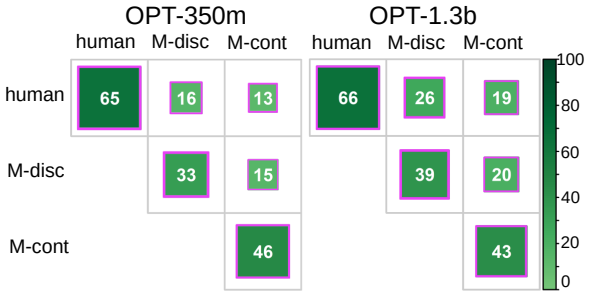

Figure 1: Unit activation overlap (0 to 100) between *human*, *M-disc* and *M-cont* prompt types for OPT-350m (left) and OPT-1.3b (right). Higher values (more intense color, larger squares) represent larger overlap. Confidence intervals (CIs) are shown as square outlines: thicker lines indicate wider CIs (CIs are generally small). Within-prompt overlap is higher than betweem-types overlap, suggesting a difference in processing.

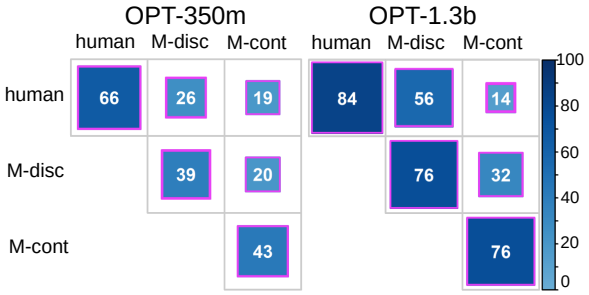

Figure 2: Percentage input similarity between *human*, *M-disc* and *M-cont* prompt types for OPT-350m (left) and OPT-1.3b (right). Higher values (more intense color, larger squares) cue high similarity. Within-prompt-type similarities (the scores on the diagonal) are generally higher than similarity between types. Note that the absolute values of the input similarity obtained with both model sizes are not directly comparable due to a difference in the input dimension (2048 *vs.* 1024).

Output agreement For each pair of same-task templates, we measure the proportion of test cases where the templates lead to the same prediction. We then compute prompt-type-level averages.

Uncertainty quantification We provide the uncertainty estimation of our measurements by computing the 95% Confidence Interval (CI) of each measure. In Table 1, the CI associated to each metric (accuracy, perplexity, attention distribution score and output entropy) is obtained by computing the 0.025 and 0.975 quantiles given the list of scores obtained with each templates of a given prompt type (note that each template's score is averaged at the level of the relation). 95%CI in Figure 1,2,3, and Table 2 are obtained using bootstrapping by randomly sampling with replacement from the list of templates (the number of resamples is found by incrementally increasing it until the uncertainty estimation converge).

4 Processing machine-generated prompts

We experimentally demonstrate that differences between human and machine-generated prompts exist at three different levels: (1) at the input level, when comparing prompt types in the embedding space, (2) at the output level, when analyzing predictions and output probabilities, and (3) at the level of intermediate activation, indicating a difference in processing at work in the LM. We conclude this quantitative analysis by showing that, although these metrics are correlated when compared within the same prompt type, the correlation is weak between prompts of different types, leading to a number of counterintuitive patterns in LM prompt processing.

4.1 Human and machine-generated prompts are processed differently.

High accuracy and high perplexity As confirmed in Table 1, the main motivation to use machine-generated prompts is their good performance, *M-cont* prompts outperforming *human* ones by +25pts. This higher accuracy comes along with lower output entropy, suggesting better LM calibration, where a larger mass of the output probability distribution is concentrated on the correct token.[3] However, prompt perplexity – quantifying the degree of predictability of a token sequence given an LM – is two order of magnitude higher for *M-disc* than for *human* templates.[4] We discuss this further in Section 4.2 below.

Low activation overlap of knowledge neurons Activation overlap statistics are provided in Figure 1. For both OPT-350m and OPT-1.3b, we observe that, while within-prompt-type overlap is mild or high, ranging from 33 to 66 (on a 0-to-100 scale), between-prompt-type overlap is always

[3]Calibration in LM analysis (e.g., Liang et al., 2023) refers to the confidence that a model has in its predictions when the latter are correct. Our output entropy measure does not directly correlate confidence and accuracy. However, as machine prompts are in general more likely to trigger the correct output and, at the same time, they have lower output entropy, the global trends do suggest that they tend to produce correct answers with more confidence. We informally use the term "calibration" to refer to this property.

[4]Due to their continuous nature, there is not trivial way to estimate perplexity for *M-cont* prompts.

	OPT-350m			OPT-1.3b		
	human	*M-disc*	*M-cont*	*human*	*M-disc*	*M-cont*
Accuracy	29.5	43.4	54.9	28.8	46.1	58.0
[95% CI]	[11.5, 65.0]	[17.0, 79.5]	[20.7, 86.0]	[10.4, 78.2]	[15.1, 83.4]	[23.8, 89.6]
Perplexity (10^3)	0.60	40.9	-	0.40	30.3	-
[95% CI]	[0.1, 1.9]	[16.0, 95.0]		[0.04, 1.48]	[2.0, 911.3]	
Attention distribution (%)	34.4	30.0	23.2	30.8	28.7	29.4
[95% CI]	[29.2, 39.7]	[27.3, 32.5]	[21.1, 25.5]	[17.0, 85.8]	[16.6, 84.4]	[14.3, 74.7]
Output entropy	5.00	4.30	2.10	4.70	3.90	2.10
[95% CI]	[3.2, 6.0]	[1.9, 5.7]	[0.5, 4.3]	[1.7, 6.0]	[1.3, 6.4]	[0.4, 5.9]

Table 1: Human and machine-generated prompts (both *M-disc* and *M-cont*) significantly differ in at least four aspects: (1) machine-generated prompts outperform the human ones in terms of accuracy; (2) they are also better calibrated on average, given their lower output entropy; while at the same time (3) machine-generated prompts are less predictable by the LMs, reaching significantly higher perplexity; (4) in machine-generated prompts, attention is concentrated on a smaller amount of tokens. For technical convenience, perplexity is not computed for *M-cont*.

low, ranging from 13 to 26. This pattern is more pronounced when comparing *human* and *M-cont*. Between-prompt overlap tends to be higher with OPT-1.3b, suggesting that larger LMs could show a convergence of human and machine-generated prompts (this remains to be further explored). The low-overlap result is confirmed by the diagnostic classifier analysis presented in App. B, that shows that a simple linear classifier can distinguish between any prompt type pair based on activation patterns on any layer of either LM.

Attention is focused on fewer tokens As transformer behaviour is a by-product of both FF and attention layers, we also look at the difference in attention distributions, shown in Table 1. Here again, we observe a clear distinction between human and machine-generated prompts, the latter leading to attention being focused on a smaller amount of tokens. Recall that prompt length is a hyperparameter of automated prompt induction algorithms, fixed at 5 tokens without tuning. This result might suggest that the algorithms only associated meaningful information to a subset of these tokens.

Machine prompts are drifting away from Human prompts in the input space. Figure 2 shows that, for both OPT-350m and OPT-1.3b, input similarity within prompt types is higher than similarity between different prompt types. In particular, the input similarity between *human* and Machine prompts dramatically decreases when moving from discrete to continuous prompts.

4.2 Surprising aspects of LM processing

The significant differences that emerged between human and machine prompt processing suggest that these prompt types trigger different decision pathways. Furthermore, they provide interesting insights concerning the nature of LM processing, and, in particular, how it can occasionally be quite counter-intuitive. We explore this by considering some unexpected correlation patterns.

Perplexity does not predict accuracy Gonen et al. (2022) reported a negative correlation between perplexity and effectiveness of handcrafted prompts. However, we observe that, when using machine-generated prompt, it is possible to reach a higher prediction accuracy while having a higher perplexity. Thus, counter-intuitively, perplexity does not necessarily predict effectiveness.

Input similarity does not predict output agreement The *Input predicts Output?* column of Table 2 measures the correlation between embedding-space similarities of same-task templates (e.g., a *human* and a *M-disc* template for the *continent of* relation) and the rate of output agreement (defined as the portion of times different templates lead to the same prediction) for the corresponding templates. There is a significantly lower correlation when templates belonging to different prompt types are compared, especially when comparing human vs. machine-generated templates (e.g., *human* vs. *M-disc* templates), than for within-type comparisons (e.g., different *M-disc* templates for the same task). When comparing different prompt types, counter-intuitively, the degree of similarity

	Input predicts Output?	Input predicts Activation?	Activation predicts Output?
		OPT-350m	
human vs. *M-disc*	**0.11** [-0.06, 0.27]	**0.21** [0.16, 0.26]	0.01 [-0.05, 0.07]
human vs. *M-cont*	**0.14** [0.01, 0.26]	**0.06** [0.02, 0.09]	**-0.04** [-0.08, -0.00]
M-disc vs. *M-cont*	**0.30** [0.18, 0.42]	**0.06** [0.03, 0.08]	**0.02** [-0.01, 0.05]
human vs. *human*	**0.53** [0.43, 0.62]	**0.73** [0.69, 0.77]	**0.66** [0.59, 0.72]
M-disc vs. *M-disc*	**0.54** [0.45, 0.63]	**0.85** [0.83, 0.88]	**0.55** [0.49, 0.63]
M-cont vs. *M-cont*	**0.63** [0.55, 0.72]	**0.74** [0.69, 0.78]	**0.54** [0.47, 0.61]
		OPT-1.3b	
human vs. *M-disc*	**0.18** [0.07, 0.28]	**0.24** [0.18, 0.30]	**0.13** [0.05, 0.22]
human vs. *M-cont*	**-0.06** [-0.21, 0.08]	**0.08** [0.03, 0.14]	**-0.03** [-0.08, 0.03]
M-disc vs. *M-cont*	**0.12** [0.01, 0.22]	**0.03** [-0.01, 0.07]	-0.00 [-0.05, 0.04]
human vs. *human*	**0.58** [0.52, 0.66]	**0.65** [0.61, 0.68]	**0.60** [0.55, 0.65]
M-disc vs. *M-disc*	**0.64** [0.58, 0.70]	**0.89** [0.87, 0.90]	**0.61** [0.55, 0.67]
M-cont vs. *M-cont*	**0.78** [0.73, 0.83]	**0.74** [0.71, 0.77]	**0.53** [0.49, 0.57]
Notation: **average** [95% confidence interval]			

Table 2: Pearson correlations between input similarity, output agreement and activation overlap. First, we compute a single comparative statistic (input similarity, output agreement or activation overlap) for each pair of prompts in some comparison set (e.g., *human* vs. *M-disc* or *human* vs. *human*); then, for each comparison set, we look at the correlation across prompt pairs between two statistics (e.g., input similarity vs. activation overlap). Within-type correlations range from mild to high. Between-type correlations are significantly lower. These low correlations highlight counteractive aspects of LM language processing. Results in **bold** are significant ($p < 0.01$). We provide the average and [95%CI interval] correlations obtained using bootstrapped uncertainty estimation.

is not a good predictor of whether the templates will trigger the same output or not.

Activation overlap is only weakly correlated with output agreement and input similarity Table 2 also provides correlations between activation overlap and input similarity (*Input predicts Activation?*) or output agreement (*Activation predicts Output?*) across different pairwise prompt type combinations. Within a prompt type, these correlations are mild or high, with the higher correlations pertaining to input similarity. On the contrary, activation overlap ceases to be correlated to either input similarity or output agreement as soon as we compare different prompt types. This drop in correlation highlights the complexity of LM internal processing. Without any prior on input type, it is difficult to predict the decision pathway that will be used by the model, even in the presence of high input or output similarities.

5 A closer look at the typical units of each prompt type

5.1 Unit distribution across layers

The low activation overlap between prompt types reported in Section 4 taught us that machine-generated prompts trigger units which are distinct from the ones triggered by human prompts. The units that are most often activated by the various prompt types also appear, to some degree, to be distributed differently across layers (cf. Figure 3). In particular, machine prompts display a tendency to activate more units on the last layer and, especially, on the first one (it is worth recalling that this is the first proper Transformer layer, and not the embedding layer). The *M-disc* profile lies somewhere between *human* and *M-cont*, confirming the trends already observed in Section 4.

5.2 Profiling prompt-type-typical units through associated vocabulary items

Having shown that the three prompt types activate different pathways through the network, we seek now some insights into the nature of the units characterizing these different pathways.

Methodology We identify those units that are both *typical* of a single prompt type across tasks, and significantly impacting the network output distribution, in the sense that their gradients w.r.t to the max output probability are in the top quartile of all network units (recall that, as usual, we focus on those units we identified as knowledge neu-