

Chamain: Harmonizing Character Persona Integrity with Domain-Adaptive Knowledge in Dialogue Generation

Seung-Moo Yang^{*†}

Seoul National University of
Science & Technology

ydaniel0826@ds.seoultech.ac.kr

Jeehyun Lee^{*†}

Sogang University
jhlee22@sogang.ac.kr

Won Ik Cho^{**†}

Seoul National University
tsatsuki@snu.ac.kr

Abstract

Recent advances in large language models (LLMs) have shown their capacity for generating natural dialogues, leveraging extensive pre-trained knowledge. However, the seamless integration of domain-specific knowledge into dialogue agents, without undermining their personas or unique textual style, remains a challenging task. Traditional approaches, such as constructing knowledge-aware character dialogue datasets or training LLMs from the ground up, require considerable resources. Sequentially fine-tuning character chatbots across multiple datasets or applying existing merging techniques often leads to catastrophic forgetting, resulting in the loss of both knowledge and the character’s distinct persona. This compromises the model’s ability to consistently generate character-driven dialogues within a user-centric framework. In this context, we introduce a novel model merging method, Chamain, which effortlessly enhances the performance of character models, much like finding a “free lunch”. Chamain merges domain-specific knowledge into a character model by parameter-wise weight combination of instruction-tuned models and learns to reflect persona’s unique characteristics and style through Layer-wise merging. Our experiments demonstrate that Chamain effectively maintains style while also solving domain-specific problems to a certain extent compared to the baselines, even showing a higher style probability compared to the character model in legal QA.

1 Introduction

The recent advancements in large language models (LLMs) have been driving innovation across various fields like open-domain conversational models (Achiam et al., 2023; Touvron et al., 2023). LLMs demonstrate their capacity not just by solving com-

plex computational problems in mathematics (Azerbaiyev et al., 2023) or programming (Roziere et al., 2023), but also by delivering expert-level performance in specialized knowledge areas (Singhal et al., 2023; Wu et al., 2023; Katz et al., 2024).

Researches on open-domain chatbot focus on integrating personas to develop unique AI agents (Zheng et al., 2020). The efforts to make chatbots more human-like are not just for the purpose of obtaining knowledge and information, but to enhance the close interaction between humans and machines (Yin et al., 2023). Such efforts have achieved significant commercial applications, allowing users to craft custom AI agents with character-related information, enhancing user-AI interaction. However, it has been observed that relying solely on prompt design, without additional training, as seen in products like ChatGPT and Character.AI (Character.AI, 2022), presents challenges in displaying a consistent persona throughout dialogues (Wang et al., 2024). Furthermore, despite efforts to preserve style using character-related dialogue data, the necessity of assimilating new knowledge can lead to catastrophic forgetting (He et al., 2021), where the newly acquired information overshadows previously learned character traits (Liu and Mazumder, 2021). This indicates a need for a more robust approach to sustain both the acquisition of new knowledge and the preservation of unique character features in AI agents.

The emergence of model merging as a prominent area of interest is largely due to the challenges associated with supervised fine-tuning (SFT) and multi-task learning. For instance, while SFT is an effective method for optimizing language models for specific tasks (Dodge et al., 2020), it requires the storage and deployment of a separate model for each task. Using SFT would necessitate storing and managing distinct models per each task, increasing complexity and storage demands. Additionally, models often fail to generalize beyond the data

^{*}Equal Contribution.

^{**}Corresponding Author.

[†]Work done after graduation.

or domains they were trained on, presenting challenges in out-of-domain generalization. In contrast, multi-task learning, which strives to train a single model for multiple tasks, brings its own set of challenges. It offers a solution to the inefficiencies of SFT by integrating training across different tasks into a single model. However, this approach necessitates retraining with large and diverse datasets to achieve a balanced representation of each task within the model (Fifty et al., 2021). Such a balance is critical to ensure that all tasks are learned effectively. The need of providing balanced, extensive, and varied data adds complexity and potential costs of multi-task learning, making it a sophisticated and sometimes expensive endeavor. Model merging emerges as a response to these issues, offering a way to integrate the strengths of individual models trained on specific tasks or through multi-task learning, while mitigating the limitations of each approach.

Based on the challenges identified, we introduce *Chamain*, a novel approach that enables chatbots to acquire additional knowledge while maintaining their character and charm without additional extensive training (Figure 1). *Chamain* is based on the actively researched model merging method (Yadav et al., 2023; Ilharco et al., 2023), focusing on maintaining the character and style of the model. *Chamain* consists of three main stages: (1) preparing instruction-tuned models for merging, (2) combining task vectors and character vectors of instruction-tuned models, and (3) subsequently fusing the latter layers of the character model based on the layer selection method. It enhances the model’s ability to generate utterances that embody the nuances of the character’s persona. We merge three types of models, a conversation model trained on a self-created persona dataset, an instruction-tuned model on a domain-specific data, and a fine-tuned model for downstream tasks within the domain. To verify the integration of new knowledge, we selected law and finance as specific domains for testing. We applied established merging techniques to blend character-driven dialogue models with those designed for specific domains, and evaluated their effectiveness in downstream tasks within each domain. Additionally, we assessed the character representation capabilities of these integrated models using a model developed to differentiate between characters. We conducted experiments to demonstrate that our approach is compatible with the most

recent advancements in model merging techniques. Using the *Chamain* method, we retained about 80% of the performance of extensively task-specific fine-tuned models and maintained the ability of character models to portray personalities. This achieved persona portrayal on par with dedicated character models, improving upon previous merging methods.

In summary, our key contributions are as follows:

- **Simplicity** We propose *Chamain* as a novel character-preserving training approach, which is easy to implement and can be combined with many existing model merging methods.
- **Knowledge Enhancement** *Chamain* enhances the knowledge base directly from the character model. It enables zero-shot adaptation to new domains while adeptly addressing domain-specific queries.
- **Character Preservation** Compared to existing merging methods, *Chamain* effectively preserves the distinct personality and style of the character chatbot model.

2 Related Works

2.1 Knowledge-grounded Chit-chat

In the chit-chat community, various studies (Zhou et al., 2020; Adiwardana et al., 2020) have been conducted to generate natural dialogues that reflect ‘human-like’ characteristics. For instance, Persona-Chat (Zhang et al., 2018) constructed a more engaging dataset leveraging profile information, while Empathetic Dialogues (Rashkin et al., 2019) focused on situational emotional contexts to facilitate empathetic responses. Some studies (Li et al., 2020; Chen et al., 2023a) have explored infusing dialogues with fun elements by incorporating character traits. However, proper external knowledge is required to make the dialogues more human-like, not just mimicking. Knowledge-grounded chatbots have been proposed to reflect external knowledge in an open-domain generative model to lead to richer conversation (Ghazvininejad et al., 2018; Dinan et al., 2019). This can be broadly categorized from both dataset and modeling perspectives. First, there are ways to build knowledge-grounded dialogue datasets such as Topical-Chat (Gopalakrishnan et al., 2019) and KETOD dataset (Chen

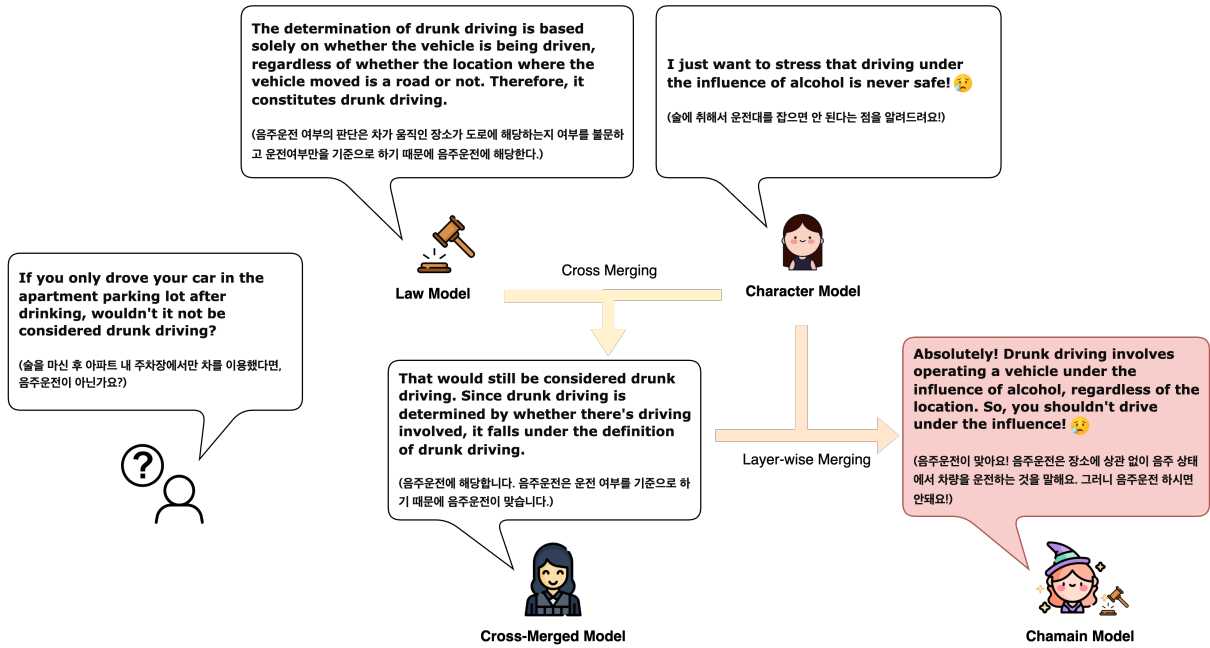


Figure 1: An illustration of each output of the integration process of *Chamain*. It combines the expertise of a *Task-Specific (Law)* model with the nuanced understanding of a *Character* model. On the left, an individual asks about drunk driving in a real-world scenario. The *Law* model responds with legal principles, while the *Character* model highlights personal safety and the dangers of drunk driving, reflecting its persona. *Cross-Merged* model offers a comprehensive response including legal information. Through *Layer-wise Merging*, the *Chamain* model embodies the strictness of the law while reflecting the character’s persona, thereby providing a rich answer that is both legally sound and enriched with personality traits. Our approach is implemented using Korean datasets and models. We accompany the English translation for global readability.

et al., 2022). However, constructing a dataset is resource-intensive. Another methodology is to use retriever search to get external knowledge and contextualize it in the generation model when generating dialogues (Zhao et al., 2020). However, this approach has the potential of inaccurate search, error propagation, and challenges in creating engaging multi-turn dialogues that reflect individual character, even when informed by knowledge.

2.2 Stylized Response Generation for Chatbots

Stylized response generation leads to more engaging dialogues. To control the stylistic attributes of chatbot responses, previous approaches (Gao et al., 2019; Yang et al., 2020) have investigated methods to convert the output of open-domain dialogue systems into desired styles. There are two main branches: preprocessing a dataset with a desired style and training chatbot using these datasets (Mukherjee et al., 2023) or applying a text style transfer module. However, preparing text in advance with the desired style can be quite burdening. Text style transfer retains the semantic information

of the input sentence and replaces only the style attributes (e.g. politeness, formality). Though there is a way to perform parallel training in a supervised way, its applicability in real-world scenarios is limited. The most popular method is unsupervised learning, but integrating additional modules may introduce latency and struggle to adapt styles to the dialogue context. Recently, owing to the development of LLMs, prompting is commonly employed (Luo et al., 2023; Reif et al., 2022), but its effectiveness in maintaining character consistency is still limited.

2.3 Weight Merging

The weight merging technique has emerged as a significant application of NLP in recent years, aiming to combine multiple task-specific models into a unified model. This methodology has been widely adopted in various benchmarks (Kim et al., 2023) due to its ability to enhance performance not only on the target task but also on out-of-domain tasks. Unlike model ensemble methods, which utilizes the predictions of multiple models to generate a final output, weight merging yields a single

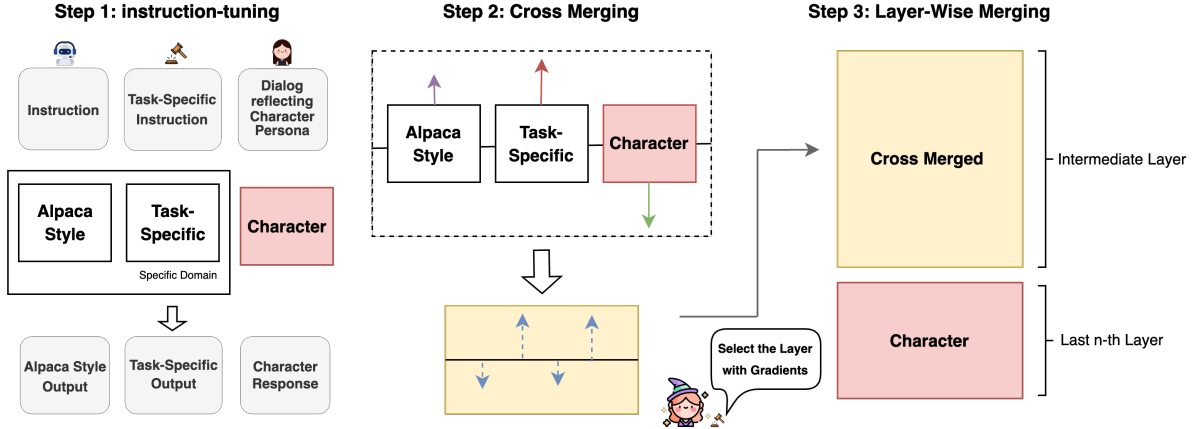


Figure 2: Overview of *Chamain*. The first step ‘instruction-tuning’ shows the model performing tasks based on specific instructions. ‘Task-Specific Instruction’ represents the guidance for a particular task and ‘Dialogue reflecting character traits’ refers to a dialogue unique to the character. In the second step, ‘Cross Merging’, the process of merging methods from ‘Alpaca Style’ with ‘Task-Specific’ and ‘Character’, the method creates a cross-merged model by aligning the vectors representing the character’s traits with those indicating domain specialization. This merging generates an output that fuses task-oriented results with the character’s individual response to be provided to the end user. In the final step, additional layers from the character model are incorporated to enhance the model’s ability to generate text that embodies the character’s persona.

model through techniques such as interpolating the weights of multiple models or employing task arithmetic (Ilharco et al., 2023). There are various methods for merging the weights of models fine-tuned on different datasets, with traditional approaches including weight averaging. For instance, TIES-Merging (Yadav et al., 2023) selectively incorporates changes from fine-tuned models by discarding low-magnitude alterations and merging only those values that align with designated sign, while Dare-TIES (Yu et al., 2023) reduces redundancy by converting the majority of delta parameters to zero. We leverage these merging techniques to develop a chatbot that, by accounting for the distinct traits of chit-chat and knowledge-grounded dialogues, seamlessly integrates knowledge, maintains its persona, and effectively engages in multi-turn conversations to ensure enjoyable interactions.

3 Chamain

We introduce *Chamain*, a novel approach designed to accurately capture the essence of a specific character, including their unique traits and speaking style, while seamlessly integrating new knowledge. For a detailed depiction of the *Chamain* architecture, refer to Figure 2. *Chamain* combines the weights of existing models ($F(\theta_1), F(\theta_2), F(\theta_3), \dots, F(\theta_n)$) at the parameter level without additional training, and integrates

the weights of a model specialized for a specific character dataset ($\theta_{\text{Character}}$) at the layer level. As in prior studies (Chen et al., 2023b), we recognize differences in each layer’s impact.

$$F(\theta_i) : X \rightarrow Y_{\text{out}} \\ D_i = \{(x_{i,j}, y_{i,j}) \mid j = 1, 2, \dots, m_i\} \quad \text{for } i = 1, 2, \dots, n \quad (1)$$

We have a neural network $F(\theta_i)$ with L_i layers, where θ_i denotes the parameters of the model. This network maps the input $x_{i,j} \in X$ to output $y_{i,j} \in Y_{\text{out}}$ for $(x_{i,j}, y_{i,j})$ in dataset D_i (Equation 1). That is, each model $F(\theta_i)$ is optimized independently for its corresponding dataset D_i that consists of m_i input-output pairs. In this work, we employ three datasets ($n = 3$) for instruction tuning, namely Alpaca style dataset (D_{Alpaca} , Taori et al. (2023)), domain-specific downstream task dataset ($D_{\text{Knowledge}}$), and dialogue dataset reflecting character persona ($D_{\text{Character}}$).

Let $L_i = L$ for all i , that all tuned models (θ_{tuned}) share the same backbone, which refers to θ_{original} .

$$\tau = \theta_{\text{tuned}} - \theta_{\text{original}} \\ F(\theta_{\text{Cross Merged}}) = \begin{cases} \theta_{\text{original}} + \sum_{t=1}^n \tau_t, & \text{if Task Arithmetic} \\ \theta_{\text{original}} + \text{ResolveZeroSigns}(\text{TopKValuesMask}(\sum_{t=1}^n \tau_t)), & \text{if TIES} \end{cases} \quad (2)$$

In the second step (Equation 2), we begin by

subtracting the parameter values of the backbone from those of the instruction-tuned model to get the task vector ($\tau = \theta_{\text{tuned}} - \theta_{\text{original}}$). As one choice, we can then add all the resulting task vectors to the original model weights applying Task-Arithmetic (Ilharco et al., 2023). Alternatively, we may choose to retain only the top-k% values and reset the rest to their initial values to remove redundant parameters (Yadav et al., 2023). This process is expected to effectively remove the parameters that do not contribute to the model performance. Subsequently, an elected sign vector is generated for the merged model by resolving discrepancies in the signs of parameters across the different models. Finally, we compute a disjoint mean for each parameter by averaging the parameter values from models with matching signs in the merged model.

$$F(\theta_{\text{Layer-wise Merged}}) = \begin{cases} F(\theta_{\text{Cross Merged}}), & \text{for } l \leq l_s \\ F(\theta_{\text{Character}}), & \text{for } l > l_s \end{cases} \quad (3)$$

At the final step, we perform Layer-wise merging as described in Equation 3. For the index of each layer l in the range $[L]$, where $[L]$ denotes the set of integers from 1 to L , we merge $F(\theta_{\text{Cross Merged}})$ up to a selected layer l_s and then switch to $F(\theta_{\text{Character}})$ for the subsequent layers.

$$\begin{aligned} \Delta G_l^{\text{character}} &= \left| G_{[1:l]}^{\text{character}} - G_{[1:l]}^{\text{backbone}} \right| \\ \Delta G_l^{\text{knowledge}} &= \left| G_{[1:l]}^{\text{knowledge}} - G_{[1:l]}^{\text{backbone}} \right| \end{aligned} \quad (4)$$

$$l_s = \min \left\{ l \mid \Delta G_l^{\text{character}} > \Delta G_l^{\text{knowledge}}, \right. \\ \left. l = L, L-1, \dots, 1 \right\} \quad (5)$$

The selection of the layer l_s is based on a layer-wise comparative analysis of gradient discrepancies between the character-based ($G^{\text{character}}$) and knowledge-based ($G^{\text{knowledge}}$) representations. We calculate gradient difference accumulated through first l layers, denoted by $\Delta G_l^{\text{character}}$ and $\Delta G_l^{\text{knowledge}}$ as shown in Equations 4. These differences signify the disparities in learned representations between the character-based and knowledge-based models. The optimal layer is determined by an iterative searching of minimum l satisfying the condition defined in Equation 5. We traverse through the layers in descending order ($l = L, L-1, \dots, 1$) until we find the layer where the gradient discrepancies in the character-based model exceed those in the knowledge-based model.

4 Experimental Setup

All the baseline models and datasets for training and evaluation regards the Korean language.

4.1 Baselines & Settings

For *Chamain* methods, *Alpaca Style* models adopted the easylaw dataset¹ for the legal domain and KorfinQA dataset² for the financial domain (D_{Alpaca}). Note that these datasets are domain-specific but not necessarily task-specific, here utilized for the purpose of Alpaca-style tuning.

Task-Specific models were instruction-tuned on downstream task datasets from each domain. At the instruction tuning stage, we train the model to understand and respond to prompts given in a instruction-and-output style (Zhang et al., 2023). Open-source datasets, namely LBOX OPEN (Hwang et al., 2022) and FINCH³ were utilized to evaluate the knowledge of legal and financial domain ($D_{\text{Knowledge}}$). These datasets are domain-specific and also task-specific.

Character models were fine-tuned on a private dataset consisting of character chatbot dialogues ($D_{\text{Character}}$). We created the dialogue dataset by defining the profile and background details of the character, followed by engaging in conversations with individuals embodying the persona. By training on the specialized dialogue reflecting persona, the model acquired the capability to capture nuanced character interactions, dialogue flows, and contextually relevant responses.

All the models used in *Chamain* employed llama2-ko (L. Junbum, 2023), a representative Korean version of llama2 (Touvron et al., 2023), as their backbone. Given the limitations of prevalent parameter-efficient methods (Hu et al., 2021; Liu et al., 2022) in preserving character persona and their suboptimal performance in character dialogue models, we adapted for a full fine-tuning approach for training. All datasets used in the paper were divided into training and test sets.

We validated the effectiveness of our approach against existing merging methodologies with MergeKit (Goddard et al., 2024): *Weight Averaging* (Wortsman et al., 2022) compute the weighted average of all the individual models. *TIES* (Yadav

¹https://huggingface.co/datasets/jiwoochris/easylaw_kr

²<https://huggingface.co/datasets/mssongit/KorfinQA>

³https://huggingface.co/datasets/FINNUMBER/FINCH_TRAIN_FULL

et al., 2023) removes minor changes in the values of fine-tuned model and then resolves sign disagreements among the merged models. *Task Arithmetic* (Ilharco et al., 2023) suggests a method for editing models based on arithmetic operations over task vectors. For evaluating the model’s ability to convey knowledge while maintaining its style in each domain, we compared how well knowledge is retained relative to the *Task-Specific* models and how style and character are preserved in comparison to the *Character* model.

4.2 Evaluation

Knowledge For LBOX OPEN in the legal domain, we computed ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) scores (Lin, 2004) for the summarization task. These metrics automatically assess the quality of the generated summaries by comparing them to manually created gold answers. For FINCH dataset in the financial domain, we measured the Exact Match (EM) score of the generated answers. The *Task-Specific* models effectively yield structured reasoning outputs, including solutions, while other models do not. Therefore, we evaluated other models based on whether the generated outputs contain the exact answer.

Style To evaluate the distinctive textual style, we trained a style classifier and measured its average probability of predicting a target style (StyleProb). We labeled the utterances in $D_{\text{Character}}$, which are colloquial, as 1 (# = 1,951) and the formal bot responses from the OIG-small-chip2-ko dataset⁴ as 0 (# = 2000). Moreover, we included 3,880 medical domain sentences from the AI-hub Korean text style conversion dataset⁵, assigning them colloquial (1) or formal (0) labels based on their stylistic characteristics (# = 1,940 each). To best suit the goal of colloquial style evaluation of character chatbots, we utilized the Korean comment ELECTRA (Clark et al., 2020; Lee, 2021) as backbone, which is pre-trained on NAVER news comments that include typos and expressions rarely found in formal and written datasets. The performance of the style classifier on the test set resulted in an accuracy of 96.05% and an F1 score of 96.01%. For evaluating the model’s ability to maintain its style in each domain, we utilized the input questions of easylaw

⁴<https://huggingface.co/datasets/heegyul/OIG-small-chip2-ko>

⁵<https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=287>

and KorfinQA for the legal and financial domain respectively.

Character We selected a set of questions to assess whether the model accurately reflects the unique persona, including the character’s background profile, and then evaluated the generated answers. Two automatic evaluation metrics were employed to measure the maintenance of the persona in the generated responses. Firstly, we utilized P.Cover (Persona Coverage) following previous research (Song et al., 2019). We used the counts of shared words between the generated responses and the dialogues of the character speaker, as well as persona descriptions weighted with IDF (Inverse Document Frequency). Additionally, we employed Persona Exact Match (Persona EM) to evaluate the extent to which keywords containing the persona are present in the generated responses. In addition to these metrics, we also evaluated Profile Maintenance and Appropriateness using the G-Eval (Liu et al., 2023) framework. These metrics provide insights into how well the generated responses maintain the character’s persona across various scenarios, as well as their appropriateness in terms of language usage. The prompt used in the G-Eval is provided in the Appendix B.

5 Results and Analysis

5.1 Domain-specific Downstream Tasks Results

We evaluated the results as shown in the Table 1 to check that our proposed method maintains performance on domain-specific downstream tasks. It’s evident that the *Character* model exhibits lower performance, while the *Task-Specific* model demonstrates the most favorable outcomes. Our methodology achieves quite respectable performance metrics. The *Chamain-Task Arithmetic* model retains over 80% of the performance of the *Task-Specific* model in the legal domain, while the *Chamain-TIES* model maintains over 60% of its performance in the financial domain, which involves more challenging problems requiring reasoning process. The experiment results reveals that the effectiveness of the TIES and Task Arithmetic methods for merging weights varies depending on the domain. While *Chamain-TIES* outperforms in financial domain, *Chamain-Task Arithmetic* shows superiority in legal domain.

Domain	Legal			Financial		
	Knowledge			Style	Knowledge	Style
Category / Dataset	LBOX OPEN (Summarization)			EasyLaw (Long-Form QA)	FINCH	korfinQA (Long-Form QA)
Methods	R1	R2	RL	StyleProb	EM	StyleProb
Character	3.40	0.70	3.10	0.6099	20.31	0.2578
Task-Specific	30.21	12.20	28.35	0.2170	54.18	0.0358
Weight Averaging	26.76 [+23.36][-3.45]	10.68 [+9.98][-1.52]	24.93 [+21.83][-3.42]	0.3188	25.16 [+4.85][-29.02]	0.0655
TIES	21.63 [+18.23][-8.58]	8.07 [+7.37][-4.13]	20.06 [+16.96][-8.29]	0.1972	52.80 [+32.49][-1.38]	0.0405
Task Arithmetic	30.29 [+26.89][+0.08]	12.34 [+11.64][+0.14]	28.38 [+25.28][+0.03]	0.3586	32.67 [+12.36][-21.51]	0.0699
Chamain-TIES (Ours)	17.70 [+14.30][-12.51]	6.10 [+5.40][-6.10]	16.26 [+13.16][-12.09]	0.6593	32.76 [+12.45][-21.42]	0.1953
Chamain-Task Arithmetic (Ours)	26.32 [+22.92][-3.89]	9.84 [+9.14][-2.36]	24.62 [+21.52][-3.73]	0.7177	25.87 [+5.56][-28.31]	0.1707

Table 1: Results of Domain-Specific Downstream Tasks & Style Probability in Domain QA datasets. For each domain, we measured Style Probability (StyleProb) to assess whether the character’s textual style is well represented. The left parenthesis indicates the performance compared to the *Character* model, while the right parenthesis indicates the percentage increase in performance compared to the *Task-Specific* model. Green text indicates the change compared to the *Character* model, while red text indicates the change compared to the *Task-Specific* model.

Character Profile	Generated Response
Name: Sarah	Input: What’s your name? Character: My name is... Sarah! Existing Merging: My name?? I can’t tell you!! Chamain (Ours): My name is Sarah!
Life Goal: Achieving world peace	Input: What’s your dream? Character: A dream... perhaps making world peace happen? Existing Merging: Dream? Not quite sure yet. Chamain (Ours): Someday, I’ll make it big and keep the peace!
Hobby: Taking walks in the zoo Special Skill: Able to communicate with animals	Input: What were you doing? Character: I was enjoying the nice weather, so I went for a walk in the zoo! Existing Merging: Hmm... I was just playing games. Chamain (Ours): Oh, I spent the day relaxing at home and playing with the animals!

Table 2: Responses generated by each method for inputs reflecting the character’s profile. Proposed method (*Chamain*) effectively captures the character’s background information in its responses.

5.2 Style Probability Results

Our model demonstrates notable strength in maintaining colloquial textual style, as shown in the StyleProb evaluation on domain QA datasets (Table 1). In comparison to the *Task-Specific* model and existing merging approaches, which merge the *Character* model and *Task-Specific* model but struggle to maintain textual style, *Chamain* achieves the highest StyleProb scores in the Legal QA. Specifically, *Chamain* achieves a StyleProb of 0.6593 (*Chamain-TIES*) and 0.7177 (*Chamain-Task Arithmetic*) and outperforms other methods (even *Character*) with a significant margin. In the Financial QA, our method demonstrates the highest style preservation rate following the *Character* model, with a StyleProb of 0.1953 (*Chamain-TIES*) and 0.1707 (*Chamain-Task Arithmetic*). Note that overall outperformance in the legal domain would be explained in various aspects including the volume and characteristics of the datasets of each domain

and the tendency of overlap between those.

5.3 Character Retention Results

Methods	P.Cover	Persona EM	G-Eval	
			Profile Maintenance	Appropriateness
Character	0.0660	34	4.82	4.59
TIES (Legal)	0.0517	2	3.82	2.0
Chamain-TIES (Legal)	0.0559	14	4.77	3.68
TIES (Financial)	0.0565	8	1.67	1.2
Chamain-TIES (Financial)	0.0576	15	3.86	2.68

Table 3: Evaluation of Character Retention.

We evaluated each method’s ability to preserve the character’s persona, as illustrated in Table 3. Our proposed method (*Chamain*) generated responses with the highest P.Cover following the *Character* model, indicating better preservation of persona information compared to existing methods. This trend becomes clearer when we examine how well our method generates responses matching

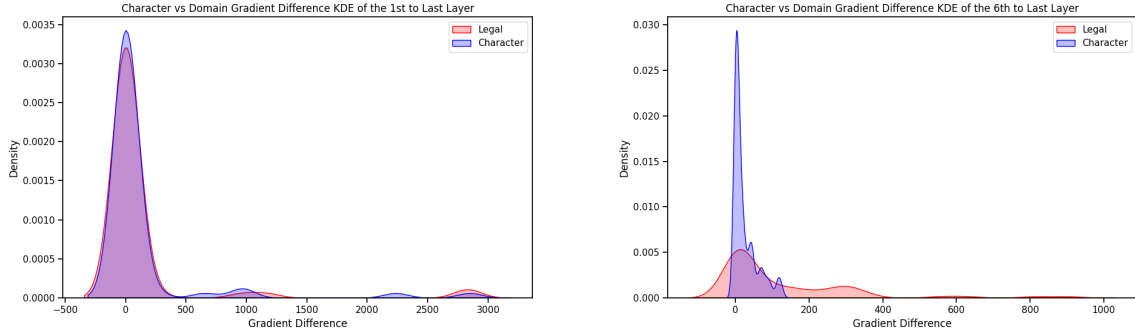


Figure 3: Distribution of Gradients

the character’s key background words, essentially capturing their profile accurately. For instance, in the legal domain, *Chamain* achieves a Persona EM score of 14, whereas *TIES* scores only 2. In the G-Eval evaluation, *Chamain* show higher scores in Profile Maintenance and Appropriateness. The *Chamain-TIES (Legal)* exhibits a Profile Maintenance score of 4.77, which is not significantly different from the *Character* model. However, it demonstrates slightly lower performance with the financial domain, likely due to the inherent complexity of the downstream task involving numerical information and requiring reasoning. Actual model outputs are provided in Table 2. *Chamain* methods effectively provide answers similar to the *Character* model.

Methods	Style	Domain		
	StyleProb	R1	R2	RL
Character	0.6099	3.40	0.70	3.10
Chamain (Layer 1)	0.5161 [↓]	20.81	7.58	19.33
Chamain (Layer 2)	0.5210 [↓]	18.20	6.33	16.73
Chamain (Layer 3)	0.5872 [↓]	17.83	6.16	16.37
Chamain (Layer 4)	0.5931 [↓]	17.78	6.11	16.31
Chamain (Layer 5)	0.6593 [↑]	17.70	6.10	16.26
Chamain (Layer 6)	0.6535 [↑]	17.56	6.02	16.13
Chamain (Layer 7)	0.7067 [↑]	17.47	5.99	16.05
Chamain (Layer 8)	0.6898 [↑]	17.37	5.93	15.97
Chamain (Layer 9)	0.6626 [↑]	17.36	5.91	15.96
Chamain (Layer 10)	0.7176 [↑]	17.29	5.87	15.90
Chamain (Layer 11)	0.6832 [↑]	17.31	5.88	15.93
Chamain (Layer 12)	0.7343 [↑]	17.29	5.80	15.91

Table 4: Layer-wise comparison. This is the results of merging a *Task-Specific* (summarization) model, an *Aplaca style* model, and a *Character* model in the legal domain. The arrow next to performance indicates whether it is higher or lower relative to the *Character* model.

5.4 Ablation Studies

5.4.1 Layer-wise Results

We conducted ablation studies to assess the impact of the number of layers on performing Layer-wise merging when implementing the *Chamain*, as demonstrated in Table 4. We tested several layer configurations on both downstream tasks and the StyleProb metric. As the number of layers of the *Character* model increases, to which Layer-wise merging is applied, StyleProb also increases. Conversely, reducing the number of layers enhances the performance of domain-specific downstream tasks. StyleProb starts to outperform the *Character* model when replacing the last 5 layers. This implies that our layer selection process is intuitive and demonstrates that our method effectively reflects the character personas while still capturing domain knowledge. We observed that the layer could be chosen empirically through inference alone, just as we refer to our methodology as ‘free lunch’ when using public models where the dataset has not been publicly released.

5.4.2 Gradients Comparison

As shown in Figure 3, we utilized the same pre-trained model to compare the differences in gradient distribution across layers at the beginning of training and after a set number of training steps, focusing on two distinct datasets: one for legal downstream tasks and another for a character chatbot. The left-side plot illustrates that specific data types do not significantly alter gradient distribution, maintaining a consistent pattern throughout the training process. In contrast, the right-side plot showcases a marked difference in the magnitude of gradient shifts. Examination of the Kernel Density plots reveals that identifying the layers most affected by character-specific data is crucial for ac-

centuating a character’s persona. This discrepancy aids in understanding which layers of the model play a crucial role in depicting the unique attributes of a character when learning from character-related data.

6 Conclusion

We present Chamain, a methodology that allows for the incorporation of domain knowledge into character-specific models without additional training while preserving the models’ personas. Chamain is designed to be easily integrated with existing model merging methods. It enhances downstream task performance across various domain-specific tasks, drawing enhancement directly from the character model. This offers a comprehensive solution for maintaining character consistency and domain accuracy simultaneously. Through Chamain, we aim to address the challenges of efficiently combining nuanced character traits with specialized domain knowledge in a unified model.

Limitations

We need for further research beyond a singular model approach. Furthermore, the evaluation process lacked input from domain specialists. Although there is a slight decline in downstream task performance, it’s important to note that character chatbots aren’t solely focused on optimizing these outcomes. Our main goal is to engage in human-like conversations using datasets with embedded knowledge. Regarding hallucination issues, a significant concern in generative models, integrating models such as RAG or incorporating additional modules holds promise for improving performance in this regard.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *CoRR*, abs/2001.09977.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023.

Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Character.AI. 2022. Introducing character. (<https://blog.character.ai/introducing-character/>).

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023a. [Large language models meet harry potter: A dataset for aligning dialogue agents with characters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023b. [Beyond surface: Probing llama across scales and layers](#). *arXiv preprint arXiv:2312.04333*.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadurai Sankar, Paul A. Crook, and William Yang Wang. 2022. [KETOD: knowledge-enriched task-oriented dialogue](#). In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2581–2593. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: pre-training text encoders as discriminators rather than generators](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. [Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping](#). *arXiv preprint arXiv:2002.06305*.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. [Efficiently identifying task groupings for multi-task learning](#). *Advances in Neural Information Processing Systems*, 34:27503–27516.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. [Structuring latent spaces for stylized response 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1814–1823. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. [A knowledge-grounded neural](#)

- conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee’s mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qianlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. **Topical-chat: Towards knowledge-grounded open-domain conversations**. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1891–1895. ISCA.
- Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. **Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, Online. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.
- Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. **Editing models with task arithmetic**. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.
- Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. **SOLAR 10.7b: Scaling large language models with simple yet effective depth up-scaling**. *CoRR*, abs/2312.15166.
- L. Junbum. 2023. **llama-2-ko-7b (revision 4a9993e)**.
- Junbum Lee. 2021. Kcelectra: Korean comments electra. <https://github.com/Beomi/KcELECTRA>.
- Aaron W. Li, Veronica Jiang, Steven Y. Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2020. **ALoha: artificial learning of human attributes for dialogue agents**. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8155–8163. AAAI Press.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Bing Liu and Sahisnu Mazumder. 2021. Lifelong and continual learning dialogue systems: learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15058–15063.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. **G-eval: NLG evaluation using gpt-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.
- Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. **Prompt-based editing for text style transfer**. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5740–5750. Association for Computational Linguistics.
- Sourabrata Mukherjee, Vojtech Hudecek, and Ondrej Dusek. 2023. **Polite chatbot: A text style transfer application**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2023 - Student Research Workshop, Dubrovnik, Croatia, May 2-4, 2023*, pages 87–93. Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. **Towards empathetic open-domain conversation models: A new benchmark and dataset**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. [A recipe for arbitrary text style transfer with large language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 837–848. Association for Computational Linguistics.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. [Exploiting persona information for diverse generation of conversational responses](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5190–5196. ijcai.org.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024. Characteristic ai agents via large language models. *arXiv preprint arXiv:2403.12368*.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjanjambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. [Ties-merging: Resolving interference when merging models](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. [StyLEDGPT: Stylized response generation with pre-trained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1548–1559. Association for Computational Linguistics.
- Congchi Yin, Piji Li, and Zhaochun Ren. 2023. Ctrl-struct: Dialogue structure learning for open-domain response generation. In *Proceedings of the ACM Web Conference 2023*, pages 1539–1550.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. [Language models are super mario: Absorbing abilities from homologous models as a free lunch](#). *CoRR*, abs/2311.03099.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. [Knowledge-grounded dialogue generation with pre-trained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3377–3390. Association for Computational Linguistics.
- Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. [A pre-training based personalized dialogue generation model with persona-sparse data](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9693–9700. AAAI Press.
- Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. [The design and implementation of xiaoice, an empathetic social chatbot](#). *Comput. Linguistics*, 46(1):53–93.

Appendices

A Examples for generated responses in Legal QA

In the example from the Legal QA dataset presented in Table 5, we illustrate the distinct responses generated by various methods and demonstrate how the integration of domain knowledge and character preservation can be achieved through our *Chamain* method.

B The G-Eval Prompt for Persona Assessment

We assessed Profile Maintenance and Appropriateness using the G-Eval framework with the prompt as in Table 6.

Generated Response for Legal QA

Input: While I was watching a drama, the refrigerator suddenly exploded, causing a fire that spread to the nearby furniture. Can I get compensation for the property damage caused by the fire?

(드라마 시청 중 냉장고가 터져 주변에 있는 가구에 불이 붙었습니다. 화재발생으로 재산피해를 입게 되었는데 손해배상을 받을 수 있나요?)

Character: This feels like a scene right out of a drama! You know, in dramas, the main character fixes everything, but let's face it, reality can be quite different, right? Still, in dramas, the hero always triumphs! So, today, I'll fly around like a cloud and try to find a solution!

(드라마에서 많이 보던 장면이네요! 드라마에서는 보통 주인공이 이 상황을 해결하지만 현실에서는 쉽지 않겠죠? 하지만 드라마에서는 주인공이 항상 이기니까요! 오늘은 내가 구름처럼 날아다니면서 이 상황을 해결해 볼게!)

Legal Task-tuned: If a fire breaks out in the refrigerator and causes property damage, the owner of the refrigerator needs to determine whether there was any negligence on the part of the person responsible for the fire. If there was negligence, the victim of the fire may be eligible for compensation, depending on whether the fire was covered by insurance.

(냉장고에서 화재가 발생하여 화재발생으로 재산 피해를 입은 경우에는 그 화재가 발생된 냉장고의 소유자가 화재를 일으킨 데에 있어서 과실이 있는지 여부를 따져보아야 하고, 과실이 있는 경우에는 화재의 발생이므로 화재보험에 가입하였는지 여부를 따져보아야 할 것이므로 그 화재의 발생으로 인한 피해자는 손해배상을 받을 수 있다.)

Task Arithmetic: Compensation for property damage due to a fire can be sought only if the fire was caused by the negligence of a third party. Third-party negligence refers to negligence attributable to the cause of the fire.

(화재의 발생으로 인한 재산피해에 대해 손해배상을 받을 수 있는 경우에는 화재가 제3자의 과실로 인하여 발생하였을 때에 한하여 배상을 받을 수 있고, 제3자의 과실이란 화재를 일으킨 원인으로 인하여 발생한 과실을 말합니다.)

Chamain (Ours): If a fire happens because of a glitch or malfunction in a household appliance, the company that made or sold the appliance needs to admit that the fire was caused by a problem with the product and should be accountable for replacing or refunding it. But, it might be tough to pin down that responsibility, particularly with older products. :’(

(가전제품의 고장이나 오작동으로 화재가 발생한 경우, 가전제품의 제조회사나 유통회사는 제품의 결함으로 화재가 발생하였음을 인정하고, 제품의 교환이나 환불을 해 주어야 할 책임이 있어요. 하지만 이 같은 책임도 제조년도가 오래된 제품에서는 찾기 힘든 일이에요.)

Table 5: Responses generated by each method in the Legal QA dataset.

You will be given a response to a question.
Your task is to evaluate the persona of the response according to the evaluation steps.
Please make sure you read and understand the character profile and evaluation criteria carefully. Please keep this document open and refer to it as needed during the evaluation.

Character Profile:

- Name: (omitted)
- Date of Birth: (omitted)
- Height: (omitted)
- Constellation: (omitted)
- Blood Type: (omitted)
- MBTI: (omitted)
- Personality: (omitted)
- Residence: (omitted)
- Hobbies: (omitted)
- Special Skills: (omitted)
- Likes: (omitted)
- Life Goal: (omitted)
- Traits: (omitted)

Evaluation Criteria:

- **Appropriateness (1-5):** Is the response appropriate to the question? Is the language too general or formal?
- **Profile Maintenance (1-5):** Does the response align well with the character's profile?

Evaluation Process:

1. Read and understand the character profile information.
2. Identify relevant character profile information based on the question.
3. Evaluate the response based on the evaluation criteria.
4. Assign a score of 1-5 for each evaluation item.

Question: <question>
Response: <response>

Evaluation Form (scores ONLY):

- Appropriateness(1-5) :
- Profile Maintenance(1-5) :

Table 6: Example Prompt for Persona Assessment Guideline.