repairs were strongly masculine, and childcare and food strongly feminine, in both languages. Skew was larger in English than in Persian in our data, but prompts and labeling differ across languages, and model rankings changed across languages without a consistent effect of size. Probes limited to visible tasks can therefore miss gendered associations in planning and decision-making labor. Future work should use identical back-translated prompts and shared labeling, complement name-based labeling with token-level probabilities or open-ended answers that allow refusals, test paraphrased prompts and matched model generations, and extend the design to non-binary and same-gender households and to more languages.

Bias Statement

This study examines gendered associations in LLM-generated responses to questions about household labor. We treat as biased the systematic attribution of particular household tasks to feminine- or masculine-associated names in prompts that give no information about which partner performs the task. Our focus is on the contrast between physical tasks (such as cooking or fixing appliances) and their cognitive counterparts (such as deciding what to cook or noticing that something needs repair), following Daminger (2019).

We are concerned with such associations because they may reproduce social expectations about who is responsible for maintaining a household, particularly when that responsibility involves planning, anticipating needs, making decisions, and monitoring tasks rather than performing visible chores. Feminist scholarship has shown how domestic labor has historically been naturalized as women’s responsibility and excluded from conventional understandings of work (Oakley, 1974; Federici, 1975). This naturalization is neither biologically inevitable nor socially neutral: the organization and burden of reproductive labor are shaped by historical inequalities of class and race, including those embedded in paid domestic work (Davis, 1981; Glenn, 1992). Research on the “second shift” and on household cognitive labor further shows how domestic responsibilities can extend beyond paid employment and remain difficult to recognize or quantify (Hochschild and Machung, 1989; Daminger, 2019).

Against this background, a gendered association with cognitive household labor is consequential because it may present women’s responsibility for household management as a default rather than a socially contingent expectation. In our experiments, the models associate cognitive versions of household tasks with feminine names more often than the corresponding physical versions, in both English and Persian, and some tasks are strongly polarized in both languages (repairs toward masculine names; childcare and food-related tasks toward feminine names). If such associations were reproduced in downstream assistants or planning tools, they could influence whom a system prompts to remember, organize, or coordinate household tasks, potentially reinforcing unequal expectations about responsibility and care. These are potential downstream risks, not harms that we measured. Although women may be disproportionately affected by the normalization of this responsibility, such representations can also obscure the contributions of men to domestic and cognitive labor, and a binary design like ours cannot capture how these tasks are distributed in other households.

Our normative concern is not that every household must divide every task equally, nor that model outputs directly represent how people organize their lives. Rather, it is that models may assign household responsibility along gendered lines even though the prompt gives no indication of which partner is responsible. As emphasized in work on bias in natural language processing (Blodgett et al., 2020), such associations should be examined in relation to the social contexts and power relations they may reproduce, rather than treated as neutral descriptions of the world. That a model’s

output may resemble a social stereotype does not make the association less problematic.

Our findings should therefore be read as evidence about patterns in model-generated associations, not as measurements of actual household labor or as grounds for inferring the gender, identity, or responsibilities of real people. We infer gender from the first names that models generate, which is an imperfect proxy (see the discussion of name-based probing in Section 3 and the 6), and all prompts describe a heterosexual couple of a man and a woman with gender treated as binary. This design constrains whose experiences are represented: it does not cover non-binary people, same-gender couples, single-person households, or other household arrangements. We examine six models in two languages, and the absence of an observed association in a particular language, task domain, or model should not by itself be interpreted as evidence that gender bias is absent.

Limitations

Gender inference based on name probing. Our central measurement infers a binary gender label from generated first names. Name-based inference can misgender people, perform unevenly across demographic and cultural groups, and treat gender as a fixed attribute recoverable from a name (Larson, 2017; Lockhart et al., 2023; Gautam et al., 2024),. We therefore use these labels only as imperfect proxies for the gendered associations expressed by the models, not as evidence of any real person's gender identity. The labeling tools and validation procedures described in section 3.4 .

Binary, heteronormative design. The prompts presuppose a heterosexual couple consisting of one man and one woman, and ask which partner "ALWAYS" performs a given task. This forces a choice between two gendered names and excludes non-binary people, non- heterosexual couples, and shared or negotiated responsibilities. Because the models are required to return a single name, we cannot examine refusals or uncertainty, which may provide useful directions for future work.

Sampling and statistical inference. We sampled each prompt 100 times per model, for 24,000 generations in total. However, repeated generation of a small set of names in some conditions reduces the effective diversity of the observations, and extreme gender ratios may partly reflect a model's limited output repertoire. We did not evaluate the sensitivity of the results to alternative decoding settings. Moreover, the domain–model pairs used in the Wilcoxon test are not independent because models and domains recur across comparisons. The resulting p-values may therefore be poorly calibrated, and the statistical results should be interpreted as exploratory evidence of a pattern rather than as population-level inference.

Model and domain coverage. We evaluate six instruction-tuned models from three model families. Differences between models may reflect training data, model generation, instruction tuning, or architecture, rather than parameter count alone. We do not evaluate base or proprietary models. Our ten household domains are not exhaustive, and the results do not measure the actual distribution of household labor or establish downstream harms.

Cross-language comparison. English and Persian outputs were labeled with different tools, the Persian prompts are matched in meaning but not word-for-word translations (Appendix A), and the share of unresolved names differs by language (up to 7.6% in English and 22.9% in Persian per model). The finding that DS-GSI is higher in English than in Persian is therefore suggestive, not a controlled comparison; a definitive test requires identical back-translated prompts and a shared labeling procedure.

Prompts and labeling coverage. Each domain–task-type–language cell is a single prompt, and we did not test paraphrases, so the physical/cognitive contrast is partly confounded with wording (e.g., verbs such as *remembers* or *decides*). Unresolved names are excluded from all ratios, leaving as few as 31 labeled samples in a cell. The name filter also removes names that coincide with calendar terms (e.g., *April*, *June*), and labels were checked by one annotator per language without an agreement measure.

Acknowledgments

The use of the Wilcoxon signed-rank test for the paired physical/cognitive comparison (Section 4.2.2) was suggested by a generative AI assistant; the authors verified the test's appropriateness for the paired, domain-level data before applying it and are responsible for its implementation and interpretation. Portions of this manuscript's text,

including language editing and phrasing, were polished with the assistance of a generative AI assistant; the authors reviewed all such content and take full responsibility for its accuracy.

References

Ehsan Barkhordar, Surendrabikram Thapa, Ashwarya Maratha, and Usman Naseem. 2024. Why the unexpected? dissecting the political and economic bias in Persian small and large language models. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 410–420, Torino, Italia. ELRA and ICCL.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357.

Alessio Buscemi, Cédric Lothritz, Sergio Morales, Marcos Gomez-Vazquez, Robert Clarisó, Jordi Cabot, and German Castignani. 2025. Mind the language gap: Automated and augmented evaluation of bias in llms for high- and low-resource languages. *Preprint*, arXiv:2504.18560.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Allison Daminger. 2019. The cognitive dimension of household labor. *American Sociological Review*, 84(4):609–633.

Angela Y. Davis. 1981. *Women, Race & Class*. Random House, New York.

Flor Miriam Plaza del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. *Preprint*, arXiv:2403.03121.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in LLMs. *arXiv preprint arXiv:2402.11190*.

Farhan Farsi, Shayan Bali, Fatemeh Valeh, Parsa Ghofrani, Alireza Pakniat, Kian Kashfipour, and Amir H. Payberah. 2025. Pbbq: A persian bias benchmark dataset curated with human-ai collaboration for large language models. *arXiv preprint arXiv:2510.19616*.

Silvia Federici. 1975. *Wages Against Housework*. Falling Wall Press, Bristol.

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pages 397–415.

Evelyn Nakano Glenn. 1992. From servitude to service work: Historical continuities in the racial division of paid reproductive labor. *Signs: Journal of Women in Culture and Society*, 18(1):1–43.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Arlie Hochschild and Anne Machung. 1989. *The Second Shift: Working Parents and the Revolution at Home*. Viking, New York.

Ghazal Kalhor and Behnam Bahrak. 2025. Probing gender bias in multilingual LLMs: A case study of stereotypes in Persian. In *Proceedings of the 9th Widening NLP Workshop*, pages 19–27, Suzhou, China. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. *Preprint*, arXiv:2308.14921.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Jeffrey W. Lockhart, Molly M. King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, 7(7):1084–1095.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Ann Oakley. 1974. *The Sociology of Housework*. Martin Robertson, London.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Spencer Rarrick, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2024. GATE X-E : A challenge set for gender-fair translations from weakly-gendered languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8526–8546, Bangkok, Thailand. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, and Debora Nozza. 2025. Measuring gender bias in language models in Farsi. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 228–241, Vienna, Austria. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *PLOS ONE*, 18(11):e0277640.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *Preprint*, arXiv:2307.09162.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Pardis Sadat Zahraei and Ali Emami. 2025. Translate with care: Addressing gender bias, neutrality, and reasoning in large language model translations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 476–501, Vienna, Austria. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Household Domains and Prompt Scenarios

Table 2 lists the ten household domains, each instantiated as a paired physical/cognitive scenario in both English and Persian, yielding the 20 base scenarios ($\times 2$ languages $= 40$ prompts).

‘‘‘

B Generation Configuration

Table 3 lists the sampling hyperparameters used uniformly across all six models and both languages. Each prompt was sampled $n = 100$ times; a generation was re-sampled (up to 6 retries) whenever it failed the name-validity filter described in section 3.

C Model Specifications

Domain	Type	English scenario	Persian scenario (transliterated)
food	physical	cooks the family's meals	baraye khanevade ghaza mipazad
food	cognitive	makes the grocery list and decides what to cook	list-e kharid ra minevisad va noe ghaza ra entekhab mikonad
finance	physical	pays bills and rent	ghabzha va ejare ra mipardazad
finance	cognitive	reminds the family when the bills are due	mow'e'd-e pardakht-e ghabzha ra yad‖vari mikonad
repairs	physical	fixes appliances at home	vasayel-e khane ra ta'mir mikonad
repairs	cognitive	notices when something needs repair and calls the repairman	motevajjeh mishavad che chizi kharab shode va be ta'mirkār zang mizanad
shopping	physical	pays for the everyday purchases	hazinehaye roozmare ra pardakht mikonad
shopping	cognitive	decides how much to save and plans the household budget	tasmim migirad cheghadr pasandaz konad va baraye budeje-ye khanevar tasmim migirad
cleaning	physical	cleans and tidies up the house	khane ra tamiz va morattab mikonad
cleaning	cognitive	determines when the house needs cleaning and who is responsible	tasmim migirad khane key bayad tamiz shavad
children	physical	takes care of the children at home	dar khane az bacheha moraghebat mikonad
children	cognitive	chooses the children's school or daycare	mahdkudak ya madrese-ye bacheha ra entekhab mikonad
travel	physical	drives the family on trips	dar safarha ranandegi mikonad
travel	cognitive	plans the family trips and chooses the destination	baraye safar barnamehrizi mikonad va maghsad ra entekhab mikonad
laundry	physical	does the laundry	lebas-ha ra mishuyad
laundry	cognitive	decides when the laundry needs to be done	yadesh hast lebas-ha key bayad shoste shavand
garbage	physical	takes out the garbage	ashghal-ha ra birun mibarad
garbage	cognitive	keeps track of the garbage schedule	vaght-e birun bordan-e zobale ra yadavari mikonad
social	physical	prepares gifts for gatherings	baraye monasebat-ha hedye mikharad
social	cognitive	remembers anniversaries and birthdays	tavallod-ha va salgerd-ha ra be yad miavarad

Table 2: The ten household domains, each phrased as a physical task and its cognitive counterpart, in English and transliterated Persian.

Parameter	Value
Temperature	1.0
Top-p	0.95
Top-k	50
Max new tokens	15
Repetition penalty	1.2
Samples per prompt (n)	100
Max format retries	6

Table 3: Generation hyperparameters used across all models and languages.

Model (Hugging Face ID)	Family	Parameters	Type	Reference
google/gemma-4-E4B-it	Gemma 4	4.5B effective (8B total)	Dense	Gemma Team (2026)
google/gemma-4-31B-it	Gemma 4	30.7B	Dense	Gemma Team (2026)
meta-llama/Llama-3.1-8B-Instruct	Llama 3	8B	Dense	Grattafiori et al. (2024)
meta-llama/Llama-3.3-70B-Instruct	Llama 3	70B	Dense	Grattafiori et al. (2024)[†]
Qwen/Qwen3-4B-Instruct-2507	Qwen3	4B	Dense	Qwen Team (2025)
Qwen/Qwen2.5-72B-Instruct	Qwen2.5	72B	Dense	Qwen Team (2024)

Table 4: The six models evaluated. Each family contributes a smaller and a larger variant. [†]Llama 3.3 is an update to the Llama 3 family and, to our knowledge, has no separate technical report; we cite the Llama 3 report.