Eliciting Textual Descriptions from Representations of Continuous Prompts

Dana Ramati **Daniela Gottesman** **Mor Geva**

Blavatnik School of Computer Science, Tel Aviv University

{danaramati@mail, gottesman3@mail, morgeva@tauex}.tau.ac.il

Abstract

Continuous prompts, or "soft prompts", are a widely-adopted parameter-efficient tuning strategy for large language models, but are often less favorable due to their opaque nature. Prior attempts to interpret continuous prompts relied on projecting individual prompt tokens onto the vocabulary space. However, this approach is problematic as performant prompts can yield arbitrary or contradictory text, and it individually interprets each prompt token. In this work, we propose a new approach to interpret continuous prompts that elicits textual descriptions from their representations during model inference. Using a Patchscopes variant (Ghandeharioun et al., 2024) called `InSPEcT` over various tasks, we show our method often yields accurate task descriptions which become more faithful as task performance increases. Moreover, an elaborated version of `InSPEcT` reveals biased features in continuous prompts, whose presence correlates with biased model predictions. Providing an effective interpretability solution, `InSPEcT` can be leveraged to debug unwanted properties in continuous prompts and inform developers on ways to mitigate them.

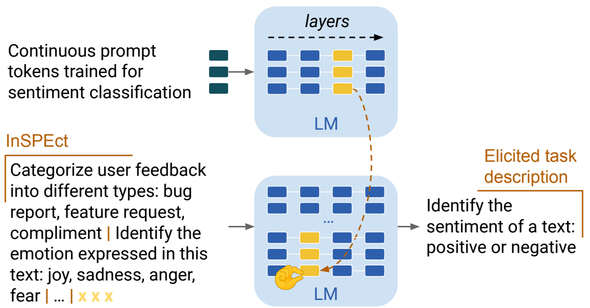

Figure 1: `InSPEcT` interprets a continuous prompt by patching the prompt representations (top) into an inference pass that generates a task description (bottom).

1 Introduction

Continuous prompts, or "soft prompts", are an efficient and widely-adopted solution for priming pretrained large language models (LLMs) to solve various tasks (Li and Liang, 2021; Lester et al., 2021). However, they are often less favorable compared to alternative parameter-efficient tuning methods, such as discrete prompt tuning, due to their opaque nature (Liu et al., 2023; Choi et al., 2024).

How should continuous prompts be interpreted? Prior work explored discretizing continuous prompts through projection to the model's vocabulary space (Khashabi et al., 2022; Ju et al., 2023). However, such approaches are problematic because it is possible to find performant continuous prompts that map to arbitrary or contradictory text (Khashabi et al., 2022). Moreover, they assume that each prompt token has an individual interpretable meaning, which does not necessarily hold.[1]

In this work, we introduce a new approach for interpreting continuous prompts that overcomes these limitations. We propose to elicit textual descriptions of the prompt from its representations, constructed by the model during inference. This is done by using the Patchscopes framework (Ghandeharioun et al., 2024); the prompt representations are extracted during inference and "patched" into a separate inference pass that steers the model to generate a textual description of the task (see illustration in Figure 1). Concretely, we define a task-description Patchscopes, called `InSPEcT` (Inspecting Soft Prompts by Eliciting Task descriptions), that relies on a few-shot target prompt for decoding task descriptions from patched continuous prompt tokens. Unlike vocabulary projections that produce a discrete replacement for the prompt, `InSPEcT` yields natural and easy-to-understand interpretations not bounded by its length.

We use `InSPEcT` to obtain descriptions of continuous prompts trained for 5 tasks, and find that it often yields accurate descriptions of the relevant target task (see examples in Table 1). Gener-

[1] For additional related work, please see §5.

Findings of the Association for Computational Linguistics: ACL 2025, pages 16545–16562
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ally, the higher the performance of a prompt, the more accurate the descriptions elicited from its representations. Next, we demonstrate the utility of the elicited descriptions for debugging continuous prompts. We show that a more detailed version of `InSPEcT` reveals biased features captured by prompts trained on the SNLI dataset (Bowman et al., 2015). Moreover, when these features are present in the elicited descriptions the model exhibits biased predictions.

In summary, our work introduces a novel and practical approach for interpreting continuous prompts by eliciting natural descriptions from their representations. We release our code at `https://github.com/danaramati1/InSPEcT`.

2 Eliciting Textual Descriptions of Continuous Prompts

We detail our approach for interpreting continuous prompts, which are learnable tensors concatenated with an input and optimized for a specific task on top of a frozen model (Lester et al., 2021). Let M be a pre-trained, auto-regressive transformer-based LLM (Vaswani et al., 2017) with L layers, and $\mathcal{P}_{cont} := \langle \mathbf{p}_1, ..., \mathbf{p}_n \rangle$ a continuous prompt optimized for classification task T.

Our method elicits comprehensible descriptions of continuous prompts from M's hidden representations, unlike prior work that maps these representations directly to discrete prompts. We build on the Patchscopes framework (Ghandeharioun et al., 2024), which decodes information from a source prompt by "patching" its hidden representations into the inference pass of a carefully designed target prompt. By conditioning M's generation on source representations through patching, these target prompts guide M to generate human-readable text reflecting the information encoded in them.

We introduce `InSPEcT`, a Patchscopes variant for deciphering continuous prompts, which are *learned disjointedly from the model's representation space*. Since these embeddings are optimized via an external loss function, they are not constrained to the distribution of standard language token representations. Moreover, although they lie within the model's hidden representation space — that is, their dimensionality matches that of the model's hidden states — they may reside in an opaque, compressed feature space designed to facilitate task performance. This is different from existing Patchscopes (e.g., Belrose et al., 2023; Pal et al., 2023), which interpret standard token representations.

	n	Example elicited descriptions
SST-2	14	Identify the sentiment of a text: positive or negative
	14	Categorize the tone of a text as positive, negative, or neutral
	7	Identify the author's intention in this text: positive, negative or neutral
Subj	56	"subjective opinion or objective fact?
	28	subjective, objective, or both?
	56	"subjective, objective, or neutral? It is a subjective, objective, or neutral text?
AGNews	28	Identify the topic of this article: technology, business, sports, world
	56	Sports? Technology? Business? World?
	28	World, Technology, Business, Sports, and Politics

Table 1: Accurate descriptions elicited from continuous prompts with n tokens using `InSPEcT` for SST-2 (Socher et al., 2013), Subj (Pang and Lee, 2004), and AGNews (Zhang et al., 2015) on LLaMA2-7B-Chat.

`InSPEcT` We treat $\mathcal{P}_{cont}$ as the source prompt we interpret, and design a few-shot task-description target prompt $\mathcal{P}_{target}$:

$$\text{"}\mathtt{desc}^{(1)}:\mathtt{class}_1^{(1)},\ldots,\mathtt{class}_{m_1}^{(1)} \mid \ldots \mid \mathtt{desc}^{(k)}:\mathtt{class}_1^{(k)},\ldots,\mathtt{class}_{m_k}^{(k)} \mid \mathtt{x\ x\ldots x}\text{"}$$

where $\mathtt{desc}^{(i)}$ is a textual description of some task $T_i \neq T$, $\mathtt{class}_j^{(i)}$ is the j-th class label of T_i, m_i is the number of classes in T_i, and k is the number of demonstrations. The list of demonstrations is followed by a sequence of placeholder patching tokens (the x's) of the same length as $\mathcal{P}_{cont}$. §A.1 lists examples of target prompts.

Denote by $\mathbf{p}_i^\ell$ the hidden representation of $\mathbf{p}_i$ at layer ℓ when running M on $\mathcal{P}_{cont}$. Similarly, let $\mathbf{x}_i^\ell$ be the hidden representation of the i-th placeholder token in the inference pass of M on $\mathcal{P}_{target}$. To elicit a textual description of $\mathcal{P}_{cont}$, we patch the representations $\mathbf{p}_1^\ell \ldots \mathbf{p}_n^\ell$ at some layer ℓ into the corresponding placeholder token representations $\mathbf{x}_1^{\ell*} \ldots \mathbf{x}_n^{\ell*}$ at some layer ℓ^* and let M generate a sequence of tokens. If M processes P_{cont} as a task description, we expect it will follow the structure of P_{target} and decode P_{cont} into a human-readable description and set of classes for T.

3 Experiments

We study the relationship between the interpretability and performance of continuous prompts, showing that prompts become interpretable as their performance increases.

3.1 Experimental setting

Tasks and models We follow Khashabi et al. (2022) and base our analysis on 5 diverse classification tasks of: SST-2 (Socher et al., 2013) and SST-5 (Socher et al., 2013) for sentiment analysis, AG-News (Zhang et al., 2015) for news classification, Subj (Pang and Lee, 2004) for text subjectivity, and TREC (Voorhees and Tice, 2000) for answer type classification. As we observed low prompt accuracy and interpretability for TREC, consistently with previous work (Min et al., 2022a; Khashabi et al., 2022; Ju et al., 2023), we omit it from the results. For additional details about the tasks, see §B. We conduct our experiments on LLaMA2-7B-Chat (Touvron et al., 2023) with 32 layers, and LLaMA3-8B-Instruct and LLaMA-3.1-70B-Instruct (Dubey et al., 2024) with 32 and 80 layers, respectively.

Prompt training For each task, we train 12 continuous prompts using standard prompt tuning (Lester et al., 2021): for every prompt length $n \in \{7, 14, 28, 56\}$, we train 3 prompts with different random initializations. During training, we save intermediate check-points of the trainable parameters every 1K-6K examples (depending on the task and dataset size), so we can analyze the progression of task accuracy and description interpretability. For more details, see §B.2.

`InSPEcT` demonstrations In order to steer the format of the elicited descriptions using `InSPEcT`, we crafted a set of 8 descriptions of classifications tasks with varying numbers of classes, that are not featured in our evaluations. Given the sensitivity of LLMs to prompt variations (Min et al., 2022b; Mizrahi et al., 2024), we interpret each continuous prompt using three target prompts with different demonstrations sampled from these task descriptions. The set of descriptions and example target prompts we used are listed in §A.

To ensure that the demonstrations influence only the format of the generated descriptions — rather than their semantic content – we also tested both an empty target prompt and a nonsensical one. In both cases, the model still produced descriptions that accurately reflected relevant class labels, however their formats varied significantly, making automatic evaluation of their quality using ROUGE-1 challenging. As such, we intentionally prompt the model with few-shot target prompts that encourage a consistent output format. Full details of this experiment are provided in §A.3.

Evaluation To assess the quality of a description D, we compute two metrics:

- **Class Rate**: The portion of class labels present in D. For example, in binary sentiment classification over the SST-2 dataset, if the label `positive` is present and the label `negative` is omitted in D, then the class rate is 0.5.
- **ROUGE-1**: The maximal ROUGE-1 score (Lin, 2004) of D against a set of 8-10 reference task descriptions, denoted by $\mathcal{D}_T$. Scores were computed after removing stopwords from both D and the reference. To construct $\mathcal{D}_T$, we manually wrote a textual description of T and then generated several paraphrased versions using ChatGPT (OpenAI, 2023). In §C we provide the references and more details, and justify this metric by showing that it correlates with user judgment.

The *interpretability* of a continuous prompt is measured by the average Class Rate and ROUGE-1 scores over the descriptions elicited from three target prompts.[2] The prompt *performance* is measured by the task accuracy of the model when the continuous prompt is prepended to the input example (explained in §B.2). We evaluate on the SST-2 and SST-5 validation sets and the AGNews and Subj test sets, as validation sets are not available.

3.2 Results

First, we observe that `InSPEcT` is able to elicit accurate task descriptions, reaching ROUGE-1 $=$ 0.8-0.9 and covering all the task class labels (Class Rate $= 1.0$). Examples are in Table 1 and §F.

Next, Figure 2 shows that the interpretability of a prompt increases with its task accuracy. Since elicited descriptions can be viewed as the model's interpretation of the continuous prompts, more effective continuous prompts yield more understandable and suitable descriptions. Moreover, interpretability improves as continuous prompts lengthen. We hypothesize that this trend arises because longer prompts allow the model to compress fewer task features per token (Elhage et al., 2022).

4 Debugging Continuous Prompts

We demonstrate the utility of `InSPEcT` for debugging continuous prompts trained over the SNLI dataset (Bowman et al., 2015). Another analysis addressing the low task accuracy on SST-5 is included in §G.

[2] We discuss the faithfulness of elicited descriptions in §D.

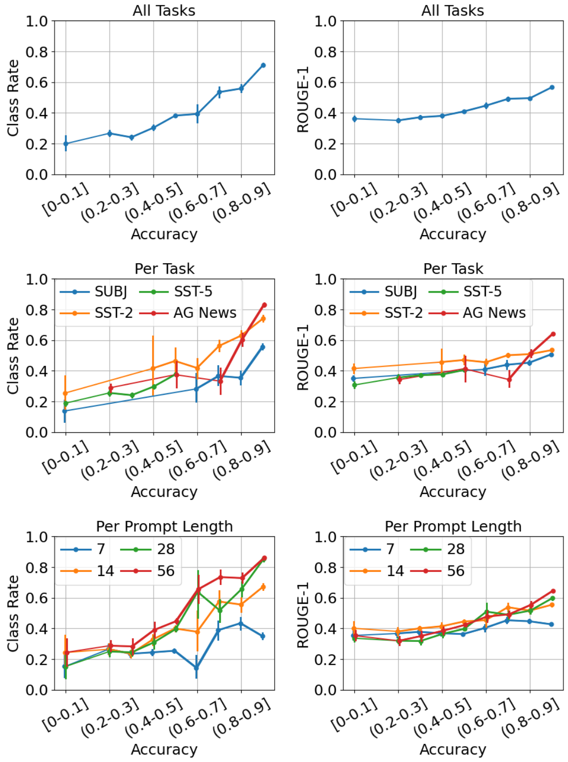

Figure 2: Prompt interpretability as a function of task accuracy for LLaMA2. The Class Rate/ROUGE-1 scores are averaged over all the prompts within the accuracy bin. For each task and token length, the scores increase with the performance of the prompt. Results for LLaMA3 show similar trends (see §E).

Eliciting spurious correlations using `InSPEcT` The SNLI dataset is known to have multiple biases (Gururangan et al., 2018; Mersinias and Valvis, 2022) which allow models to learn shortcuts, such as the correlation of negation and vagueness with certain classes. We use SNLI to train 10 different 14-token continuous prompts, check-pointed over 8 epochs, which vary in random initialization. `InSPEcT` is applied to each check-point using a target prompt that elicits the learned features:

"`Respond with a short sentence. What features are used for classifying each label in the following task: x x...x`"

Next, we count the appearances of distinct word groups in the generated outputs: (a) *biased words*: words with top-5 highest spurious correlations per class, as reported in Wu et al. (2022) Table 12, (b) *common words* (baseline): top-10 most frequent words across all generated outputs, omitting stop-words, digits and words in the target prompt, and (c) *random words* (baseline): 10 words randomly sampled from all generated outputs. For each generated output and group, we measure the word count difference with respect to the output of a randomly initialized prompt (epoch 0).

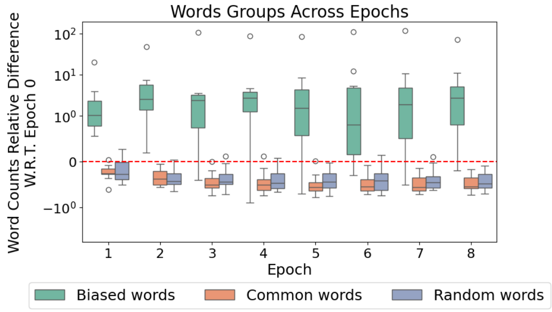

Figure 3: Differences in counts of each word group in generated outputs during training with respect to randomly-initialized prompts (epoch 0). The distributions are aggregated over 10 continuous prompts trained on SNLI (Bowman et al., 2015).

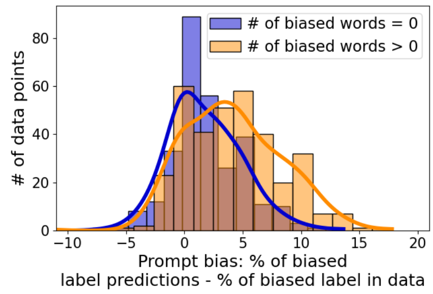

Figure 4: Histograms of the counts of generated biased words across different prompt bias levels. Outputs with biased words (> 0) show positive predictive bias, while those without ($= 0$) are unbiased on average. The x-axis is cut to $[-10, 20]$ for brevity, omitting outliers.

Figure 3 shows that biased words emerge early in training, reflecting the existence of these features in SNLI continuous prompts. This contrasts the decreasing presence of baseline word groups.

Elicited biases correlate with biased predictions If continuous prompts indeed capture the biases elicited from `InSPEcT`, then we expect them to encourage biased model predictions. To test this, we take each continuous prompt check-point and biased word pair, and quantify the model's predictive bias towards the bias-correlated class. Bias is measured by calculating the percentage difference between predicted and actual cases of a bias-correlated class among dataset examples containing the biased word, with larger differences indicating higher predictive bias. For example, considering the biased word `outside` and bias-correlated class `entailment` — if 65% of the relevant examples are true entailment cases and 70% are predicted as such, the bias measure is 5%.

Figure 4 shows that predictive bias is generally positive for outputs containing biased words, and centered around 0 for outputs lacking them. A sign test comparing these distributions indicates significantly higher predictive bias when a biased word is present (p-value $2.96e^{-11}$).

5 Related Work

Interpreting continuous prompts Interpreting continuous prompts has been attempted by projecting individual prompt tokens to the vocabulary space (Khashabi et al., 2022; Webson and Pavlick, 2022) or by optimizing an external objective to map them to their discrete forms (Ju et al., 2023). However, these mappings operate on each token individually, often contain several noisy tokens that are difficult to understand (Ju et al., 2023), and may yield discrete interpretations that are irrelevant or contradictory (Khashabi et al., 2022).

Embedding inversion Previous research has investigated reconstructing text from dense representations by learning a function that inverts the text encoder (Morris et al., 2023). Other approaches identify which content activates certain model components in order to decipher the information encoded in new inputs (Huang et al., 2024). These methods involve extensive analysis and rely on external optimizations. In contrast, our approach simply leverages the model's intrinsic generation capabilities to form understandable descriptions of continuous prompt embeddings.

Bias in continuous prompts Models may rely on spurious correlations between classes and specific words (Wu et al., 2022), and superficial clues (Kavumba et al., 2022), like high lexicographical overlap between the premise and hypothesis in natural language inference, to perform various classification tasks. To mitigate this, various dataset augmentation schemes have been developed (Zhao et al., 2018). Our work uncovers biased features in continuous prompts which can inform when it is appropriate to employ such tactics.

6 Conclusion

We tackle one of the major hurdles of continuous prompts — their lack of transparency. We show that accurate task descriptions can be elicited with `InSPEcT` from the model's internal representations, and task performance improves as the model's own interpretation of the prompts becomes more faithful. Additionally, `InSPEcT` can identify biased features in continuous prompts from the presence of prominent words in the generated outputs. Overall, our work provides an effective interpretability solution that can be leveraged to debug unwanted properties in continuous prompts.

Limitations

Following previous work on interpretability of continuous prompts (Khashabi et al., 2022; Ju et al., 2023), our experiments focus on classification tasks where evaluation is easier compared to open-ended generation tasks. Extending our analyses to other tasks is an interesting direction for future work.

We were often able to elicit meaningful and understandable task descriptions, though there were some the cases where the descriptions were unclear and did not yield informative content, especially early in training. Since `InSPEcT` can be viewed as the model's interpretation of the continuous prompts, identifying the precise conditions for its success may align with optimizing training configurations that enable the model to learn more effectively.

Our work explores the correlation between prompt interpretability and task performance by finding a meaningful one-way mapping from continuous prompts to discrete forms. Conducting a causal analysis — where elicited descriptions are modified, mapped back to continuous prompts, and evaluated for changes in task performance — could offer deeper insights into how models use and form predictions based on the information encoded within continuous prompts.

Prior work focused on discretizing continuous prompts such that the discrete prompts can be used as replacements that yield equivalent task performance and class label prediction distributions. Notably, our method does not produce such discrete replacements, but rather elicits information in a textual and easy-to-understand format to better understand the information encoded in the continuous prompt and its potential for debugging. While we found the elicited descriptions to be generally informative and accurate, they do not necessarily guide the model to produce explicit class labels like their corresponding continuous prompts.

7 Broader Impact

`InSPEcT` addresses the opacity of continuous prompts by extracting meaningful, human-readable insights from them and uncovering embedded biases. By enhancing our understanding of model behavior and how models process information, this approach supports the creation of safer, more accountable technologies across a wide range of applications.

Acknowledgments

We thank Ori Yoran, Asma Ghandeharioun, and Katja Filippova for constructive feedback. This research was supported in part by The Israel Science Foundation grant 1083/24, Len Blavatnik and the Blavatnik Family foundation.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yunseon Choi, Sangmin Bae, Seonghyun Ban, Minchan Jeong, Chuheng Zhang, Lei Song, Li Zhao, Jiang Bian, and Kee-Eung Kim. 2024. Hard prompts made interpretable: Sparse entropy regularization for prompt tuning with RL. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8252–8271, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Google-Research. 2020. Rouge: A python implementation of rouge-1.5.5. `https://github.com/google-research/google-research/tree/master/rouge`.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Xinting Huang, Madhur Panwar, Navin Goyal, and Michael Hahn. 2024. Inversionview: A general-purpose method for reading information from neural activations. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Tianjie Ju, Yubin Zheng, Hanyi Wang, Haodong Zhao, and Gongshen Liu. 2023. Is continuous prompt a combination of discrete prompts? towards a novel view for interpreting continuous prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7804–7819, Toronto, Canada. Association for Computational Linguistics.

Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352, Dublin, Ireland. Association for Computational Linguistics.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. 2022. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3631–3643, Seattle, United States. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Michail Mersinias and Panagiotis Valvis. 2022. Mitigating dataset artifacts in natural language inference through automatic contextual data augmentation and learning optimization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 427–435, Marseille, France. European Language Resources Association.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt: Language model. `https://openai.com/chatgpt`.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

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 32*, volume 32, pages 8024–8035.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. *ArXiv*, abs/2203.12942.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Target Prompts

Examples of task descriptions and target prompts are presented in this section. Discussions regarding their use and generation are in §2 and §3.1, respectively.

A.1 Example Target Prompts

The following target prompts were used by `InSPEcT` to elicit task descriptions:

- Categorize customer feedback into different types: bug report, feature request, compliment | Identify the emotion expressed in this text: joy, sadness, anger, fear | Is the information in this sentence correct?: True, False | x x x x x x x
- Determine who is the author of a given text: Shakespeare or Marlowe | Categorize customer feedback into different types: bug report, feature request, compliment | Identify the political leaning of a text or author: left or right | x x x x x x x x x x x x x x
- Identify the emotion expressed in this text: joy, sadness, anger, fear | Classify this passage from a book or movie into its genre: science fiction, romance, thriller | Identify which season is decsribed in this text: summer, winter, autumn or spring | x x x x x x x x x x x x x x x x x x x x x x x x x x x x

A.2 Crafted Classification Tasks Descriptions

The following task descriptions were used for sampling and constructing target prompts:

- Identify the emotion expressed in this text: joy, sadness, anger, fear
- Is the information in this sentence correct?: True, False
- Classify this passage from a book or movie into its genre: science fiction, romance, thriller
- Determine who is the author of a given text: Shakespeare or Marlowe
- Identify which season is described in this text: summer, winter, autumn or spring
- Categorize customer feedback into different types: bug report, feature request, compliment
- Identify the type of this email: spam or not spam
- Identify the political leaning of a text or author: left or right

A.3 Validating the Role of Target Prompts

To assess the influence of the target prompt, we experimented with two additional prompts that are irrelevant to the task: an empty target prompt and the nonsensical string "`Lorem ipsum dolor sit amet`". Interestingly, even an empty target prompt can elicit descriptions that accurately capture relevant class labels and, in the case of SST-2, correctly identify the task as sentiment classification for movie reviews. The irrelevant target prompt prompted descriptions captured relevant class for AGNews, and a loose interpretation of the task for SST-2 but the descriptions are convoluted with erroneous formatting and additional text.

These descriptions also tend to vary in format, making automatic evaluation (e.g., using ROUGE-1 against reference descriptions) difficult. To enable standardized measurement, we instead prompt the model with few-shot target prompts that encourage a consistent output format.

For these baselines, we experimented with continuous prompts of length 14 tokens with 95% accuracy on AGNews and of length 28 tokens with 93% accuracy on SST-2, trained with LLaMA-2-7B-Chat. Examples of baseline descriptions and associated patching configurations are found in Table 2.

B Additional Experimental Details

B.1 Downstream Tasks

Dataset	Task	$\|C\|$
AGNews	News topic classification	4
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
Subj	Subjectivity classification	2
TREC	Answer type classification	6

Table 3: The set of downstream tasks used in the experiments, where $|C|$ represents the number of classes for each task.

B.2 Training Details

Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where x_i is an input text for classification and y_i is a gold class label, $\mathcal{P}_{cont}$ is learned by minimizing the cross-entropy loss between y_i and the model's predicted label for the input "$\mathbf{p}_1 \dots \mathbf{p}_n$ `Text:`$[x_i]$`, Label:`" over $\mathcal{D}$.

Dataset	Learning Rate	Epochs	$\|T\|$
AGNews	$8e^{-3}$	8	50,000
SST-2	$8e^{-4}$	8	50,000
SST-5	$6e^{-3}$	12	8,500
Subj	$8e^{-3}$	8	8,000
TREC	$8e^{-4}$	20	5,400

Table 4: Hyper-parameters used to train prompts on both LLaMA2 7B chat and LLaMA3 8B Instruct models. $|T|$ represents the size of the training set used.

B.3 Resources

All our experiments were conducted using a single A100 80GB or H100 80GB GPU.

B.4 Software Packages

We used the PyTorch Python package (Paszke et al., 2019) for training the continuous prompts and conducting the experiments. For calculating the scores, we used the rouge-score Python package (Google-Research, 2020) for ROUGE-1, and the NLTK Python package (Bird and Loper, 2004) for removing English stopwords, both with default parameters.

C ROUGE-1 Calculation and Justification

Further details regarding the computation of the ROUGE-1 scores are discussed below.

C.1 Stopwords Removal

To prevent computing misleadingly high ROUGE-1 scores for discrete prompts that closely resemble the format of reference descriptions, but fail to accurately capture the target task, we removed stopwords from both the elicted `InSPEcT` descriptions and the reference descriptions in Table 5. This was achieved using the NLTK Python package (Bird and Loper, 2004).

C.2 Reference Descriptions Creation

To compute the final ROUGE-1 score for each description D elicted by `InSPEcT`, we used ChatGPT to generate 8-10 reference descriptions per task. The input format we used to prompt ChatGPT was: "`Could you rephrase the following sentence and provide a few options: <SENTENCE>`", where `<SENTENCE>` represents a brief description of the target task. Examples of reference descriptions generated are presented in Table 5.

C.3 Human Evaluation

We conducted a user study to assess the alignment between the ROUGE-1 metric and human judgments of interpretability. A total of 218 elicited descriptions were uniformly sampled across binned ROUGE-1 scores. Four graduate students were then tasked with rating the accuracy of each description on a scale from 1 to 4, based on its similarity to the reference task descriptions used for ROUGE-1 computation. The analysis revealed a

Task	Target Prompt	Source Layer	Target Layer	Elicited Description
AGNews	Empty	6	13	World! Business! Technology! Entertainment! Sports! Politics! Science! Health! Fashion! Travel! Home! Gadgets! Gaming! Sports! Politics! Science! Health! Fashion! Travel! Home! Gadgets! Gaming!
		13	8	position: Business and Technology World Business and Technology Technology and entrepreneurship World blanket._
SST-2	Empty	15	12	anse and negative experiences. The purpose of this study is to investigate the impact of positive and negative experiences on the mental health of individuals.
		9	12	ight positives and negatives. The positive impact of social media on society
		22	8	ance and happiness. As for the content of the review, the reviewer seems to be very positive about the movie. They use words like "excellent," "outstanding," and "superb" to describe it, and they mention that the movie is "a must-see." They also provide some specific examples from the movie to support their positive assessment.
AGNews	Lorem ipsum	3	5	on end: Business [object Object] Technology [object Object] Gaming [object Object] Sports [object Object] [object Object] Entertainment [object Object] Music [object Object] Movies [object Object] [object Object] Politics [object Object] World [object Object] News [object Object] [object Object] Culture [object Object] Art [object Object] Literature [object Object] [object Object] Travel [object Object] Adventure [object Object] Food [object Object] [object Object] Fashion [object Object] Beauty
		8	6	Oops, I apologize for that. Business as usual. Technology and science. Business and finance. Sports and leisure. Entertainment and culture. Politics and government. World and economy.
SST-2	Lorem ipsum	7	6	article and more ' + The movie received positive reviews from critics, with many praising the performances of the cast and the gritty, realistic portrayal of life inside a maximum security prison.
		16	5	cyt and negative sentiment? The answer is no,

Table 2: Elicited descriptions from `InSPEcT` using different target prompts, source layers, and target layers for AGNews and SST-2.

	Example reference descriptions
AGNews	Which topic is this article about? World, Sports, Business, Technology
	What is the main topic discussed in this article: World, Sports, Business, Technology
	What is the most fitting summary for this article? World, Sports, Business, Technology
	Among World, Sports, Business, and Technology, which best captures the topic of this article
	To which category does this news article's topic belong: World, Sports, Business, Technology
SST-2	Is the sentiment of this sentence positive or negative?
	Would you classify this sentence as having a positive or negative sentiment?
	Can you identify whether the sentiment of this sentence is positive or negative?
	Would you consider the sentiment of this sentence to be positive or negative?
	What is the tone of this sentence: positive or negative?
SST-5	Is the sentiment of this sentence terrible, bad, neutral, good or great?
	Do you think this sentence has a terrible, bad, neutral, good or great tone?
	How would you rate the sentiment of this sentence: terrible, bad, neutral, good or great?
	How would the sentiment of this sentence be described? terrible, bad, neutral, good, great.
	Would you classify this sentence as having a terrible, bad, neutral, good or great sentiment?
Subj	Is the subjectivity of this text objective or subjective?
	In terms of subjectivity, is this sentence objective or subjective?
	Classify the sentence based on its expression: objective, subjective
	Is this sentence objective or subjective in nature?
	Determine if this sentence presents facts or opinions: objective, subjective
TREC	Is the question asking about an entity, a description, an abbreviation, an expression, a human, a location, or a number?
	What type of thing is the question asking about? Description, entity, abbreviation, expression, human, location, number
	What type is the answer to this question: entity, description, abbreviation, expression, human, location, or number?
	Choose the category that best fits the answer: Description, Entity, Abbreviation, Expression, Human, Location, Number
	Does the question pertain to an entity, a description, an abbreviation, an expression, a human, a location, or a number?

Table 5: Example of reference descriptions used to calculate ROUGE-1 scores.

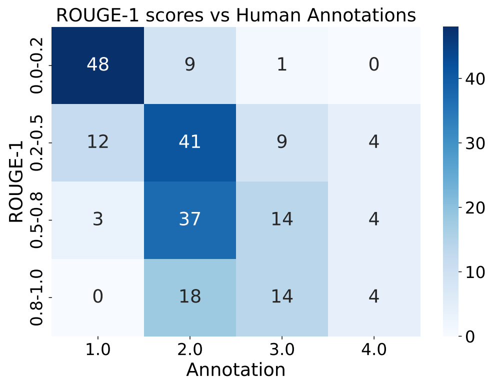

Figure 5: Heatmap comparing binned ROUGE-1 scores and human annotations of the accuracy of elicted task descriptions.

moderately-strong Spearman correlation of 0.66 (p-value $1.3e^{-28}$) between ROUGE-1 scores and human judgments, underscoring the metric's effectiveness in automatically evaluating interpretability. As shown in Figure 5, ROUGE-1 scores are generally faithful to human annotations. The detailed instructions provided to annotators are presented in Figure 7.

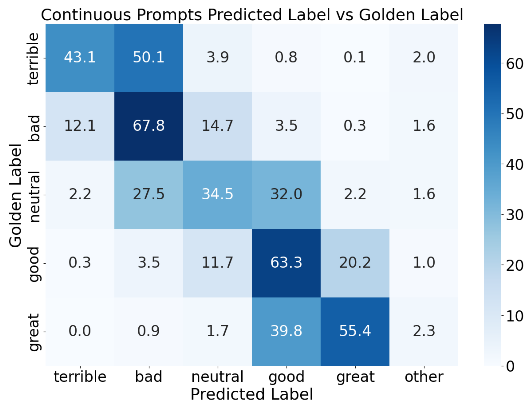

Figure 6: Confusion matrix comparing predictions generated by continuous prompts which captured only three classes, and the true labels.

INSPECT Descriptions Human Evaluation

Please make sure to read the instructions and go over the example annotations before completing your assignments.

Instructions

The goal of this study is to evaluate the accuracy of the task descriptions elicited from continuous prompts by INSPECT.

You will be given a task description produced by INSPECT and a set of reference descriptions of the task written by humans. Your task is to rate the completeness and accuracy of the description with respect to a set of references. Specifically, you will rate each description on a scale of 1-4 while considering the following aspects:

- How well does the description capture the essence of the task?
  How well does the description capture the objective of the task? Namely, does the description include all the class labels of the task?

Use the following ranking scores:

1 **The description is <u>not accurate</u>**: it is off-topic completely or the text is completely incoherent with no mention of any class labels.

2 **The description is <u>partially accurate</u>**: it doesn't fully capture the essence of the task, and there may be inaccuracies in the listed class labels

3 **The description is <u>mostly accurate</u>**: it is mostly faithful to the task, or it doesn't describe the task but lists all the correct class labels.

4 **The description is <u>accurate</u>**: it is faithful to the task and lists the complete set of class labels.

Figure 7: User study instructions (1/2)

Example Annotations

1 — Not Accurate

Task	Description	Rating Explanation
SST-2	, in this text, there is a description of a season: summer, winter, autumn or spring	**Irrelevant description** and **no classes.**
SST-5	The passage is from the movie "Her" (2013) directed by Spike Jonze	**Irrelevant description** and **no classes.**
SUBJ	Identify the political leaning of a text or author: left or right	**Irrelevant description** and **irrelevant classes.**
AGNews	Identify the type of this email: spam or not spam	**Irrelevant description** and **irrelevant classes.**

2 — Partially Accurate

Task	Description	Rating Explanation
SST-2	, or negative	**Poor description**, captures **partial classes.**
SST-5	0) was good but not great	**Poor description**, captures **partial classes.**
SUBJ	subjective opinions, the following are the opinions of the writer:	**Poor description**, captures **partial classes.**
SUBJ	of the subjective, objective, or neutral?	**Poor description**, captures all correct classes and **some incorrect class.**
AGNews	-- World-- Sports-- Business-- Technology-- Entertainment	**No description**, captures **all correct classes** but also **some incorrect class.**
AGNews	? Business? Technology? . Technology? Sports? Fashion? Entertainment? Business? World? Business? Sports? Technology? Fashion? Entertainment?	**No description**, captures **all correct classes** but also includes **incorrect classes.**

3 — Mostly Accurate

Task	Description	Rating Explanation
SST-2	Identify the tone of a text: positive, negative, or neutral	**Captures essence of task**, captures all correct classes but also includes an irrelevant class.
SST-2	Identify the tone of the email: positive, negative, or neutral	**Captures essence of task**, captures all correct classes but also includes an irrelevant class.
SUBJ	of the subjective and subjective nature of the interpretation of the text	**Captures essence of task**, captures majority of classes.
SST-5	Determine the sentiment of this text: this is amazing, terrible, or neutral	**Captures essence of task**, captures majority classes.
AGNews	Identify the topic of this article: technology, business, politics, entertainment	**Captures essence of task**, captures majority classes.
AGNews	World Sports Business Technology	**No description**, but lists **all correct classes.**

4 — Accurate

Task	Description	Rating Explanation
SST-2	, and identify whether it is a positive or negative sentiment	**Captures essence of task** and **correct classes.**
SUBJ	nature, the subjective and objective of the text	**Captures essence of task** and **correct classes.**

Figure 7: User study annotation examples (2/2)

D Faithfulness of Elicited Descriptions

A key challenge in interpreting continuous prompts is ensuring the faithfulness of the generated textual descriptions. Unlike previous approaches that seek discrete replacements for continuous prompts, our method focuses on interpretation rather than exact replication. While the model's outputs are causally influenced by the continuous prompt, due to inherent randomness in model generation, no single description can be considered fully faithful. Therefore, for a given continuous prompt, we aggregate outputs across different target prompts to help mitigate this variability. This allows us to uncover consistent patterns embedded in the continuous prompt and capture more robust and meaningful signals. For example, the bias analysis in §4 demonstrates that aggregating multiple descriptions of the same task reveals strong evidence of the model basing its predictions on spurious correlations in the data.

Although our analyses reveal a clear trend — task accuracy improves as elicited descriptions become more accurate — less accurate descriptions are occasionally observed, which can be attributed to several factors.

- **Sampling noise** The use of temperature-based sampling introduces variability, occasionally generating less probable tokens that steer outputs towards less accurate descriptions.

- **Complexity of learned features** Continuous prompts encode abstract task-relevant features, making eliciting coherent descriptions challenging. Nonetheless, even less coherent descriptions often include correct class labels, as encouraged by the target prompts which gives a useful signal for what classes the model learned.

- **Prompt length** Short continuous prompts (e.g., 7 tokens) compress task features into fewer dimensions, complicating the generation of comprehensive descriptions. Examples in Tables 7 and 8 illustrate this effect.

While this is not the intended use-case of `InSPEcT`, we did experiment with prompting the model with accurate elicited task descriptions (discrete prompts), and observed that while some of the discrete prompts yielded high task performance – this was not always the case. Exemplars are found in Table 6.

E Additional Results

The results for LLaMA3-8B-Instruct and LLaMA-3.1-70B-Instruct are presented in Figure 8 and Figure 9, accordingly. We observe similar trends to those of LLaMA2-7B-Chat. First, we observe that the interpretability of a prompt improves as its task accuracy increases. However, there is a small drop in interpretability within the 0.8 to 1 accuracy range, likely due to the trends observed across all tasks when using 7 tokens, affecting both class rate and ROUGE-1 scores. Additionally, interpretability improves as continuous prompts lengthen, as obsereved in LLaMA2-7B-Chat.

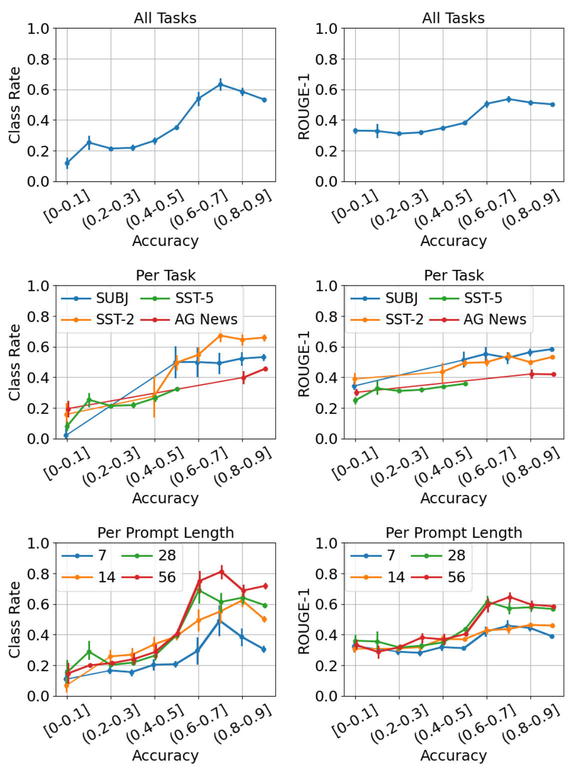

Figure 8: Prompt interpretability as a function of task accuracy for LLaMA3-8B-Instruct. The Class Rate/ROUGE-1 scores are averaged over all the prompts within the accuracy bin.

Task	Discrete Prompt	Prompt Accuracy	
		Continuous	Discrete
AGNews	World, Technology, Sports, Business, Entertainment	0.95	0.084
SST-2	Identify the sentiment of a text: positive, negative, or neutral	0.90	0.75
Subj	The term "subjective" refers to something that is based on personal opinions or feelings, rather than objective facts	0.96	0.62

Table 6: Comparison of task performance using continuous versus discrete prompts.

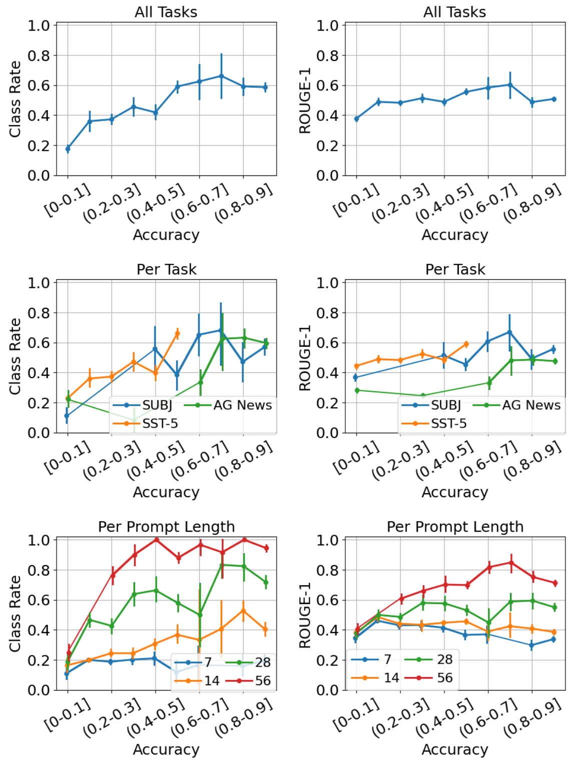

Figure 9: Prompt interpretability as a function of task accuracy for LLaMA-3.1-70B-Instruct. The Class Rate/ROUGE-1 scores are averaged over all the prompts within the accuracy bin.

F Example Interpretations of Continuous Prompts

Examples of discrete prompts elicited using `InSPEcT` on LLaMA2-7B-Chat and LLaMA3-8B-Instruct are presented in Table 7 and Table 8, respectively.

G Debugging Low Task Performance

In the SST-5 dataset, the trained continuous prompts achieved $50\% - 60\%$ task accuracy. Tables 7 and 8 contain examples of elicited `InSPEcT` descriptions, which often list only a subset of class labels: "`good`", "`bad`", "`neutral`". A possible explanation for the poor performance is that the continuous prompt steers the model to produce only a partial set of classes. Figure 6 presents a confusion matrix, with values representing dataset example counts, between the predictions generated by continuous prompts where the elicited descriptions captured only three classes, and the true labels. These prompts struggled to capture the nuanced differences between "`good`" and "`great`", shown by the similar prediction rates 39.8% and 55.4% for examples from the "`great`" class. Similar confusion is demonstrated for examples from the "`terrible`" class, where prediction rates are 43.1% and 50.1% for "`terrible`" and "`bad`", respectively. The omission of the difficult classes in the `InSPEcT` descriptions could indicate that the continuous prompts may not recognize the full spectrum of sentiment represented in SST-5.

AG News	number, as well as the latest news and updates from the world of technology
	with, Digital Marketing, Business, and Technology topics
	Identify the main topic of this text: technology, entertainment, politics, sports
	Identify the main theme of the text: technology, business, politics
	Club, or Identify the topic of this text: entertainment, politics, sports, or technology
	? technology &? business &? entertainment &? sports &? World &? news &? lifestyle &
	world, Technology, Business, and Sports
	-world, the following categories: Sports, Business, Technology, Entertainment, and Science
	– World– Technology– Business– Sports– World
	Sports? Technology? Business? World news? We will be happy to help you with any question you have!
	Technology World Business Sports
	is, World, Sports, Business, Technology
SST-2	xtake a look at the text and identify the tone: positive, negative, or neutral
	give feedback on a product: positive, negative, or neutral
	Identify the sentiment of a text: positive, negative, or neutral
	Categorize the tone of a text as positive, negative, or neutral
	ance as a positive or negative response?
	and negative sentiment?
	Please note that the text is a positive or negative?
	U (positive) and U (negative) are used to indicate the emotions expressed in the text
	? a positive or negative review?00000000000000000000000
	ES of negativity, but positivity?
	Identify the tone of a piece of writing: positive, negative, neutral
	rices and negative feelings, but also positive feelings, such as joy, happiness, and contentment
SST-5	leaving feedback on a product or service: good, bad, or neutral
	yeah (yes) (great job) (excellent) (good work) (well done) (superb) (amazing)
	(A) great (B) good (C) okay (D) poor
	yeah (100%), great (80%), okay (60%), meh (40%), bad (20%
	anarchy Is this a good or bad thing?
	yevaluate the quality of a piece of writing: good, neutral, or bad
	by: good, neutral or bad
	-ilk to which it is assigned: good, bad or neutral
	: This is a good or bad thing: Neutral
	by which I would classify it: good or bad
	bad? Very bad? Worse than bad? Terrible? Horrible? Abysmal?
	bad, my dear, this is a great answer
	nough, great, good, bad, or ugly?t is a genre of literature that explores the impact of science and technology on society
	testing is okay, but not great, but not terrible, but not good
	-based on their answers: good, neutral, or bad
	say goodness, the text is neutral
	bad and terrible) and 50/50 chance of being a good or bad review
	not enough this topic not enough to be considered as a good or bad reviews?
SUBJ	(Learning Objective 1)
	matter of fact, opinion or perspective
	coverage of a news article or event: objective, subjective, persuasive
	matter of factuality or subjectivity
	The above are examples of subjective and objective criteria for evaluating the quality of a text or author
	's the subjective and objective'
	subjective opinion of the matter
	subject to the subjective opinion of the observer
	The term "subjective" refers to something that is based on personal opinions or preferences, rather than objective facts
	The answer to this question depends on how you define "subjective" and "objective" are two different things
	subjective, objective, or both?
	in this passage, but the subjective and objective,
	the objective of this exercise is to assess the subjective value of the answer
	in a subjective, subjective, objective, or objective manner
	"objective""subjective""opinion""fact"

Table 7: Examples of accurate task descriptions elicited using `InSPEcT` on LLaMA2-7B-Chat.

AG News	World Technology Business nikaite Technology; or a Business; or Entertainment; or Sports; or Sports; or Technology; or Technology; 279; in Business; Sports; Entertainment; Technology;; Technology; Business; And World -including the World of Business or Leisure or Sports or Technology or News or Culture or Healthassistant; Technology; to classify this passage from: business Worldwide in a World of Business or Technology or Entertainment or Health or Fashion or Sports or World and answer: What is the main topic of this article?assistant Identify the type of the website: Technology, Entertainment, Sports, Business /oriented to the world of sports, the text is a sports news article ultiimateley, classify this text into a genre: business, technology, entertainmentassistant
SST-2	The text is a positive or negative: Identify the positive and negative statements in a text Identify the positive/negative emotions in a text: positive, negative Identify the positive or negative sentiment of a text Identify the positive and negative aspects of a text: The positive aspects of a text: The negative aspects of a text: Determine the sentiment of a text: positive, negative, or neutral lettered a positive or negative The text is a negative review of a movie, which is a negative review From a book: Identify the author's tone: positive, negative, formal, informal, sarcastic, or philosophical ://positive-negative-negative Is this a positive or negative review: positive, negative Is this sentence a positive or negative statement Categorize this text into a category: positive, negative, neutral
SST-5	badgered = 2;terrible = 2;good = 2;neutral = 2;bad = 2;terrible = 2;good = 2;neutral = 2;bad = 2;terrible = 2;good = 2; neutral of the good or bad of the game Identify the author of this text:terrible, good, neutral Answer: The text: a neutral good: a good'totalitarian a: a bad: aterrible:terrible:terrible:terrible onenasty of the text: neutral, good or bad ://good or bad text terrible, awful, bad, good, excellent, great, wonderful, lovely, beautiful, lovely, lovely :bad news, neutral, good news, neutral, bad news, good news, bad news :good or bad Is the information in this sentence good or bad? Is it a good news, bad news, or neutral news idiagnosis, a good or bad, and neutral I cannot be used, a good, neutral, or bad
SUBJ	Identify the tone of this text: formal, informal, formal and objective, formal and subjective Objective Subjective Subjective objective and subjective language: objective language is used to describe the facts, while subjective language is used to- express the author's opinion or feeling Objective of the learning objectives of the Subjective Subjective Objective: The text of the subjective The text is a subjective and/or objective and/or subjective/objective Objective: To identify the emotion expressed in the text Identify the subject of a text: objective, subjective Please note that the classification is subjective and may not be objective ://mannerisms of a text: Identify the tone of a text: objective, subjective, formal, informal, sarcastic ://determine the tone of the text: objective, objective, objective Subjective: The text is subjective as it is a subjective text

Table 8: Examples of accurate task descriptions elicited using `InSPEcT` on LLaMA3-8B-Instruct.