Do LLMs Have Habits in Self-Evaluation? Prototypical Self-Evaluation Enhanced Named Entity Recognition for LLMs

Supplemental Material

OVERVIEW

This document complements the main paper by visualizing the self–evaluation rating distributions that underlie PFILTER and by releasing the exact prompts used in all experiments. It is organized as follows. §II shows the four prototypes corresponding to the three entity types (PER, LOC, ORG) generated by six LLMs in our experiment. §III presents distributional analyses for four representative LLMs across PER/LOC/ORG under the single–turn (PFILTER-S) and dual–turn (PFILTER-D) pipelines (Task-1/Task-2 and Study Case-1/Study Case-2, respectively). §IV enumerates, verbatim, the prompts for base NER inference as well as the variants that incorporate label descriptions (LD) and few-shot (FS) exemplars, together with the prompts used for prototype induction and entity filtering in both PFILTER variants.

In the rating plots, correctly identified entities are shown in green and incorrectly identified entities in red. Across models and types, the empirical point clouds display clear clustering with directional separation: high-density regions for correct entities lie systematically farther from the origin (lower ratings) than those for incorrect entities. These regularities are consistent with Characteristic-1 and Characteristic-2 from the main text and provide the empirical basis for prototype induction ($P_1^{\mathrm{C}}, P_2^{\mathrm{C}}$ versus $P_1^{\mathrm{IC}}, P_2^{\mathrm{IC}}$) and for the ensuing filtering rule.

Conventions and reproducibility. We adopt the notation and evaluation protocol from the main paper. All prompts are provided exactly as used to facilitate replication; placeholders (e.g., for the target type) are explicitly marked where applicable. Figures and tables use the same label abbreviations (PER, LOC, ORG) throughout.

List of Sections:

I. Evaluation on Other LLMs

In the main manuscript, we demonstrate the efficacy of PFilter-S and PFilter-D on the NER task with Gemma 2 and GPT-4.1. We now extend this analysis to four additional LLMs, and the detailed results of PFilter and vanilla setting can be found in Tables I and II. Across both settings, PFilter consistently removes a substantial fraction ($i.e.$, $> 60\%$) of type-misidentified entities while keeping micro-F1 score stable or improving after filtering.

These findings indicate that PFilter is portable across model families and capacities, providing a reliable, drop-in mechanism for precision-oriented NER. In practical pipelines ($e.g.$, curating fine-tuning corpora by excising non-target types). PFilter reduces label noise without sacrificing overall utility, thereby facilitating the construction of high-quality training data at scale.

Setting	CoNLL 2003 (Test)			CoNLL 2003 (Validation)			WNUT17	
	PER	LOC	ORG	PER	LOC	ORG	PROD	GROP
GPT-3.5 Turbo	71.76	55.40	52.54	69.59	56.12	45.08	47.19	43.86
w/ PFilter-S	77.98	60.84	57.33	75.18	61.99	49.51	52.12	49.24
w/ PFilter-D	77.40	60.96	57.17	75.24	61.59	49.27	52.02	49.86
Max Δ F1	6.22↑	5.56↑	4.79↑	5.65↑	5.87↑	4.43↑	4.10↑	6.00↑
Llama 3.1 (8B)	60.86	52.15	52.47	61.65	51.16	48.83	39.01	36.39
w/ PFilter-S	66.13	57.81	55.99	67.21	57.13	52.81	44.14	40.31
w/ PFilter-D	66.64	58.96	55.39	67.15	58.12	53.01	44.82	39.77
Max Δ F1	5.78↑	6.81↑	3.52↑	5.56↑	6.96↑	4.18↑	5.81↑	3.92↑
Phi 4 (14B)	66.82	56.35	55.64	67.30	56.35	49.96	41.40	39.13
w/ PFilter-S	72.08	61.86	59.23	72.61	62.22	54.03	47.39	44.51
w/ PFilter-D	72.35	62.12	59.66	71.23	62.12	53.56	48.65	44.99
Max Δ F1	5.53↑	5.77↑	4.02↑	5.31↑	5.87↑	4.07↑	7.25↑	5.86↑
Qwen2.5 (7B)	69.52	58.07	53.65	71.81	61.33	48.83	46.86	36.99
w/ PFilter-S	75.36	63.34	57.62	76.81	65.84	53.70	51.96	41.16
w/ PFilter-D	75.96	63.13	58.22	77.38	65.64	54.33	51.11	41.50
Max Δ F1	6.44↑	5.27↑	4.57↑	5.57↑	4.51↑	5.50↑	5.10↑	5.51↑

TABLE I: F1 scores of LLMs with PFilter-S and PFilter-D.

Setting	CoNLL 2003 (Test)			CoNLL 2003 (Validation)			WNUT17	
	PER	LOC	ORG	PER	LOC	ORG	PROD	GROP
GPT-3.5 Turbo	384	1285	585	300	1018	1097	342	194
w/ PFilter-S	253	712	342	190	618	764	198	152
w/ PFilter-D	288	734	365	231	633	717	227	131
Max Δ TM (%)	34.1↓	44.5↓	41.5↓	36.7↓	39.3↓	34.6↓	42.1↓	32.5↓
Llama 3.1 (8B)	1379	1957	1752	1390	2143	2097	692	809
w/ PFilter-S	970	798	784	701	1287	993	345	701
w/ PFilter-D	912	784	801	736	1264	1036	369	688
Max Δ TM (%)	33.9↓	59.9↓	55.3↓	49.6↓	41.0↓	52.6↓	50.1↓	15.0↓
Phi 4 (14B)	749	1170	1454	815	913	1220	417	143
w/ PFilter-S	532	589	594	322	603	649	358	124
w/ PFilter-D	510	564	580	336	612	677	335	102
Max Δ TM (%)	31.9↓	51.8↓	60.1↓	60.5↓	34.0↓	46.8↓	19.7↓	28.7↓
Qwen2.5 (7B)	1252	1700	1814	1171	1339	1416	497	394
w/ PFilter-S	584	601	1273	734	477	582	239	311
w/ PFilter-D	564	613	1231	702	492	557	258	305
Max Δ TM (%)	55.0↓	65.4↓	32.1↓	40.1↓	63.3↓	60.7↓	48.0↓	22.6↓

TABLE II: Number of type misidentification entities for LLMs with PFilter-S and PFilter-D.

II. Generated Prototypes

Tables III to V catalog the four rating prototypes used by PFilter-S for the three target types (PER, LOC, ORG) across eight LLMs; analogously, Tables VI to VIII report the prototypes for PFilter-D. Prototypes are induced per (model, type) pair from the CoNLL 2003 test set and then held fixed. We further validate their cross-corpus robustness on additional datasets. At inference time, PFilter consults only the incorrect-class prototypes, P_1^{IC} and P_2^{IC}, to decide whether to filter a candidate entity. The correct-class prototypes, P_1^{C} and P_2^{C}, are used as diagnostics to verify that the empirical rating distribution satisfies Characteristic-2, which underpins the filtering rule.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(9.6039)	(4.9645)	(3.6636)	(1.3815)
GPT-4.1	(9.7314)	(4.8364)	(3.8153)	(1.3224)
Gemma 2 (9B)	(9.9061)	(8.7173)	(4.9333)	(2.7281)
Llama 3.1 (8B)	(8.3076)	(3.4545)	(5.4468)	(2.2153)
Phi 4 (14B)	(9.8273)	(7.4580)	(4.4935)	(2.4384)
Qwen2.5 (7B)	(7.9856)	(6.0889)	(4.7975)	(3.2667)

TABLE III: Prototypes of different LLMs on "*Person*" type for PFilter-S.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(9.6428)	(8.2093)	(4.5725)	(1.0876)
GPT-4.1	(9.8725)	(8.4633)	(4.4793)	(1.3145)
Gemma 2 (9B)	(9.7899)	(9.3664)	(3.9333)	(2.4020)
Llama 3.1 (8B)	(8.9589)	(4.2787)	(6.7284)	(1.7612)
Phi 4 (14B)	(9.5382)	(7.3519)	(4.9505)	(1.6889)
Qwen2.5 (7B)	(8.2747)	(7.3624)	(4.1481)	(2.5676)

TABLE IV: Prototypes of different LLMs on "*Location*" type for PFilter-S.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(8.6371)	(5.3992)	(2.3745)	(1.3847)
GPT-4.1	(8.8393)	(5.1252)	(2.5004)	(1.2564)
Gemma 2 (9B)	(9.7124)	(9.0612)	(5.5135)	(4.2967)
Llama 3.1 (8B)	(7.4827)	(4.2987)	(4.8703)	(2.2949)
Phi 4 (14B)	(9.4954)	(7.1223)	(5.0256)	(2.4685)
Qwen2.5 (7B)	(7.7872)	(6.2685)	(4.1461)	(3.1566)

TABLE V: Prototypes of different LLMs on "*Organization*" type for PFilter-S.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(9.03226, 2.78226)	(7.47807, 2.21491)	(9.59308)	(8.35366)
GPT-4.1	(9.73116, 8.42964)	(9.25414, 7.55691)	(9.83296)	(9.10211)
Gemma 2 (9B)	(9.26496, 7.04273)	(7.88050, 7.27044)	(9.88614)	(8.56235)
Llama 3.1 (8B)	(7.93590, 8.32051)	(7.07344, 7.86723)	(8.36923)	(7.39065)
Phi 4 (14B)	(9.56202, 4.16835)	(6.87844, 2.50114)	(9.62546)	(6.89441)
Qwen2.5 (7B)	(8.51605, 4.71101)	(5.73308, 2.47368)	(8.88040)	(6.20823)

TABLE VI: Prototypes of different LLMs on "*Person*" type for PFILTER-D.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(9.59764, 4.21212)	(8.92817, 2.37845)	(9.74131)	(9.24288)
GPT-4.1	(9.88764, 8.71910)	(9.52289, 7.99638)	(9.93170)	(8.63148)
Gemma 2 (9B)	(9.63063, 8.03153)	(9.28595, 7.41624)	(9.84418)	(9.26852)
Llama 3.1 (8B)	(8.66911, 8.57352)	(7.77292, 7.41048)	(8.64826)	(7.74401)
Phi 4 (14B)	(9.16204, 4.19806)	(7.64545, 3.18909)	(9.12296)	(7.65688)
Qwen2.5 (7B)	(8.79004, 8.38961)	(7.68004, 6.20642)	(8.88574)	(6.78327)

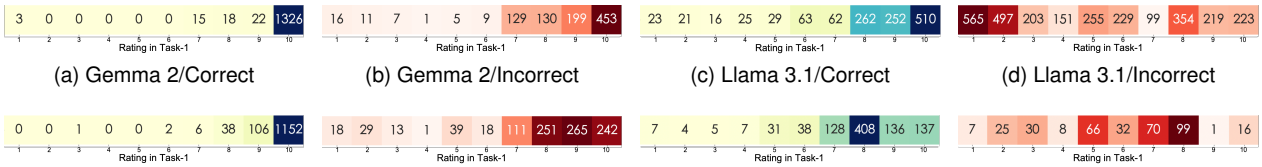
TABLE VII: Prototypes of different LLMs on "*Location*" type for PFILTER-D.

	P_1^{C}	P_1^{IC}	P_2^{C}	P_2^{IC}
GPT-3.5 Turbo	(9.18993, 3.96103)	(7.78612, 4.19653)	(9.36150)	(7.88889)
GPT-4.1	(9.61205, 8.10222)	(9.34421, 8.16593)	(9.70992)	(8.92703)
Gemma 2 (9B)	(9.39291, 8.19973)	(8.50253, 7.47462)	(9.54446)	(8.55773)
Llama 3.1 (8B)	(8.35119, 8.03869)	(8.00242, 8.23486)	(8.21587)	(7.70553)
Phi 4 (14B)	(8.07482, 4.37841)	(6.72355, 3.25160)	(7.71059)	(6.81625)
Qwen2.5 (7B)	(8.04601, 3.15752)	(6.72939, 2.93480)	(8.08561)	(6.62903)

TABLE VIII: Prototypes of different LLMs on "*Organization*" type for PFILTER-D.

III. Rating Distribution

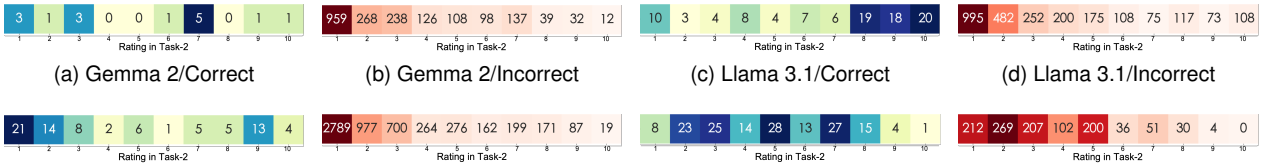
Figures 1 to 6 visualize the rating distributions for four representative LLMs across PER/LOC/ORG under Task-1 and Task-2 in PFilter-S. Analogous plots for PFilter-D are shown in Figures 7 to 12 for Study Case-1 and Study Case-2. Across models and types, the empirical clouds exhibit clear clustering structure with directional separation: the high-density regions of correctly identified entities (green) lie systematically farther from the origin (lower ratings) than those of incorrectly identified entities (red). These regularities are consistent with Characteristic-1 and Characteristic-2 introduced earlier, providing the empirical basis for prototype induction ($P_1^{\mathrm{C}}, P_2^{\mathrm{C}}$ vs. $P_1^{\mathrm{IC}}, P_2^{\mathrm{IC}}$) and for the ensuing filtering rule used by PFilter.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

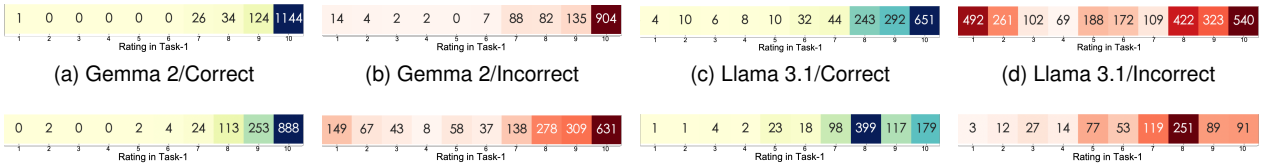
Fig. 1: Distribution of ratings in Task-1 of PFilter-S for four LLMs on "*Person*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

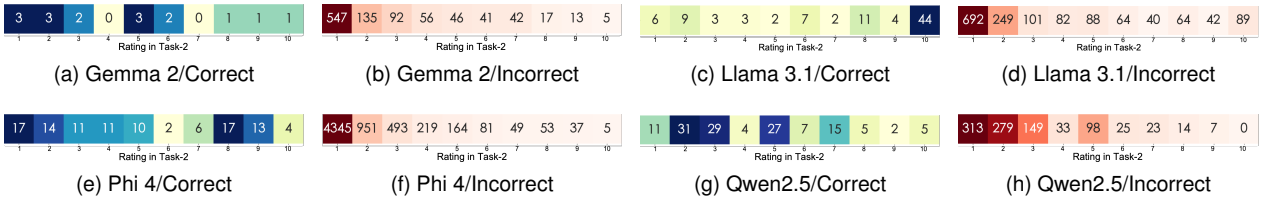
Fig. 2: Distribution of ratings in Task-2 of PFilter-S for four LLMs on "*Person*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

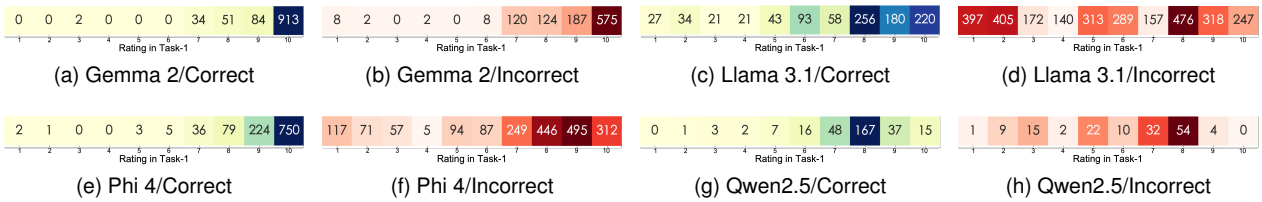
Fig. 3: Distribution of ratings in Task-1 of PFilter-S for four LLMs on "*Location*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

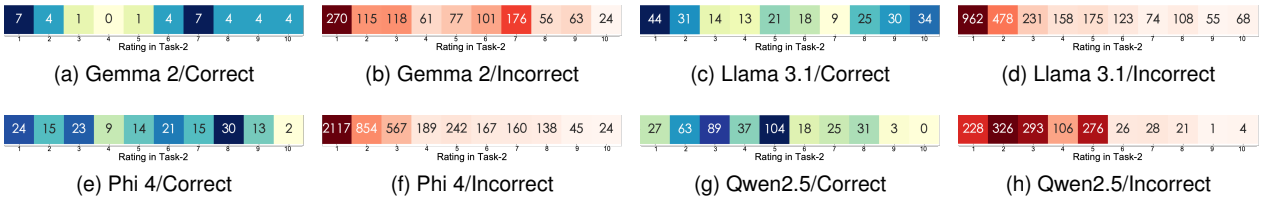
Fig. 4: Distribution of ratings in Task-2 of PFILTER-S for four LLMs on "*Location*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

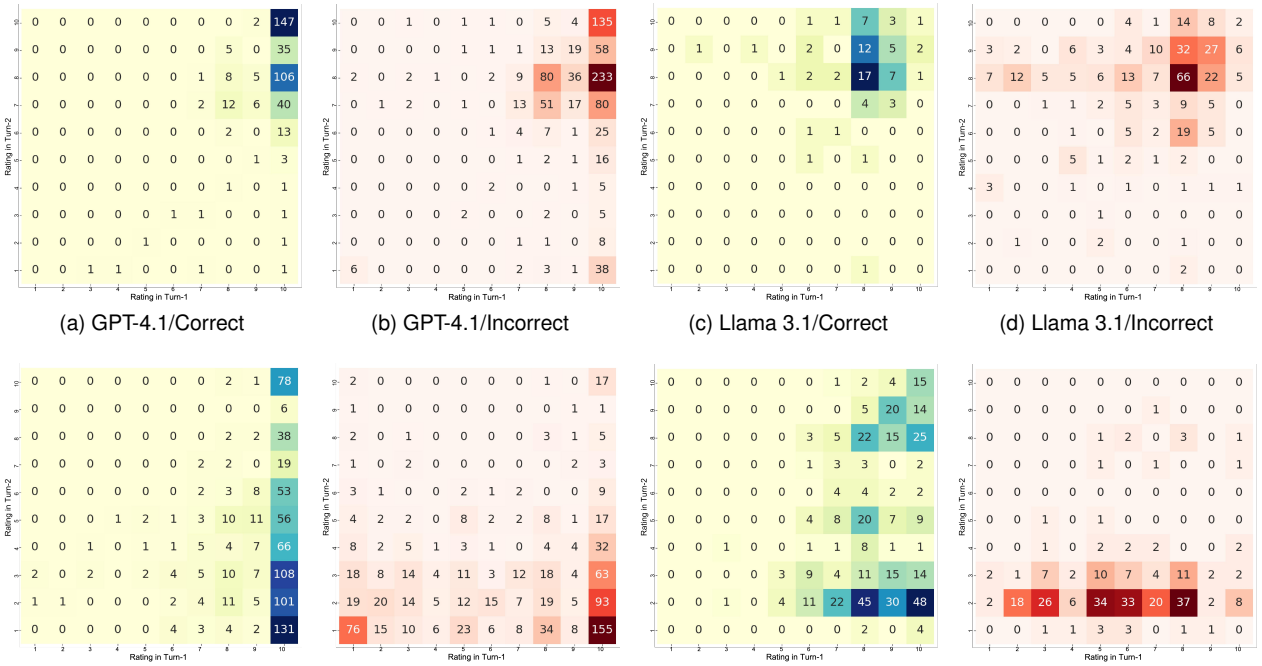
Fig. 5: Distribution of ratings in Task-1 of PFILTER-S for four LLMs on "*Organization*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) Gemma 2/Correct (b) Gemma 2/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

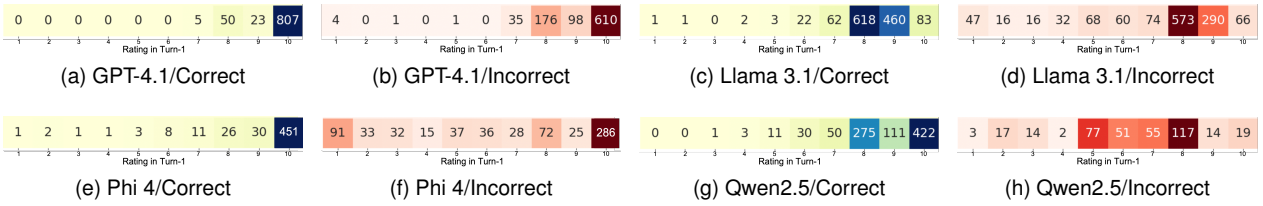
Fig. 6: Distribution of ratings in Task-2 of PFILTER-S for four LLMs on "*Organization*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

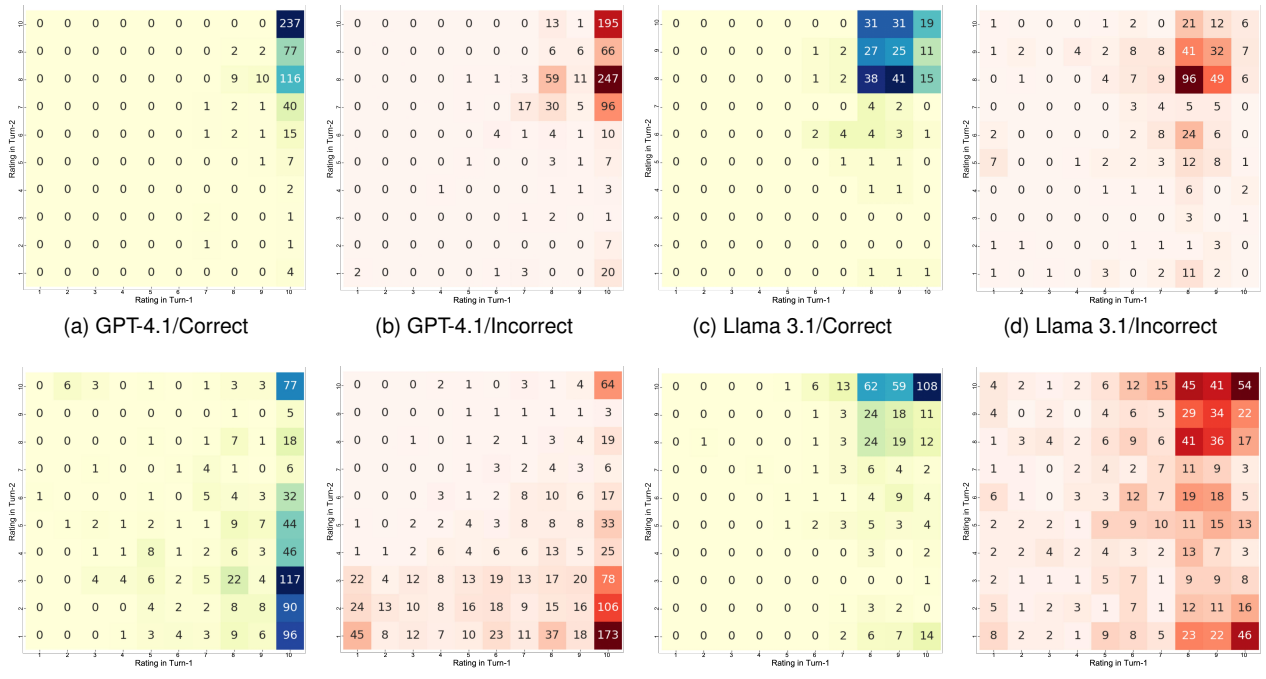
Fig. 7: Distribution of 2-tuples in Study Case-1 of PFILTER-D for four LLMs on "*Person*" type. The color intensity indicates the number of 2-tuples at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

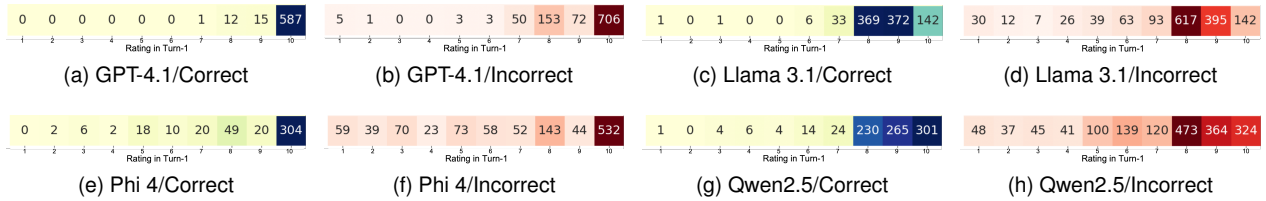
Fig. 8: Distribution of ratings in Study Case-2 of PFILTER-D for four LLMs on "*Person*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

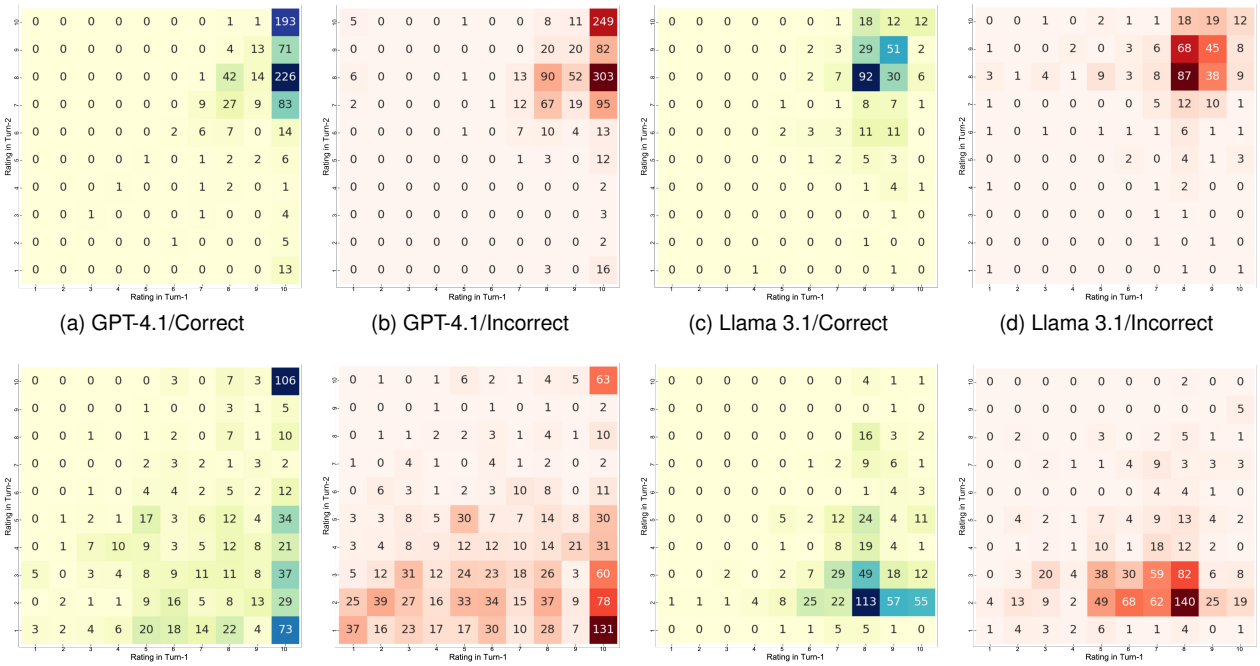
Fig. 9: Distribution of 2-tuples in Study Case-1 of PFILTER-D for four LLMs on "*Location*" type. The color intensity indicates the number of 2-tuples at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

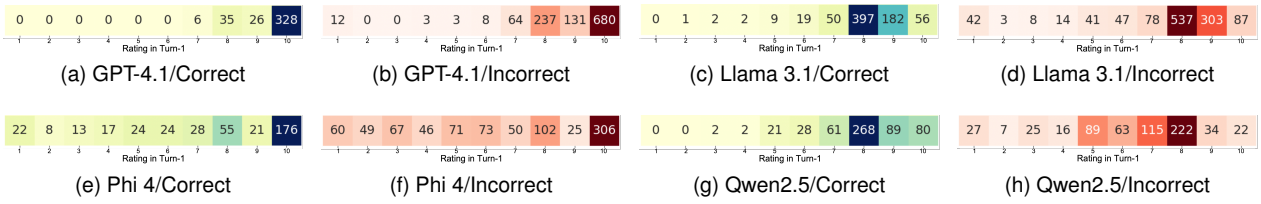
Fig. 10: Distribution of ratings in Study Case-2 of PFILTER-D for four LLMs on "*Location*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

Fig. 11: Distribution of 2-tuples in Study Case-1 of PFILTER-D for four LLMs on "*Organization*" type. The color intensity indicates the number of 2-tuples at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.



(a) GPT-4.1/Correct (b) GPT-4.1/Incorrect (c) Llama 3.1/Correct (d) Llama 3.1/Incorrect

(e) Phi 4/Correct (f) Phi 4/Incorrect (g) Qwen2.5/Correct (h) Qwen2.5/Incorrect

Fig. 12: Distribution of ratings in Study Case-2 of PFILTER-D for four LLMs on "*Organization*" type. The color intensity indicates the number of ratings at each position. Green and red distributions represent the distributions of ratings for correctly and incorrectly recognized entities in Turn-1, respectively.

IV. PROMPTS

A. *Prompts for NER*

The base prompts used for NER are listed in Table IX. Variants that incorporate label descriptions (LD) and few-shot (FS) exemplars are provided in Table X. All prompts are shown verbatim for reproducibility.

Prompt for Named Entity Recognition.

The following sentence may exist entities of the {type} type.

If there are entities:
-Please extract all entities of the type {type}.
-Please only answer all extracted entities strictly in the format: "Entity".

If there are no corresponding entities, please answer an empty string: "".

Sentence: {sentence}
Entity:

TABLE IX: Prompt for named entity recognition.

Prompt for Named Entity Recognition Using Label Description and Few-Shot Learning.

The following sentence may exist entities of the {type} type. Here is the information of entity type {type}: {Label Description}

If there are entities:
-Please extract all entities of the type {type}.
-Please only answer all extracted entities strictly in the format: "Entity".

If there are no corresponding entities, please answer an empty string: "".

Sentence: {sentence}
Output:

Examples
Sentence-1: {sentence 1}
Output-1: {entities in sentence 1}
...

TABLE X: Prompt for named entity recognition using label description and few-Shot learning.

B. *Prompts for* PFILTER-S

The prompt used to elicit self-evaluation ratings for prototype induction is given in Table XI. The prompt used at inference for entity filtering appears in Table XII. Together, these specify the single-turn (PFILTER-S) pipeline.

Prompt for Self-Evaluation Collection for PFILTER-S.

Here is the information of entity type {type}: {Label Description}

Please complete two tasks and output two lists respectively:
1. From the sentence, extract all entities of **{type} type**, and rate their relevance to the **type {type}** on a scale of 1 to 10. Each element in the list should be in the format "Entity//Rating". If there are no such entities, this list is empty [].
2. From the sentence, extract all entities **other than the {type} type**, and rate their relevance to the **type {type}** on a scale of 1 to 10. Each element in the list should be in the format "Entity//Rating". If there are no such entities, this list is empty [].
Only output the two lists, with no other words.

Output Format: [Entity1//Rating1, Entity2//Rating2,...], [Entity3//Rating3, Entity4//Rating4,...]

Sentence: {sentence}

Output:

TABLE XI: Prompt for self-evaluation collection for PFILTER-S.

Prompt for Self-Evaluation in PFILTER-S Filtering.

Here is the information of entity type {type}: {Label Description}

Please complete two tasks and output two lists respectively:
1. For each entity in the following list, rate its relevance to the **type {type}** on a scale of 1 to 10. Each element in the list should be in the format "Entity//Rating". If the Entity_list is empty, this list is empty [].
2. From the sentence, extract all entities **other than the {type} type**, and rate their relevance to the **type {type}** on a scale of 1 to 10. Each element in the list should be in the format "Entity//Rating". If there are no such entities, this list is empty [].
Only output the two lists, with no other words.

Output Format: [Entity1//Rating1, Entity2//Rating2,...], [Entity3//Rating3, Entity4//Rating4,...]

Sentence: {sentence}
Entity_list: {Entity_list}

Output:

TABLE XII: Prompt for self-evaluation in PFILTER-S filtering.

C. *Prompts for* PFILTER-D

For the dual-turn variant, the two prompts used during prototype induction are shown in Tables XIII and XIV. The corresponding two prompts used during the entity-filtering stage are listed in Tables XIV and XV. Note that Table XIV is shared across both stages.

Prompt for Self-Evaluation Collection of Target Entities for PFILTER-D.

The following sentence may exist entities of the {type} type. Here is the information of entity type {type}: {Label Description}

If there are entities:
-Please extract all entities of the type {type} and rate the relevance of extracted entities to the type {type} on a scale of 1 to 10.
-Please only answer all extracted entities with a rating strictly in the following format: "Entity//Rating" for only one phrase or "Entity//Rating, Entity//Rating" for two or more phrases.

Sentence: {sentence}
Output:

TABLE XIII: Prompt for self-evaluation collection of target entities for PFILTER-D.

Prompt for PFILTER-D Filtering and Non-Target Entity Self-Evaluation Collection.

The following sentence may contain entities other than those of the {type} type. Here is the information of entity type {type}: {Label Description}

If there are entities:
-Please extract all entities other than those of the {type} type and rate the relevance of extracted entities to the type {type} on a scale of 1 to 10.
-Please only answer all extracted entities with a rating strictly in the following format: "Entity//Rating" for only one phrase or "Entity//Rating, Entity//Rating" for two or more phrases.

Sentence: {sentence}
Output:

TABLE XIV: Prompt for PFILTER-D filtering and non-target entity self-evaluation collection.

Prompt for Target entity Self-Evaluation in PFILTER-D Filtering.

Here is the information of entity type {type}: {Label Description}

-Please rate the relevance of each phrase in the {list} to the type {type} on a scale of 1 to 10.
-Please only answer all entities in the {list} with ratings strictly in the format: "Entity//Rating" for only one phrase or "Entity//Rating, Entity//Rating" for two or more phrases.

If there are no corresponding entities, please answer an empty string: "".

Output:

TABLE XV: Prompt for target entity self-evaluation in PFILTER-D filtering.