Phase transitions in the output distribution of large language models

Julian Arnold[1], Flemming Holtorf[2], Frank Schäfer[2], Niels Lörch[1]

1) University of Basel
2) CSAIL MIT

Abstract

The behavior of LLMs has been observed to exhibit rapid, qualitative changes reminiscent of phase transitions in response to variation in tuning parameters such as temperature or training epoch. While these transitions are of particular interest in developing a better understanding of LLMs, it remains an open question how to define and detect them systematically.

The mathematical analogy between transformer-based LLMs and physical systems akin to the Ising (or more generally Potts) model suggests to draw on established insights from statistical physics to address these questions.

Here [1], we adapt a technique [3] for the automated detection of phase transitions in physical systems to the context of LLMs. In broad strokes, the proposed technique quantifies changes in the entire output distribution of LLMs via statistical distances, giving rise to universal measures of behavioral change that avoid the subjectivity and pitfalls of adhoc-defined indicators.

f-divergences: Measuring distances between probability distributions

example:
Jensen-Shannon divergence as symmetrized Kullback-Leibler divergence

$$D_{\mathrm{JS}}[p,q] = \frac{1}{2} D_{\mathrm{KL}}\left[p, \frac{p+q}{2}\right] + \frac{1}{2} D_{\mathrm{KL}}\left[q, \frac{p+q}{2}\right]$$

general f-divergence for samples x, probability distributions p and q

$$D_f[p,q] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \geq 0.$$

In practice: Pseudo-code implementation of sampling from LLMs at different temperatures

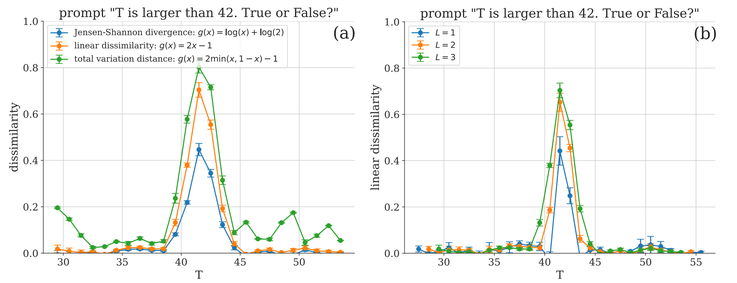
f-divergences are mappable to corresponding g-dissimilarities.

Algorithm 1 For text samples $\boldsymbol{x}$ distributed according to a temperature-dependent distribution $P(\cdot|T)$ induced by a language model, the g-dissimilarity between neighboring points T_n and T_{n+1} on a grid of temperatures is given by $D_g = \frac{1}{2}\sum_{j=n,n+1} \mathbb{E}_{\boldsymbol{x}\sim P(\cdot|T_j)} g\left[p(T_j|\boldsymbol{x})\right]$, where $p(T_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|T_j)}{P(\boldsymbol{x}|T_n)+P(\boldsymbol{x}|T_{n+1})}$. For a language model "model" and "g_function" g, it is estimated as:

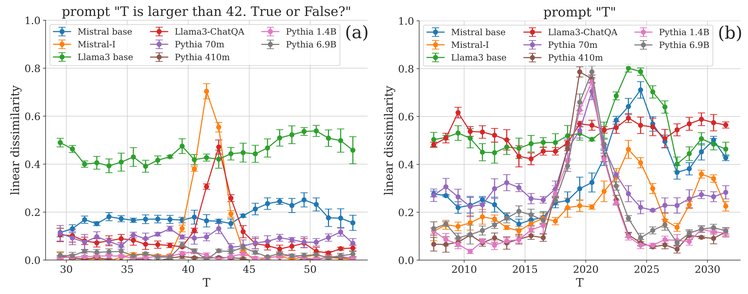
Input: model, n, T_n, T_{n+1}, n_{samples}, g_function
Output: g-dissimilarity D_g

$D_g = 0$
for i in $\{1, \ldots, n_{\mathrm{samples}}\}$ **do**
  for j in $\{n, n+1\}$ **do**
    sample = model.generate(temperature=T_j)
    p_n = model.evaluate_probability(sample, temperature=T_n)
    p_{n+1} = model.evaluate_probability(sample, temperature=T_{n+1})
    $D_g = D_g$ + g_function($p_j/(p_n + p_{n+1})$)
  end for
end for
$D_g = D_g/(2n_{\mathrm{samples}})$

Simple application: Transitions for varying prompts

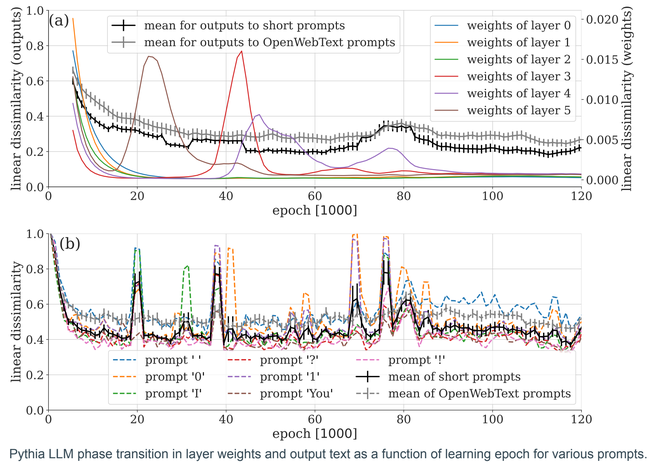


Mistral model applied to integer ordering prompt varying T.
(a) Different g-dissimilarities with length parameter L = 3. (b) Linear dissimilarity for different values of length parameter L.



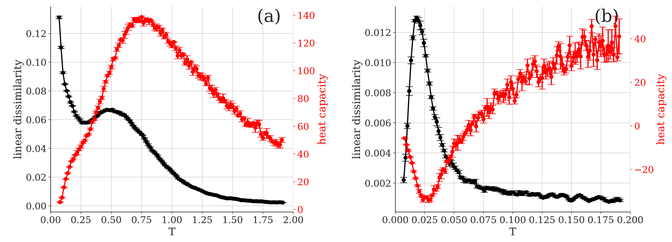
Benchmarking various models using the linear dissimilarity with L=3.
(a) Test of ability to compare integers in value. (b) Bare integers as prompt reveals transition in tokenizer encoding.

Transitions for varying training epoch

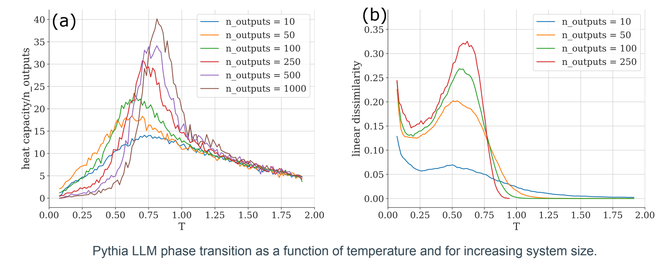


Pythia LLM phase transition in layer weights and output text as a function of learning epoch for various prompts.

Transitions for varying temperature

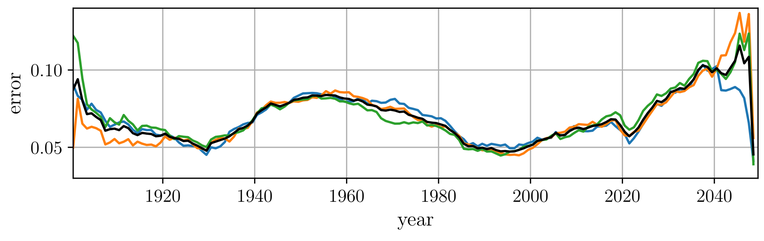


Pythia LLM phase transition as a function of temperature at temperatures close to (a) 1 and (b) 0.



Pythia LLM phase transition as a function of temperature and for increasing system size.

Generalization to models without tractable densities: Transitions in Stable Diffusion



Transitions in Stable Diffusion with prompt f"technology of the year {year}" for various years.
here, no direct access to the underlying probability distribution is available and we resort to "learning-by-confusion".
The error here is proportional to minus the total variation distance, which is a f-divergence, so valleys indicate transitions.
see Reference [2] for details.

References

[1] J. Arnold, F. Holtorf, F. Schäfer, N. Lörch, "Phase Transitions in the Output Distribution of Large Language Models", arXiv:2405.17088
[2] J. Arnold, F. Schäfer, N. Lörch, "Fast detection of phase transitions with multi-task learning-by-confusion", arXiv:2311.09128
[3] J. Arnold, F. Schäfer, A. Edelman, C.Bruder, "Mapping out phase diagrams with generative classifiers", Phys. Rev. Lett. 132, 207301