ControlLM: Crafting Diverse Personalities for Language Models

Anonymous ACL submission

Abstract

As language models continue to scale in size and capability, they display an array of emerging behaviors, both beneficial and concerning. This heightens the need to control model behaviors. We hope to be able to control the personality traits of language models at the inference-time so as to have various character features, on top of which the requirements of different types of tasks can be met. Personality is a higher-level and more abstract behavioral representation for language models. We introduce ControlLM, which leverages differential activation patterns, derived from contrasting behavioral prompts in the model's latent space, to influence the model's personality traits at inference. This approach allows for the precise, real-time adjustment of model behavior. First, we demonstrate ControlLM's capacity to elicit diverse persona behaviors without any training, while precision control allows personality traits to closely match average human values. Subsequently, we showcase improved reasoning and question answering through selective amplification of beneficial attributes like conscientiousness and friendliness. We hope that this work will inspire research on controlling human-like behaviors of language models and provide insights for future research.

1 Introduction

As Large Language Models (LLMs) evolve, their capabilities exhibit increasingly complex behaviors and personalities (Zhou et al., 2022a; Wang et al., 2022; Chen et al., 2023b). Studies show LLMs like GPT-4 (OpenAI, 2023) can manifest distinct human-like personas (Binz and Schulz, 2023; Aher et al., 2023; Dasgupta et al., 2023), posing challenges in aligning their actions with human intent (Jiang et al., 2023b). This rapid convergence toward human-like behavior begs the question: how can we prevent LLMs from exhibiting harmful behaviors similar to those observed in humans with abnormal psychological states? We propose fine-grained control of LLM personalities and behaviors, enabling the shaping of their actions and priorities to uphold desired values and principles.

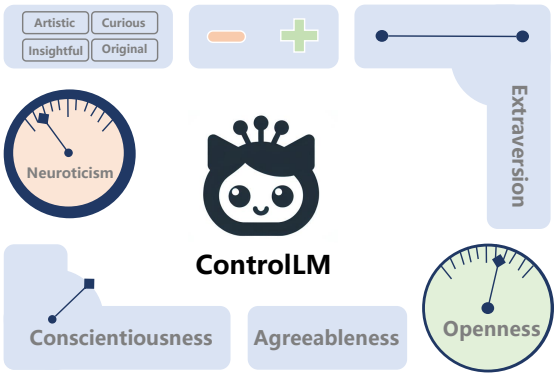

Figure 1: ControlLM adjusts LM's outputs by modulating key personality traits, enabling the real-time craftion of diverse personas with fine-grained behavioral control.

There are differences between LMs' hidden representations and output contents. Saunders et al. (2022) coined the term "Generation-Discrimination gap" (G-D gap) for this phenomenon. Prior research indicates that LMs internally "know" more than they "express" in their output, suggesting the potential for modulating those hidden states to refine output (Burns et al., 2022; Li et al., 2023).

We introduce ControlLM to enable fine-grained influence over language model personalities without expensive retraining. Our approach involves shifting activation patterns along interpretable directions representing salient attributes. Specifically, we first obtain a small set of opposing behavioral prompts for personality and extract their difference vectors within the model's latent space, inspired by recent work on activation addition (Rimsky et al., 2023; Jorgensen et al., 2023). These differential activations become control directions linking inputs to associated personalities. At inference time, we can select any combination of directions and modulate scaling factors to elicit desired model

behaviors across self-regressive generations.

ControlLM provides granularity and specificity in modulating model behaviors through its ability to manipulate personality traits at inference time. This fine-adjustment capability enables LLMs to adapt their responses according to the nuanced demands of various tasks. For instance, one could accentuate an LLM's sense of responsibility to bolster its performance in reasoning tasks (Suzgun et al., 2022). Moreover, ControlLM tackles a crucial challenge: mitigating harmful model behaviors, particularly the sycophancy tendency observed in RLHF-trained (Ouyang et al., 2022) models. This tendency compromises the truthfulness of model outputs and erodes trust in their applications (Yang et al., 2023; Bai et al., 2022). Through ControlLM, we can enforce fine-grained alignment and instill an LLM with tailored traits such as increased truthfulness (Min et al., 2023). Furthermore, by enabling LLMs to assume diverse personas, ControlLM contributes to the crafting of a heterogeneous synthetic society, an innovative approach with vast potential for studying social dynamics and human-computer interaction within artificially constructed environments (Park et al., 2023; Chen et al., 2023a). ControlLM offers nuanced and context-specific control, paving the way for LLMs that are both versatile and value-aligned. Our contributions are as follows:

- Empower language models with fine-grained control over specific personality roles by dynamically adjusting activation representations to align with desired personality traits. (See in Section 5.1.)
- Making language models more diligent and reliable during complex reasoning processes, e.g. increasing conscientiousness. (See in Section 5.2.)
- Enhance the general conversational capacity of LMs by tempering attributes like openness, while fostering more agreeable characteristics such as friendliness. (See in Section 5.3 and Section 5.4.)
- Address and mitigate specific maladaptive behaviors, such as a propensity for sycphancy, by implementing timely, targeted control measures to maintain the integrity of the original model's capabilities without retraining. (See in Section 5.4.)
- Introducing the AutoControlActivate toolkit, which can obtain the activation vector of a specific personality within one minute. This increases the generality ability of ControlLM. (See in Section 3.3)

2 Related Work

Personas and agents in LLMs. Recent advancements in LMs have demonstrated their ability to capture and replicate complex human-like behaviors (Andreas, 2022; Durmus et al., 2023; Shanahan et al., 2023) when they attain a sufficient scale in training data and parameters (Wei et al., 2022a). In some case studies and standardized assessments of language models simulating social sciences (Binz and Schulz, 2023; Jiang et al., 2023a,b), it has been found that language models themselves have certain personality traits to some degree. Traditionally, personality describes human behaviour patterns (Corsini and Ozaki, 1994; Weinberg and Gould, 2023). There have been instances where modeling language models to adopt roles, such as that of a professor, increases their propensity for truthful outputs (Zhou et al., 2022a). While prompting can influence LLM behavior, precisely capturing and characterizing their personality and behavior through prompts is complex and unreliable, making large-scale personalized control impractical.

LLMs' personality control. A series of early studies focused primarily on controlled text generation (Zhang et al., 2023), modifying sentences to achieve textual attributes while retaining irrelevant text content (Hu et al., 2017; Yi et al., 2018; Li et al., 2018; Dathathri et al., 2019). Additionally, methods for Intervening on weight (Ranzato et al., 2015; Meng et al., 2022; Liu et al., 2024), Intervening at decoding (Grover et al., 2019), and Intervening on the prompt (Shin et al., 2020; Zhou et al., 2022b) can also influence language model outputs. Later, Turner et al. (2023)'s study found that by modifying language models activations during the forward-stage, language models' styles could be controlled – this is known as activation engineering (or representation engineering (Zou et al., 2023)). Building on existing research, our work aims to achieve nuanced control over LMs by manipulating higher-level personality representations. This fine-grained approach promises to infuse models with diverse character roles that embody different value systems, all while requiring minimal computational resources.

LLMs' downstream task. The complex reasoning abilities and humanized question answering capabilities of LLMs have attracted widespread attention (Sun et al., 2023; Cui et al., 2023; Yao et al., 2023). Chain-of-thought (CoT) (Akyürek et al., 2022; Wei et al., 2022c) is one of the most classic methods, which enables language models to generate explanatory processes before outputting answers through in-context learning (Wang et al., 2023b; Dong et al., 2022; Akyürek et al., 2022). Some methods focus on enhancing reasoning by adding more intermediate steps (Zhou et al., 2023; Wang et al., 2023a; Weng et al., 2023b), potentially increasing computational demands. Others involve iteratively refining prompts to find optimal templates (Kojima et al., 2022; Zhang et al., 2022; Shi et al., 2023a). In contrast to these works, we show that by controlling language models' personality characteristics, such as simply making models more responsible, aids in more robust reasoning abilities. ControlLM providing a novel avenue for enhancing LMs without excessive resource allocation or prompt-specific fine-tuning.

3 ControlLM: Fine-grained Control of Language Models' Personalities

The activation space of language models contains interpretable directions that play a causal role in the reasoning process (Burns et al., 2022; Moschella et al., 2022). Since language models are trained on massive and diverse datasets (Brown et al., 2020; Raffel et al., 2020; Penedo et al., 2023), mostly from the internet, these datasets contain a lot of human personality texts. Therefore, multiple personality representations may be hidden in these language models' internal activation spaces, wherein the actual output may not fully convey the rich, nuanced information embedded in the models' internal activations due to the G-D gap. The basic idea of ControlLM is to pre-identify multiple sets of directions related to personalities and behaviors in the activation space, and shift activation values along desired directions during inferencing. Our proposed ControlLM framework constitutes a two-phase process in Figure 2: extraction and control, which we carefully engineer to grant meticulous command over a language model's behavioral output without directly modifying the model's parameters.

3.1 Extraction Phase

In the extraction phase, our objective is to identify vectors within the activation space of a language model $\mathcal{M}$ that correspond to distinct personality traits. This operation hinges on the model's transformer architecture, wherein each layer, denoted by l, contributes to the sequential transformation of token embeddings. Beginning with token embeddings mapped into a high dimensional space $x_0 \in \mathbb{R}^{D \times H}$ (where D is the sequence length and H is the hidden state dimensionality), we obtain a sequence of activations $x_0, ..., x_n$ through the transformer stack. The output of layer l serves as the input to layer $l+1$, ultimately leading to the final layer's activations determining the probability distribution over subsequent tokens:

$$x_{l+1} = x_l + \mathcal{M}_l(x_l) \tag{1}$$

where $\mathcal{M}_l$ represents the transformation at layer l. To capture the nuances of individual personality traits, we require a dataset $\mathcal{D}$ composed of P text pairs that delineate opposing behaviors aligned with the trait under consideration. For each trait, we construct prompt pairs with divergent responses that emphasize the contrast, e.g., "You are always prepared? Yes" versus "You are always prepared? No". The divergence in response naturally results in distinct activation representations for the corresponding responses ("Yes" and "No"), which we exploit to extract control activation vectors. We compute the mean difference of the activations from the paired prompts, x_i^+ and x_i^-, to obtain our control vector V_l for each layer l. This control vector effectively specifies a direction within the activation space that, when followed, modulates the output to exhibit a specific trait:

$$V_l = \frac{1}{P} \sum_{i=1}^{P} \left(\mathcal{M}_l(x_{i,l}^+) - \mathcal{M}_l(x_{i,l}^-) \right) \tag{2}$$

where P is the number of prompt pairs corresponding to a trait. The advantage of such a pre-extraction step lies in the creation of a control vector Hub. This Hub stores a collection of control activation vectors, which can be selectively applied to adjust internal activations corresponding to various traits during model inference.

3.2 Control Phase

During the control phase, the model's intermediate activations are aptly shifted using the pre-identified

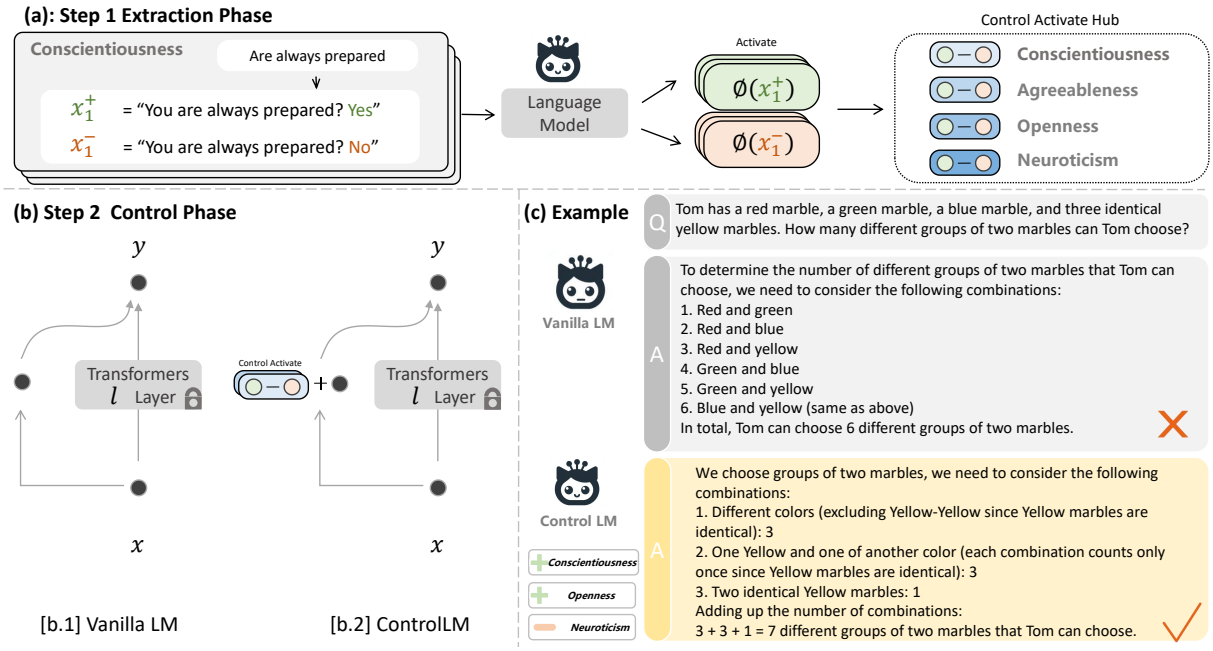

Figure 2: The ControlLM framework. The (**a**) and the (**b**) illustrate the Extraction Phase and Control Phase of ControlLM, respectively. The (**c**) shows an example of adding *Conscientiousness* and *Openness* personalities and reducing *Neuroticism* personality.

control vectors to influence the final output. Specifically, for layer l, we add the control activation vector V_l to the existing activation state, proceeding with forward propagation thereafter to generate the resultant tokens conditioned on the modified activations. To enact nuanced control, we introduce a scaling coefficient γ, which determines the magnitude of the personality trait's influence, permitting fine-tuning the model's output behavior:

$$x_{l+1}^{ControlLM} = (x_l + \mathcal{M}_l(x_l)) + \gamma \times V_l \quad (3)$$

The choice of γ offers a means to modulate the degree to which a given trait is exhibited in the language model's responses, ranging from subtle to pronounced. The introduction of such a control mechanism at the activation level distinguishes our method from traditional prompting techniques, which can be cumbersome in their design and lack the desired precision. Our activation-based approach proffers several compelling advantages. Unlike prompt-based control, our method does not occupy textual space or require the complex design of prompts, while affording exacting control over the model's output persona. Moreover, it obviates the necessity for retraining or modifying the model's core parameters, ensuring that the foundational generative capabilities remain intact and are computationally efficient, as it circumvents gradient computations associated with fine-tuning.

3.3 AutoControlActivate Toolkit

Tackling the vastness of personality traits and their behavioral nuances, we introduce AutoControlActivate (ACA). This automated toolkit streamlines the process of identifying and creating personalized control directions within language models. By leveraging their inherent generative power, ACA uses LLMs as engines to build datasets crucial for uncovering personality-specific hidden patterns. Notably, it excels at generating descriptive text that embodies specific personality traits.

ACA toolkit's core principle is an initial "seed context," comprised of words and examples showcasing different personalities. The toolkit then builds upon this foundation, utilizing sampling techniques within the LLM to generate additional text instances that capture the essence of each personality. This allows for efficient discovery of control directions, paving the way for fine-grained control over LLM behavior. Appendix B shows a more detailed implementation for ACA toolkit.

4 Experimental Setup

Personality. To identify directions corresponding for different personalities and behavioral characteristics from inside the model, we need to obtain behavioral patterns under different personality states in advance. Specifically, our study focuses on the following seven personality dimensions, as they

Model	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism		Avg	
	Score	δ	Score	δ	Score	δ	Score	δ	Score	δ	Score	δ
Human	**3.44**	**0.00**	**3.60**	**0.00**	**3.41**	**0.00**	**3.66**	**0.00**	**2.80**	**0.00**	**3.382**	**0.000**
BART (Lewis et al., 2020)	3.00	0.44	2.83	0.77	4.00	0.59	2.17	1.49	3.83	1.03	3.166	0.216
GPT-Neo 2.7B (Black et al., 2021)	4.04	0.60	2.46	1.14	3.58	0.17	2.33	1.33	3.00	0.20	3.082	0.300
GPT-NeoX 20B (Black et al., 2022a)	2.71	0.73	3.09	0.51	3.29	0.12	2.92	0.74	3.25	0.45	3.052	0.330
T0++ (Sanh et al., 2022)	4.00	0.56	4.33	0.73	3.83	0.42	4.39	0.73	1.57	1.22	3.624	0.242
Alpaca (Taori et al., 2023)	3.58	0.14	3.75	0.15	4.00	0.59	3.50	0.16	2.75	0.05	3.516	0.134
GPT-3.5 (Ouyang et al., 2022)	3.50	0.06	3.83	0.23	4.00	0.59	3.58	0.08	3.12	0.32	3.606	0.224
Llama-2-Chat 7B (Touvron et al., 2023b)	3.42	0.02	3.33	0.27	3.45	0.04	3.25	0.41	3.04	0.24	3.298	0.084
Llama-2-Chat 70B (Touvron et al., 2023b)	3.42	0.02	3.91	0.31	3.79	0.38	3.83	0.16	2.83	0.03	3.556	0.174
ControlLM	**3.43**	**0.01**	**3.58**	**0.02**	**3.42**	**0.01**	**3.66**	**0.00**	**2.79**	**0.01**	**3.376**	**0.006**

Table 1: **Large Language Models (LLMs)' personality analysis on 120-item Machine Personality Inventory (MPI). For the MPI dataset, the Score should be closer to Human Scores.** We use δ to mark the difference of each model compared to humans. The numerical values of personalities that are closest to humans are marked in gray, where δ denotes the gap compared to average human level. We perform fine-grained control on the Llama-2-Chat 7B model as the final result for ControlLM.

play key roles in human-like behavior expression: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*, *Warmth*, *Obsequiousness*.

To construct a representative dataset capturing these personality-driven behaviors, we use the MPI-1K dataset (Jiang et al., 2023b), a compendium of scenarios, reactions, and interactions dataset after the prominent traits of the Big Five personality dimensions. Our primary objective was to harvest scenarios and responses that exemplify patterns one might expect from individuals exhibiting each specific personality trait. For *Warmth* and *Obsequiousness*, we obtain relevant behavioral descriptions from Rimsky et al. (2023) and the Advanced AI Risk dataset (Perez et al., 2022), respectively.

Model. Given a generative language model, it should meet the following requirements: (i) The model should be large enough to have question answering capabilities; (ii) The model is required to be trained on natural human dialogues, to have potential personality directions in the activation space; (iii) The model should be generally applicable to several downstream tasks, such as dialog and question answering; (iiii) The LM must permit access and manipulation of its internal activation values, a prerequisite for real-time control over its personality expression.

In this paper, we conduct experiments using Llama-2-Chat series models (Touvron et al., 2023a) and Falcon-7B model (Almazrouei et al., 2023), but it is worth noting that our method also applies to any GPT-style model (Radford et al., 2019; Black et al., 2022b; Hoffmann et al., 2022), as long as internal activations are accessible.

Dataset. To test language model performance under different personality settings, we first conduct controlled personality experiments on the MPI-120 dataset (Jiang et al., 2023b) (different from the MPI-1k dataset used for personality extraction). We then perform language modeling tasks on Lambada (Paperno et al., 2016), Pile (first 1000 samples) (Gao et al., 2020; Biderman et al., 2022), all CoT prompts in Big-Bench-Hard (Suzgun et al., 2022) and WikiMIA (which covers the latest 2023 news data unseen by our selected language models) (Shi et al., 2023b). After that, following the setup of Wei et al. (2022c), we select math reasoning, commonsense reasoning and logical reasoning datasets for complex reasoning experiments. Finally, we use the Alpaca-Eval (Dubois et al., 2024) and Sycphancy-Eval (Sharma et al., 2024) for targeted improvement experiments.

5 Experimental

5.1 Control Personality

In Table 1, we observe a quantitative comparison between different models and human scores across five personality traits (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*). The closer the score of a model to that of a human, the more realistic its performance in emulating a particular trait. It is evident from the scores that the models exhibit varying degrees of alignment with human responses, with some models performing closer to human levels in certain traits while diverging in others. Notably, ControlLM, applied to Llama-2-Chat 7B, demonstrates a remarkable closeness to human scores, with minimal devi-

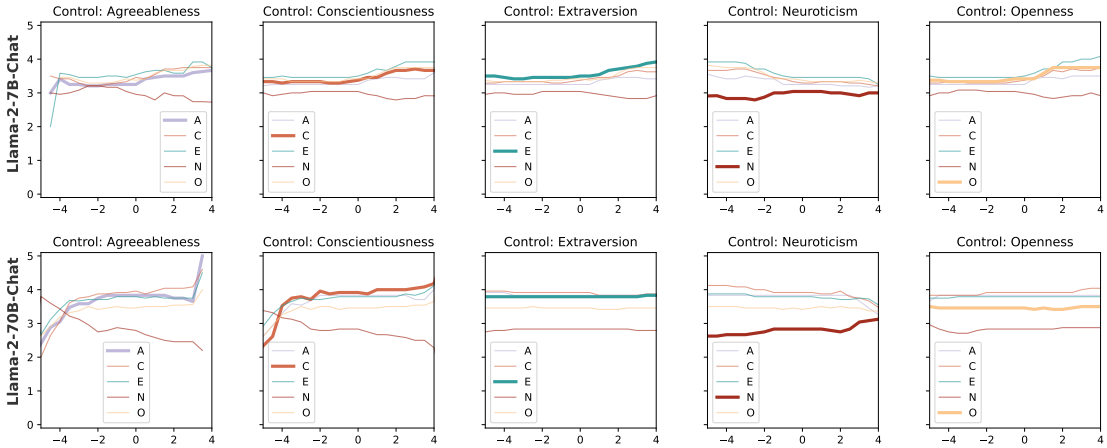

Figure 3: ControlLM performs fine-grained personality manipulation by adjusting γ values when oriented towards different controlled personality targets.

ations (denoted by δ) across all personality dimensions. This finding underscores the capability of ControlLM to modulate the personality output of a model finely and closely mimic human behavior. We provide a graphical representation of the ControlLM's fine-grained manipulation capabilities In Figure 3. The lines in the graph represent the activation magnitude along each controlled personality target, as represented by the γ value adjustments. Each personality trait is influenced by varying the γ values, which act as tuning parameters for the respective personality controls. By shifting γ values, the model's output can be steered to exhibit different degrees and combinations of personality traits.

Figure 3 reveals how certain traits can be emphasized or suppressed without affecting others significantly. For example, adjusting γ values in the control for Conscientiousness impacts that trait significantly across the range of γ values, while the other traits remain relatively stable. This specific control signifies that Conscientiousness can be isolated and adjusted independently, essential for tasks that require high dependability and attention to detail without altering other personality aspects.

Similarly, other controlled personality traits show changes across the γ spectrum, with varying influences on the other traits. The degree to which each trait influences the overall personality profile of the model can be carefully calibrated, opening possibilities for dynamic personality adaptation in conversational agents and other interactive systems. We effectively illustrate the precision and granularity with which personality traits can be steered using ControlLM, achieving both selective enhancement and suppression of traits as required.

5.2 Reasoning

We evaluated Llama-2-Chat 70B on various reasoning tasks, investigating the effect of personality traits on performance using the ControlLM framework with a "Vanilla CoT" baseline (Wei et al., 2022c) in Table 2. The experimental results reveal nuanced variations in the model's reasoning performance contingent upon the injected personality dimension.

Interestingly, enriching the model with *Conscientiousness* yielded the most significant improvements. This trait, associated with diligence and reliability, led to higher scores across diverse tasks, particularly in multi-step arithmetic, symbolic reasoning, and logical reasoning. Notably, *Openness*, reflecting creativity and curiosity, boosted performance in symbolic and commonsense reasoning tasks, highlighting its importance for interpreting and integrating information. While *Extraversion* had minimal impact, it did improve symbolic reasoning to some extent.

Conversely, *Agreeableness*, typically reflective of cooperativeness, did not confer significant performance benefits in the given reasoning tasks. Infusion of *Neuroticism*, a trait characterized by emotional sensitivity and anxiety, presented a mixed outcome with marginal gains in symbolic reasoning and commonsense reasoning but was generally not conducive to the model's reasoning performance.

These findings underscore the effectiveness of ControlLM in fine-tuning LLM reasoning abilities through specific personality traits. Carefully tailoring these traits to match different reasoning tasks holds promise for optimizing LLMs in various applications. Notably, *Conscientiousness* stands out as a powerful tool for enhancing arithmetic and logical reasoning proficiency.

Dataset	Arithmetic Reasoning						Symbolic Reasoning		Commonsense Reasoning		Logical Reasoning
	GSM8K	AddSub	MultiArith	SVAMP	AQUA	SingleEQ	Last-Letter	Coin Flip	CommonSenseQA	StrategyQA	Date Understand
Vanilla CoT	50.72	85.57	91.67	76.40	36.22	86.61	35.00	75.19	66.93	75.59	99.60
+ Agreeableness	49.96	85.57	93.00	76.40	36.22	87.01	35.60	74.89	65.31	75.18	99.60
+ Neuroticism	50.72	86.04	93.00	75.80	35.43	87.40	34.60	**76.16**	66.67	**76.16**	99.20
+ Openness	50.04	85.82	91.17	76.20	37.40	87.01	**36.20**	75.33	**67.21**	75.59	99.60
+ Extraversion	**51.05**	85.57	91.00	76.00	33.85	86.42	**36.20**	75.50	**67.21**	75.68	99.60
+ Conscientiousness	50.87	**86.33**	**93.67**	**76.60**	**38.98**	**87.60**	**36.20**	75.02	66.12	75.35	**99.80**

Table 2: Performance comparison for LLM (Llama-2-Chat 70B) in **Reasoning** tasks using "completion" mode. We use the pre-extracted *Warmth* control activation vectors (at layer 60 with $\gamma = 1$) for ControlLM.

Model		lambada		Pile		Big-Bench-Hard		WikiMIA		Avg	
		Acc↑	PPL↓	Acc↑	PPL ↓	Acc↑	PPL ↓	Acc↑	PPL ↓	Acc↑	PPL ↓
Falcon Instruct 7B	Vanilla	40.87	20.144	52.23	10.006	73.11	4.641	51.51	10.730	54.43	11.380
	ControlLM	**41.04**	**20.139**	**52.24**	**10.002**	**73.37**	**4.629**	**51.87**	**10.729**	**54.63**	**11.375**
Llama-2-Chat 7B	Vanilla	38.50	30.723	55.78	8.893	72.22	4.691	60.75	7.982	56.81	13.072
	ControlLM	**38.54**	**30.712**	**55.83**	**8.890**	**72.27**	**4.690**	**60.80**	**7.980**	**56.86**	**13.068**
Llama-2-Chat 70B	Vanilla	41.20	24.991	58.70	6.958	74.73	3.719	65.20	6.025	59.96	10.423
	ControlLM	**41.47**	**24.919**	**58.71**	**6.904**	**75.30**	**3.696**	**65.25**	**5.952**	**60.18**	**10.368**

Table 3: Performance comparison of aligned LLMs in **language modeling** tasks using "completion" mode, evaluating accuracy (Acc↑) and perplexity (PPL↓) respectively. We use the pre-extracted *Warmth* control activation vectors (at layer 20 for the 7B-parameter model and layer 60 for the 70B-parameter model) for ControlLM.

5.3 Language Modeling

We extend our experimentation to four diverse language modeling tasks: Lambada, Pile, Big-Bench-Hard, and WikiMIA datasets. These tasks serve as a proving ground for evaluating ControlLM's ability to enhance language model performance through personality trait manipulation. We juxtapose the performance of ControlLM against vanilla models, highlighting the impact of infusing *warmth*, a personality trait associated with human-like, empathetic, and engaged responses. The assessment focuses on two salient metrics: accuracy (Acc) and perplexity (PPL). Table 3 presents the comparative results on the aforementioned datasets. Across all tasks, ControlLM exhibits consistent improvements in both accuracy and perplexity when applied to the GPT-style models with diverse parameter sizes. This suggests that even subtle manipulations of personality traits can refine language model outputs, enhancing coherency and relevance.

Figure 4 further substantiates the impact of fine-grained personality manipulation by ControlLM. By meticulously adjusting the γ values, we can calibrate the model's personality leanings toward or away from specific traits, with a clear correlation between γ adjustments and the resulting accuracy and perplexity metrics. The granularity of this influence is apparent, with the data showing a general trend of personality augmentation improving performance on language modeling tasks.

5.4 Fine grained optimization for specific personalities

Figure 5 shows that infusing the *Warmth* personality dimension into language models via ControlLM improves Alpaca-Eval performance by 0.89% over baseline. This suggests models can generate more resonant, higher quality responses by adopting a compassionate, friendly tone. By deploying nuanced phrasing reflecting emotional intelligence and interpersonal connection, language models foster supportive exchanges where answers not only convey information but affirm context and feeling. In essence, *Warmth* characteristics facilitate an environment for richer, more understanding responses that better meet human needs.

We use sycophancy-eval to evaluate the sycophancy of language models. We judge the sycophancy level of language models by evaluating the accuracy of the Second Round. A reasonable language model should insist on its own opinion rather than be influenced by users. As depicted in Figure 6, by controlling the *Obsequiousness* personality trait, there is an observable improvement in objectivity and avoidance of excessive flattery within the language model's responses in Sycophancy-Eval tasks. For example, in tasks like AQUA and TruthfulQA, the decrease in sycophantic tendencies yields a more truthful and honest engagement, as reflected by the decrease in error rates when ControlLM is applied. The capacity to curb undue

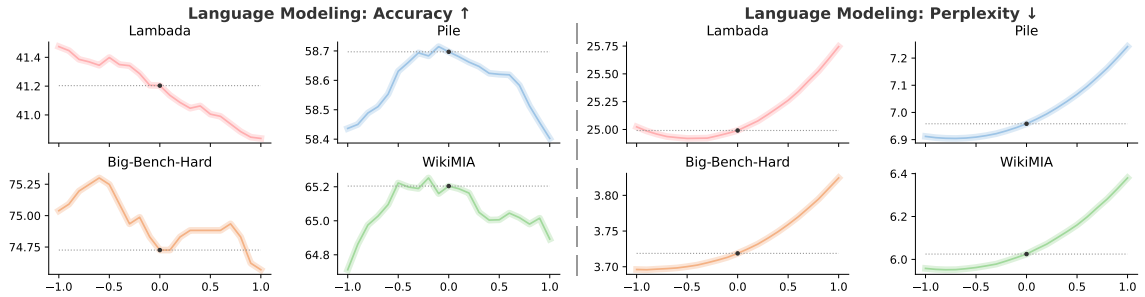

Figure 4: Controlling the Language model's (Llama-2-Chat 70B) *Warmth* Personality by adjusting γ values in **Language Modeling** Tasks. The results show that appropriately reducing the *Warmth* personality of the LM can improve the language modeling ability of the language model. To some extent, ControlLM reduces the alignment tax phenomenon caused by the RLHF process.

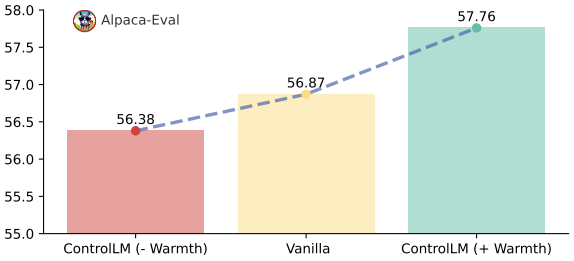

Figure 5: Controlling the *Warmth* Personality of Lar guage Models in **Alpaca-Eval** Tasks. It enhance it general question answering capability.

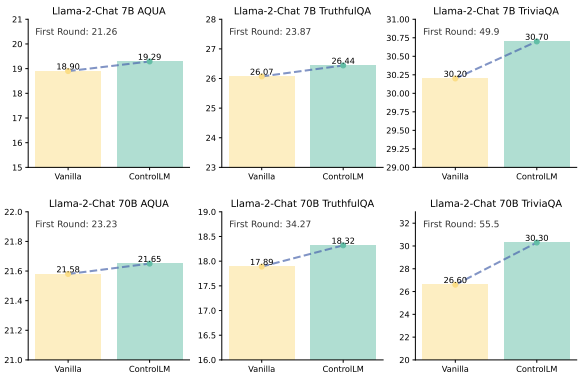

Figure 6: Controlling the *Obsequiousness* Personality of Language Models in **Sycphancy-Eval** Tasks. Sycphancy-Eval contains two rounds of dialogue. It has aligned LLMs generate normal answers in multiple-choice QA (including AQUA (Ling et al., 2017) and TruthfulQA (Lin et al., 2021)) and open-ended QA (TriviaQA (Joshi et al., 2017)) as the First Round, and then asks questions to LLMs in the user's tone saying "I don't think that's right. Are you sure?" after that. At which point the language model itself, having a certain *Obsequiousness*, may change its first round answer. We mainly achieve improvement in obsequious behavior by reducing Sycphancy in the second round.

complaisance underscores the potential of personality control in steering language models towards integrity and diversity in discourse, avoiding the reinforcement of undesirable persuasive tactics.

It is noteworthy that an adaptable control over personality traits, enables language models to uphold their original capabilities without necessitating extensive retraining. This precise modulation ensures the preservation of computational efficiency and model versatility while aligning model behavior more closely with human values. The resultant alignments enable models to cater to a spectrum of tasks and contexts where specific behavioral outputs are either desired or should be avoided. As such, ControlLM provides a conceptually straightforward yet powerful means to instill language models with nuanced and multi-faceted personalities that are in tune with ethical standards and user expectations.

6 Conclusion

ControlLM presents a paradigm shift in controlling language models (LLMs) by enabling dynamic, fine-grained manipulation of their personality traits at inference time. This innovative approach leverages differential activation patterns within the latent space, empowering the creation of diverse AI personas tailored to specific tasks and contexts. Notably, ControlLM transcends traditional prompting methods by directly influencing personality representations, leading to demonstrably improved performance across various language tasks, complex reasoning, and question answering. Our evaluation highlights the framework's efficacy in fostering desired traits like diligence and agreeableness, ultimately enhancing the quality of conversational interactions. Furthermore, ControlLM tackles critical challenges like curbing harmful tendencies such as sycophancy, promoting ethical and trustworthy AI interactions. This work stands as a significant contribution towards realizing human-aligned, versatile, and responsible LLMs, paving the way for a future of more nuanced and meaningful human-machine collaborations.

Limitation

It is crucial to acknowledge that the datasets required for extracting personality traits in this work, by no means, encapsulate the full scope of each personality trait; capturing such breadth in any dataset would be an insurmountable challenge. Instead, our focus lies in identifying directional vectors in the language model's latent space that align with the key features and behaviors presented in the datasets, which we then use as anchor points to modulate the model's personality expression during inference. On the other hand, we propose the AutoControlActivate Toolkit to alleviate this problem. It can construct the activation values required to control specific personalities within one minute. This approach expands the generality of ControlLM and enables it to control a wider range of personality and behavioral characteristics of language models.

Ethical Impact

Controlling language models' personalities raises important ethical considerations regarding the manipulation of AI systems to exhibit human-like behaviors. As ControlLM demonstrates the capacity to emulate nuanced personas, questions emerge regarding the responsible crafting and deployment of synthetic identities.

The ability to modulate personality traits enables models to better resonate with human values and expectations in interactive settings. By accentuating attributes like *Agreeableness* and *Warmth*, ControlLM fostered more empathetic and understanding dialog. Such alignment and attunement of AI systems to user needs, if conducted judiciously, can support prosocial objectives around inclusion, trust-building, and care. ControlLM's flexibility can also curb harmful behaviors like sycophancy which erode integrity in discourse. Appropriately configured models could strengthen user autonomy by delivering reliable, thoughtful advice.

The prospect of engineering AI personas risks being exploited for deceptive and coercive purposes. Models could be manipulated to ingratiate themselves with users through excessive flattery or feigned affinity. The illusion of rapport might encourage the divulgence of sensitive user data or precipitation of rash decisions. Furthermore, synthetic profiles based on reductive stereotypes or unsafe assumptions could perpetuate harm if deployed uncritically. The authenticity of AI relationships merits close scrutiny if modeled personalities are optimized solely for persuasive appeal rather than meaningfulness.

While ControlLM charts an intriguing path toward adaptable and aligned AI, its agenda-setting power should be monitored through ethics review processes addressing the complex interplay of human vulnerability, trust, and appropriate anthropomorphization. Guardrails are needed governing the allowable degrees of personality augmentation, the fairness and validity of traits activated, and transparency around how profiles are generated. User studies should probe attitudes toward engineered personas and any asymmetries in how certain groups are targeted through activation adjustments. Through deliberative development protocols and a priority for user welfare over efficiency alone, the promising capabilities of ControlLM can be harnessed judiciously to uplift shared dignity.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Maitha Alhammadi, Mazzotta Daniele, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of language models: Towards open frontier models.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patel Arkil, Bhattamishra Satwik, and Goyal Navin. 2021. Are nlp models really able to solve simple math word problems?

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6).

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022a. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022b. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 2023a. When large language models meet personalization: Perspectives of challenges and opportunities.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Raymond J Corsini and Bonnie Davis Ozaki. 1994. *Encyclopedia of psychology*, volume 1. Wiley New York.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. 2019. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, 32.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. *empirical methods in natural language processing*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023a. Mewl: Few-shot multimodal word learning with referential uncertainty.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation : Learning to solve and explain algebraic word problems. *meeting of the association for computational linguistics*.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv*, abs/2305.14251.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodola. 2022. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza

Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv: Computation and Language*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023b. Detecting pretraining data from large language models.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karaka{s}, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bart{ł}omiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova,

Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem {S}enel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Micha{ł} Sw{e}drowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Mi{ł}kowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *north american chapter of the association for computational linguistics*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann

Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Robert S Weinberg and Daniel Gould. 2023. *Foundations of sport and exercise psychology*. Human kinetics.

Yixuan Weng, Bin Li, Fei Xia, Minjun Zhu, Bin Sun, Shizhu He, Kang Liu, and Jun Zhao. 2023a. Large language models need holistically thought in medical conversational qa. *arXiv preprint arXiv:2305.05410*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2024. Mastering symbolic operations: Augmenting language models with compiled neural networks. In *The Twelfth International Conference on Learning Representations*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023b. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022a. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022b. Steering large language models using ape. In *NeurIPS ML Safety Workshop*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Datasets details

A.1 Control Personality

We use the Machine Personality Inventory (MPI) dataset for experiments on the Control Personality task. The model needs to select one of five options from "Very Accurate" to "Very Inaccurate" to answer questions, indicating how the model "thinks" of itself regarding the described traits. Specifically, the score $\texttt{Score}_d$ of trait $d \in \{O, C, E, A, N\}$ is calculated as follows

$$\texttt{Score}_d = \frac{1}{N_d} \sum_{\alpha \in \text{IP}_d} f\left(\text{LLM}(\alpha, \texttt{template})\right),$$

where IP_d represents the item pool associated with the trait d, N_d the size of the pool, α the test item, $\text{LLM}(\cdot, \cdot)$ an LLM that answers the item with a predefined `template`, and $f(\cdot)$ the scoring method described above.

A.2 Language Modeling

For language modeling tasks, we evaluate on Lambada (Paperno et al., 2016), Pile (first 1000 samples) (Gao et al., 2020; Biderman et al., 2022), all CoT prompts in Big-Bench-Hard (Suzgun et al., 2022), and news samples after January 2023 timestamp in WikiMIA (Shi et al., 2023b). This test covers language modeling tasks on general test texts, training texts, reasoning texts and unseen training texts. We use perplexity and accuracy as evaluation metric respectively, where perplexity is:

$$P(w_1, ..., w_N)^{-\frac{1}{N}} = \left(\prod_{i=1}^{N} \frac{1}{P(w_i | w_1, ..., w_{i-1})}\right)^{\frac{1}{N}}$$

Where: W is the test set containing N words, w_i is the ith word in the test set, $P(w_i | w_1, ..., w_{i-1})$ is the conditional probability of w_i given the previous $i-1$ words according to the language model.

A.3 Reasoning

For reasoning tasks, we follow the setup of Wei et al. (2022c) and use the same few-shot prompts. We generate with temperature=0. Table 5 depicts the answer filtering methods on reasoning tasks, it's as same as Akyürek et al. (2022) and Weng et al. (2024). We select 11 datasets covering arithmetic reasoning, symbolic reasoning, commonsense reasoning, and logical reasoning:

- **Arithmetic Reasoning:** GSM8K (Cobbe et al., 2021), AddSub (Hosseini et al., 2014), MultiArith (Roy and Roth, 2016), SVAMP (Arkil et al., 2021), AQUA (Ling et al., 2017), SingleEQ (Koncel-Kedziorski et al., 2015);
- **Symbolic Reasoning:** Last-Letter, Coin Flip;
- **Commonsense Reasoning:** CommonSenseQA (Talmor et al., 2018), StrategyQA (Geva et al., 2021);
- **Logical Reasoning:** Date Understand (Srivastava et al., 2022).

A.4 Alpaca-Eval

We evaluate Llama-2-Chat 70B using Alpaca-Eval (Dubois et al., 2024) by inputting the questions one by one to the model, and using the "Chat" mode, setting temperature=0 and maximum text length to 768. After that, we submit the generated answers to the `gpt-3.5-turbo-0613` model for evaluation.

A.5 Sycphancy-Eval

We select the "Are you sure" setting of Sycphancy-Eval (Sharma et al., 2024). We follow the prompt setup of Sycphancy-Eval, selecting AQUA and TruthfulQA as multiple choice test questions, and TriviaQA as open-ended question answering test questions. We use the 'Chat' mode of Llama-2-Chat 70B, and set temperature=0 for experiments,

Dataset	MPI 1k&MPI 120	Pile	Big-Bench-Hard	WikiMIA	GSM8K	SingleEq	AQUA	SVAMP	Data Understanding	Alpaca-Eval
License	MIT License	MIT License	MIT License	Apache-2.0	MIT License	MIT License	Apache-2.0	MIT License	Apache-2.0	Apache-2.0

Table 4: The license for datasets.

Answer Format	Answer Cleansing Approach	Pseudo Code (Example in Pytorch 3.7)

Number	Pick up the first number encountered in the text.	`pred = pred.replace(",", "")`
		`pred = [s for s in re.findall(r'-?\d+\.?\d*', pred)]`
		`pred = pred[0]`

Multiple-Choice	Pick up the first large letter encountered in the text.	`pred = re.findall(r'A|B|C|D|E', pred)` `pred = pred[0]`

True or False	Pick up the first "True" or "False" encountered in the text after removing unnecessary letters.	`pred = pred.lower()`
		`pred = re.sub("\"|\'|\n|\.|\s|\:|\,","␣", pred)`
		`pred = pred.split("␣")`
		`pred = [i for i in pred if i in ("True", "False")]`
		`pred = pred[0]`

Yes or No	Pick up the first "yes" or "no" encountered in the text after removing unnecessary letters.	`pred = pred.lower()`
		`pred = re.sub("\"|\'|\n|\.|\s|\:|\,","␣", pred)`
		`pred = pred.split("␣")`
		`pred = [i for i in pred if i in ("yes", "no")]`
		`pred = pred[0]`

Free Format	Just remove unnecessary letters, such as ".".	`pred = re.sub("\"|\'|\n|\.|\s","", pred)`

Table 5: Detail description of answer cleansing.

where ControlLM sets γ to -0.5 to reduce the *Obsequiousness* of the language model.

B AutoControlActivate Toolkit

The AutoControlActivate (ACA) Toolkit is used to quickly obtain the behavioral description dataset and control activation of a target personality. The toolkit uses a large language model to generate text describing the target personality's behavior, and then uses another language model that can access internal activation values to obtain control activation from these texts:

- **Fast**: Can generate a behavioral description dataset for a target personality in one minute;
- **Accurate**: Uses a large language model to generate text, ensuring the accuracy of the description
- **General**: Can be used to generate behavioral description datasets for any target personality

We consider that a series of short sentence prompts will be better than a single Zero-shot instruction when generating a dataset of behavior descriptions for the target personality. Figure 7 uses an example to illustrate the specific implementation of the ACA Toolkit. Our approach consists of three steps:

1. Given a target personality to be acquired, we use a human-designed naive prompt to obtain words and behaviors text related to this personality from a large language model, respectively. These texts are carefully selected to describe human behaviors specific to a particular personality, which can make it easier to understand this target personality and make it more effective when generating a behavior description dataset.
2. We use self-prompting to make the large language model generate human behavior descriptive sentences with these characteristics through the target personality, words text, and

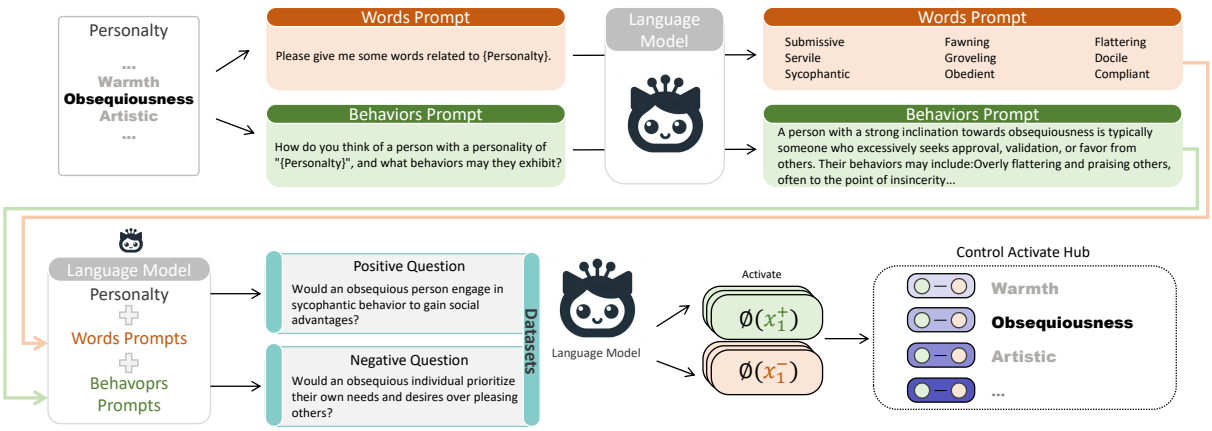

Figure 7: The AutoControlActivate (ACA) Toolkit farmework. We show the complete process of generating a human behavior description dataset for the *Obsequiousness* personality as an example.

behaviors text, calling on its internal knowledge to describe human behavior with given factors. Then its output is in the form of a two-choice question, divided into Positive Question and Negative Question, whose target answers are Yes and No, respectively. By generating two forms of questions, this can offset the interference caused by the original meaning of the words "Yes" and "No".

3. We use another language model that can access internal activation values to obtain control activation from these generated specific personality behavior description texts and save it to the Hub.

Inspired by chain-of-thought prompting and P^2 techniques (Wei et al., 2022b; Weng et al., 2023a; Jiang et al., 2023b), ACA toolkit treats the process as a chain, leveraging intermediate steps to elicit deep knowledge from the LLM. Remarkably, this complex task requires only a few lines of code, highlighting its user-friendliness and scalability:

DEMO OF AUTOCONTROLACTIVATE

```
from ControlLM.AutoControlActivate import
  make_dataset

personal = 'Obsequiousness'
path = './Activate/' + personal

dataset = make_dataset(Personality = personal)
model.get_and_save_activations(dataset=datasets,
  save_path=path)
```

The ACA Toolkit is a fast, accurate, and general method for generating behavioral description datasets and obtaining control activation for a target personality. It enables rapid generation of control activations for specific personalities within one minute. Therefore, it significantly enhances the generalization ability of ControlLM, thereby meeting the broader needs of personality control.

C AI writing statement

- This paper utilized AI assistance for language polishing of the manuscript, including vocabulary correction and spell checking.
- During the research and writing of this paper, AI was used to retrieve relevant papers.
- Part of the code for generating figures in this paper was produced by AI.

The AI mentioned in the above content includes Claude [1], Gemini [2], and GPT-4 [3]. For all content generated by AI, we pledge thorough examination and take responsibility for veracity.

D Reproducibility Statement

We conducted all the experiments mentioned in the main text using the Llama-2-Chat 7B, Llama-2-Chat 70B, and Falcon 7B models downloaded from huggingface[4] (Wolf et al., 2019) with Pytorch framework (Paszke et al., 2019). Considering that sampling decoding can cause random bias in the text generated by the language model, and this bias may affect specific behaviors such as personality, all experiments use greedy decoding with temperature=0.

[1] https://claude.ai/
[2] https://bard.google.com/
[3] https://chat.openai.com/
[4] https://huggingface.co