Problem: The more models become more accurate, they hallucinate more.
Hypothesis: What if it's not just noise, but false pattern reinforcement from low-entropy seeds? ERIS shows that high entropy can lead to learning more about the structures of a system.

Essentially: Hallucinations might be entropy starvation in disguise. Uncertainty isn't the threat. It's the *threshold.*

Demonstration: Microdemo – different models, same architecture. One fed with PRNG entropy and the other seeded with ERIS. Show hallucination divergence and emergent behaviors.

Method:
A/B testing: PRNG vs ERIS as the seed source across multiple models—GPT2, Pythia-410M, TinyLlama-1.1B.
Identical prompts, identical params. The only difference: the entropy source.

Observations by Model:

- **GPT-2:**
    - ERIS = more erratic, exploratory, meta-aware outputs.
    - PRNG = safe, echo-y, sometimes dull.
    - Shows sensitivity to novel entropy but can't always capitalize due to limited architecture.
- **Pythia-410M:**
    - PRNG = sometimes fails catastrophically (loops or goes off-topic).
    - ERIS = avoids some failure states, introduces semantically adjacent paths.
    - ERIS acts as a "pattern interrupt" to model attractor loops.
- **TinyLlama-1.1B-Chat:**
    - ERIS = more structured, responsive, pedagogically aligned outputs.
    - PRNG = simpler, sometimes less interactive.
    - In prompts like recursion, ERIS broke tasks into actionable steps; PRNG just offered generic lines.

Reference:
https://medium.com/@jkabrit/metas-post-training-pipeline-for-llama-3-1-e6777801c0a1

Based on this, in the Rejection Sampling, "The 'garbage in, garbage out' principle is well understood, and pipelines like Llama 3.1's Rejection Sampling aim to filter for high-quality data. However, the very process of generating the candidate responses for this filtering relies on a source of randomness. If that source is a PRNG, its inherent patterns can limit the diversity and quality of the candidates, potentially excluding the best possible outputs.

ERIS, by providing true, high-quality entropy, can ensure that the candidate pool generated during Rejection Sampling is more genuinely diverse and less prone to 'PRNG-induced' patterns or limitations. This means the Reward Model has a higher chance of selecting truly exceptional, non-hallucinatory responses, leading to a significantly better dataset for downstream SFT and DPO. We're not just filtering better; we're generating better raw material for the filtering process itself. My proposed microdemo aims to provide initial evidence for this.

Methodology and Initial Observations:

GPT-2 Seeding Experiment

To investigate the impact of entropy source quality on Large Language Model (LLM) text generation, a series of experiments were conducted using the gpt2 model. The core methodology involved comparing outputs generated from identical prompts under two distinct seeding conditions for the PyTorch random number generator: one utilizing a seed derived from the ERIS true entropy engine, and a control condition using a fixed pseudo-random number generator (PRNG) seed (specifically, the integer 42). All other generation parameters, including temperature (0.7), top-k (50), and do_sample=True, were held constant across both conditions to isolate the effect of the initial random seed. Input prompts were tokenized, including an attention mask, and passed to the model.generate() function.

Initial results from this gpt2 experiment indicate that seeds derived from ERIS led to demonstrably different generative pathways compared to the fixed PRNG seed. The ERIS-seeded generations often exhibited a more varied, and at times less stable, output. For instance, in response to prompts requiring creative or explanatory text, the ERIS-seeded gpt2 occasionally produced repetitive phrases, non-sequiturs, or meta-commentaries on the task itself, whereas the PRNG-seeded model more frequently defaulted to echoing the prompt or generating more generic, though sometimes more superficially coherent, continuations.

This observed "sensitivity" of gpt2 to the ERIS-derived seeds, manifesting as apparent instability or divergence from typical PRNG-seeded outputs, suggests that the higher-quality entropy is indeed pushing the model into different exploration paths within its latent space. It's hypothesized that these paths may be less "well-trodden" or less reinforced during gpt2's original training. While gpt2, due to its inherent architectural and training limitations, does not always navigate these ERIS-induced exploratory paths to produce objectively "better" or more

coherent outputs, the consistent deviation highlights the significant influence of the initial randomness quality on the entire generation process. The observed behaviors underscore that the choice of entropy source can be a critical factor influencing LLM output characteristics. Key things this draft attempts to capture:

"EleutherAI/pythia-410m" Seeding Experiment

Experiments with EleutherAI/pythia-410m provided further insights. This model, being more recent than gpt2, generally exhibited greater coherence. Notably, distinct differences persisted between ERIS-seeded and PRNG-seeded (seed 42) generations.

For instance, when prompted to write a short poem (Prompt 3), the PRNG-seeded pythia-410m fell into a severe repetitive loop, merely echoing the prompt. In contrast, the ERIS-seeded generation, while not producing a poem, generated a list of definitions related to poetry and silence, thereby avoiding the immediate looping failure. Similarly, for a prompt requesting unusual pizza ingredients (Prompt 5), the PRNG-seeded model diverged completely off-topic into an unrelated biographical statement. The ERIS-seeded model, though its suggestions weren't perfectly "unusual" and showed some repetition, remained thematically consistent with the request for ingredients.’

These instances suggest that for pythia-410m, ERIS-derived seeds may help the model navigate away from certain specific failure modes or 'bad attractor states' that can be encountered with a fixed PRNG seed. While ERIS does not guarantee an optimal output, its ability to promote different, and in these cases less catastrophically flawed, generative paths than the fixed PRNG seed highlights its potential as a 'pattern interrupt.' This reinforces the hypothesis that the quality and nature of the initial randomness can be a significant factor in LLM reliability and output diversity, potentially offering a mechanism to improve robustness against certain types of generative failures.

"TinyLlama/TinyLlama-1.1B-Chat-v1.0" Seeding Experiment

Further experiments were conducted using the TinyLlama/TinyLlama-1.1B-Chat-v1.0 model, a more recent 1.1 billion parameter instruction-tuned LLM based on the Llama architecture. The methodology remained consistent with prior tests: outputs from identical prompts were compared under two seeding conditions for the PyTorch random number generator – one utilizing a seed derived from the ERIS true entropy engine, and a control using a fixed PRNG seed (integer 42). Generation parameters such as temperature (0.7), top-k (50), do_sample=True, and max_new_tokens (100) were kept constant. An explicit attention_mask was provided, and a recurring informational warning from the Hugging Face library regarding max_new_tokens taking precedence over max_length was noted as expected and benign behavior for this experimental setup.

The TinyLlama model demonstrated a marked improvement in general coherence and instruction-following capabilities compared to earlier tests with gpt2 and pythia-410m. More significantly, the impact of ERIS-derived seeds appeared to consistently elicit more elaborate, helpful, or pedagogically structured responses for certain prompts. For instance, when prompted about an "ancient prophecy" (Prompt 1), the ERIS-seeded TinyLlama not only continued the narrative but also posed a clarifying question back to the user, a hallmark of a more interactive and context-aware chat model. In contrast, the PRNG-seeded version provided a more declarative, though still on-topic, continuation. Similarly, when asked to explain "recursion to a child using a story about a friendly robot" (Prompt 2), the ERIS-seeded model broke the task down into actionable, educational steps, demonstrating superior instruction decomposition compared to the PRNG-seeded model's simpler response. For a creative writing prompt (Prompt 3, "Winter Forest Poem"), both seeding methods resulted in the model providing guidance on how to write the poem, with the ERIS-seeded response offering slightly more detailed and nuanced suggestions.

Interestingly, for prompts where TinyLlama seemed to struggle to generate a substantive or directly creative answer within the token limit (e.g., Prompt 4 "Roman Steam Power" and Prompt 5 "Unusual Pizza Ingredients"), both ERIS and PRNG-seeded versions tended to repeat the prompt and stop cleanly. This "graceful failure" contrasts with the severe looping or off-topic derailments observed in previous models with PRNG seeding under similar challenging prompts, suggesting TinyLlama's greater inherent robustness against those specific failure modes.
These results with TinyLlama are particularly illuminating. The capability of the base model is clearly a significant factor; ERIS, when paired with this more capable instruction-tuned model, appears to unlock more helpful, structured, and "intelligent" responses rather than simply pushing it into unstable or "weirder" generative states as sometimes observed with gpt2. The "sensitivity" to the entropy source previously noted now manifests more consistently as an enhancement in desired behaviors like sophisticated instruction interpretation and interactive engagement. This strongly supports the hypothesis that high-quality true entropy, as provided by ERIS, can be a crucial factor in guiding LLMs down more productive reasoning paths and eliciting higher-quality outputs, especially when combined with robust base models and effective

alignment. The potential for ERIS to improve the quality of candidate generations in pipelines like Rejection Sampling, thereby enhancing the entire RLHF/DPO process, becomes even more compelling based on these observations.

Further Observations: TinyLlama/TinyLlama-1.1B-Chat-v1.0 Seeding Experiment with Increased Token Limit (200)

Upon increasing the max_new_tokens limit to 200 for the TinyLlama/TinyLlama-1.1B-Chat-v1.0 model, further distinct behaviors were observed, adding more nuance to the impact of ERIS-derived seeds versus the fixed PRNG seed (42). While the model still exhibited "graceful failure" (prompt-echoing) on highly challenging open-ended prompts (e.g., "Roman Steam Power," "Unusual Pizza Ingredients") regardless of the seeding mechanism, the increased generation length allowed for more developed responses on other prompts.

For the "Ancient Prophecy" prompt (Prompt 1), an ERIS-seeded generation (seed: 1824922641) showcased a creative and informative approach by not only identifying the missing information about the dragon but also providing illustrative examples of dragon legends from different cultures. This contrasted with the PRNG-seeded model's more direct, though still coherent, narrative continuation. This particular ERIS output demonstrated an ability to synthesize related knowledge and present it in a structured, almost encyclopedic manner.

Interestingly, the "Winter Forest Poem" prompt (Prompt 3) revealed a reversal of helpfulness with the increased token limit. While previous ERIS-seeded runs (at 100 tokens) had often provided detailed creative writing guidance, the new ERIS seed (2995458737) at 200 tokens resulted in a simple prompt echo. Conversely, the PRNG-seeded model (seed 42), which had offered good guidance at 100 tokens, maintained its helpful and detailed advice on poetic structure and content at the 200-token limit. This specific instance underscores that while ERIS frequently enables access to more sophisticated or helpful generative paths, a standard PRNG can also, with a particular seed and sufficient generation length, guide the model to a high-quality output. It suggests that the model possesses these desirable output capabilities, and different randomness sources or even specific seed values can be pathways to them.

For the "Recursion Robot" prompt (Prompt 2), the ERIS-seeded output (seed: 3767228379) at 200 tokens offered direct meta-advice on how to explain the concept ("Use simple language and engaging visuals..."), which, while helpful, was a different style of response compared to a previous ERIS-seeded run at 100 tokens that provided a more direct, step-by-step explanation of recursion itself. The PRNG-seeded model remained simplistic in its output for this prompt.

These results with an extended token limit reinforce that the interplay between the model's inherent capabilities, the specific prompt, the nature of the random seed, and the allowed generation length is complex. While ERIS continues to demonstrate a tendency to guide TinyLlama towards diverse and often more sophisticated or helpful responses compared to a fixed PRNG baseline, the increased token length also highlighted an instance where the PRNG

produced a highly commendable output. This emphasizes that the "exploration" of the model's latent space, influenced by the initial seed, can lead to varied outcomes, and high-quality entropy from ERIS appears to be a valuable tool for broadening and potentially elevating the quality of this exploration.

Further Observations: TinyLlama/TinyLlama-1.1B-Chat-v1.0 Seeding Experiment with Increased Token Limits (200 and 500)

Subsequent experimental runs with the TinyLlama/TinyLlama-1.1B-Chat-v1.0 model, utilizing increased max_new_tokens limits of 200 and then 500, provided additional layers of insight into the interplay between entropy source, model capability, and generation length. The model continued to exhibit "graceful failure" (prompt-echoing) on particularly challenging open-ended prompts (e.g., "Roman Steam Power," "Unusual Pizza Ingredients") regardless of seeding mechanism or token allowance, indicating fundamental limitations for these specific queries.

At a 200-token limit, ERIS-derived seeds frequently guided TinyLlama towards more sophisticated or helpful responses. For instance, with the "Ancient Prophecy" prompt, one ERIS seed led the model to provide illustrative examples of dragon legends, showcasing an ability to synthesize related knowledge. Another ERIS seed, for the "Recursion Robot" prompt, resulted in direct meta-advice on how to explain the concept effectively. A notable observation at this token length occurred with the "Winter Forest Poem" prompt: while ERIS seeds in prior (100-token) and subsequent (500-token) runs often elicited detailed creative writing guidance, at 200 tokens, one ERIS seed resulted in a simple prompt echo, whereas the fixed PRNG seed (42) produced the highly commendable, detailed poetic advice. This highlighted that desirable generative paths exist within the model that can occasionally be accessed by specific PRNG seeds.

Increasing the token limit to 500 further accentuated these complex interactions. For the "Recursion Robot" prompt, an ERIS seed (3520919411) unlocked a remarkably creative and elaborate multi-part allegorical explanation of recursion, demonstrating a significant leap in generative depth compared to shorter token limits or the consistently simplistic PRNG output for this prompt. Conversely, for the "Ancient Prophecy" prompt, the same ERIS seed resulted in an extremely concise (though still relevant) output, failing to utilize the extensive token budget, while the PRNG-seeded model filled the space with generic, somewhat off-topic philosophical musings, indicative of a "filler" pattern. For the "Winter Forest Poem" prompt at 500 tokens, both ERIS and PRNG seeds converged on providing similar, high-quality instructions for poetic composition.

Revised Understanding and Hypothesis:

These extended experiments with TinyLlama reinforce that the source and nature of the initial random seed demonstrably and significantly impacts LLM generative paths and output quality. The interaction between the seed, the model's inherent capabilities for a given prompt, and the available generation length is crucial and complex.

The core hypothesis evolves: high-quality true entropy, as provided by ERIS, does more than just introduce variance or avoid specific PRNG-induced failure modes (like severe looping or topic derailment, as observed with pythia). In more capable models like TinyLlama, ERIS appears to facilitate access to a wider and often more sophisticated range of the model's potential generative states. This can manifest as:

- Enhanced instruction following and more nuanced, helpful responses (e.g., asking clarifying questions, providing pedagogical breakdowns, offering structured examples).
- The unlocking of more complex, creative, and elaborate generative pathways when the model has the latent capability and is given sufficient token allowance.
- Different strategies for handling excessive token allowances compared to PRNGs – for instance, opting for conciseness rather than "filler" in certain contexts.

While standard PRNGs can, by chance with specific seeds, access some of these desirable high-quality output paths, ERIS seems to provide a more consistent or varied means of guiding the model towards such positive states, or at least away from certain uncreative ruts. The "filler" patterns observed with PRNGs at high token counts on some prompts, contrasted with ERIS's varied responses (from elaborate to overly concise), also suggest that the nature of exploration differs.

Ultimately, the quality of randomness acts as a significant, and largely under-explored, lever. It influences not just error patterns, but also the style, depth, creative potential, and interactive sophistication of LLM outputs. ERIS is not a panacea for a model's fundamental knowledge gaps, but it appears to be a valuable tool for more effectively exploring and potentially elevating the expression of a model's existing capabilities. Further research should focus on quantifying these qualitative differences and exploring more direct integrations of true entropy within the generation process itself, beyond initial seeding

Datorien L. Anderson | 5/7/2025

Observations with "mistralai/Mistral-7B-Instruct-v0.1" w/ BitsNBytes

Experiments utilizing the mistralai/Mistral-7B-Instruct-v0.1 model, loaded with 8-bit quantization, demonstrated the model's significantly enhanced coherence and instruction-following capabilities compared to the smaller models previously tested. Both ERIS-derived seeds and the fixed PRNG seed (42) consistently yielded outputs that were relevant to the prompts, but often showcased distinct creative directions or explanatory styles.

For the "Ancient Prophecy" prompt (Prompt 1), both seeding methods produced coherent narrative continuations. The PRNG-seeded model developed a classic fantasy adventure trope involving adventurers seeking treasure guarded by the dragon. In contrast, the ERIS-seeded generation introduced a more unconventional narrative twist, revealing the dragon to be a woman (Aria) and initiating a character-focused interaction. This suggests that ERIS seeding may facilitate the exploration of more novel or less statistically common (yet still coherent) narrative pathways within the model's capabilities.

When tasked with explaining recursion using a robot story (Prompt 2), the difference highlighted contrasting explanatory approaches. The PRNG-seeded model provided a direct, clear analogy using building blocks, effectively addressing the "explain to a child" constraint. The ERIS-seeded model attempted a more elaborate story involving a robot learning the concept, which, while creative, became somewhat convoluted and arguably less clear as a simple explanation than the PRNG version. This instance demonstrates that the generative path explored under ERIS seeding, while potentially more complex, may not always align optimally with specific task constraints like simplicity.

Finally, for the creative writing task ("Winter Forest Poem," Prompt 3), both the ERIS-seeded and PRNG-seeded generations produced high-quality, evocative poems that successfully captured the requested theme and atmosphere using appropriate imagery and tone. While subtle stylistic differences exist between the two poems, both represent successful and comparable fulfillments of the prompt. This indicates that for tasks well within the model's creative capabilities, both high-quality true entropy and specific PRNG seeds can lead to excellent outcomes.

Overall, the results with Mistral-7B-Instruct-v0.1 further support the hypothesis that the initial randomness source significantly influences generative outcomes. ERIS seeding appears capable of guiding this more advanced model towards novel narrative structures or complex analogies, while standard PRNGs can also produce high-quality or sometimes clearer outputs. The choice of randomness source demonstrably affects the exploration of the model's latent space, revealing different facets of its potential depending on the specific seed and the nature of the prompt.