

Evaluate the paper that was uploaded according to the criteria of this conference: <https://agents4science.stanford.edu/call-for-papers.html>. Every paper will also be scored on the following components: quality, clarity, significance, and originality. Check for inconsistencies and if the references are ok. Propose improvements and formulate the improvements in LaTeX so that they can be included.

I've reviewed your draft "**From Borges' Library to Procedural Universes**" against the *Agents4Science* conference criteria (quality, clarity, significance, originality). Here's a structured evaluation and improvement plan:

Evaluation

✓ Strengths

- **Quality:** The theoretical framework is rigorous. Theorems, lemmas, and propositions are well-structured and connect classical information theory, complexity, and LLM practice.
- **Clarity:** Definitions are formal and consistent; proofs are sketched clearly. Figures (entropy reduction, best-of-N, hallucination decomposition) illustrate abstract concepts effectively.
- **Significance:** Addresses *fundamental questions* about navigability, hallucination, and epistemic limits of LLMs — highly relevant to trustworthy AI research.
- **Originality:** The "procedural library" metaphor (anti-Babel) is original and positions the paper in dialogue with Borges. The decomposition of hallucination risk into coverage/abstention/error is novel.

⚠ Weaknesses

- **Inconsistencies:**
 - References are partly incomplete: some [e.g., Shannon, Solomonoff, Chaitin] are classic, but citation formatting is inconsistent (sometimes "In ..." journal, sometimes missing details like page numbers).
 - "Figure 2 (from Section 2)" is misleading; it belongs to Section 4.
 - The empirical validation is very lightweight: BASE system achieves $pf = 1.0$ across all 12 questions, which weakens the motivation for the theoretical metrics.
 - Section titles sometimes mix conceptual framing and technical detail (e.g.,

“Hallucinations as Residual Noise” vs. “Computational and Epistemic Limits”).

This makes the paper slightly uneven.

- **Clarity issues:**

- Abstract and introduction are dense; could be smoothed for readability.
- Notation: using both $\mathcal{V}f(\mathcal{O})$ and $\mathcal{N}\mathcal{I}$ creates redundancy.
- The connection between *operators* (prompt, soft prompt, retrieval) and formal *operator calculus* could be explained with a clearer bridging example.

- **References check:**


- All main references are real and relevant, but several are under-detailed:
 - [11] Rissanen (1978) → journal Automatica, but “shortest data description” is not the standard title (should be “Modeling by shortest data description”).
 - [14] Solomonoff (1964) should have two parts (Part I, II, *Information and Control*).
 - [4] FAISS reference is normally cited as a conference/workshop paper first, not just IEEE TBD.

Suggested Improvements (with LaTeX inserts)

Here are concrete improvements, already in LaTeX so you can drop them in:

1. Abstract (clarity)

latex

 Code kopieren

```
\begin{abstract}
Large Language Models (LLMs) can be understood as \emph{procedural libraries}:
instead of storing all texts, they generate strings on demand according to a
learned distribution  $P_{\theta}$  over  $\Sigma^*$ . This paper develops a
theoretical framework for such libraries, focusing on suppression,
navigability, and inherent limits. We (i) formalize typical-set suppression
that concentrates probability on coherent strings, (ii) define operators
(prompts, soft prompts, retrieval) as entropy-reducing mechanisms, (iii)
analyze navigability through success probability, hitting time, and energy
bounds, and (iv) decompose hallucination risk into coverage, abstention, and
conditional error. We also prove complexity-theoretic lower bounds, connect
retrieval to submodular information acquisition, and propose design metrics. A
lightweight empirical study illustrates how these metrics can be
operationalized. Together, our results bridge information theory and modern
LLM practice, offering principles for trustworthy and controllable generative
systems.
\end{abstract}
```

2. Reference consistency

In .bib :


```
@article{Solomonoff1964a,
  author    = {Ray Solomonoff},
  title     = {A Formal Theory of Inductive Inference. Part I},
  journal   = {Information and Control},
  volume    = {7},
  number    = {1},
  pages     = {1--22},
  year      = {1964}
}

@article{Solomonoff1964b,
  author    = {Ray Solomonoff},
  title     = {A Formal Theory of Inductive Inference. Part II},
  journal   = {Information and Control},
  volume    = {7},
  number    = {2},
  pages     = {224--254},
  year      = {1964}
}
```

3. Clarify notation

Replace dual definitions of navigability index:

latex

 Code kopieren

```
\paragraph{Navigability Index.}
For a predicate  $f$ , define
\begin{equation}
\text{NI}_f(0) := - \log p_f(\emptyset) + \log p_f(0),
\end{equation}
the log-improvement in success probability relative to the unconditional
model.
```

4. Improve Figure references


latex

 Code kopieren

As illustrated in Figure~\ref{fig:bestofn}, best-of- N sampling amplifies success probability.

with consistent labels:

latex

 Code kopieren

```
\begin{figure}[t]
```

```

\begin{figure}
\centering
\includegraphics[width=0.7\linewidth]{bestofn.pdf}
\caption{Best-of- $N$  success probability  $1-(1-p)^N$  for base probabilities  $p$ 
\in  $\{0.05, 0.10, 0.20\}$ }.}
\label{fig:bestofn}
\end{figure}

```

5. Discussion expansion

latex

 Code kopieren

```
\section{Limitations}
```

Our empirical validation is limited to a toy dataset where the BASE condition already achieves perfect accuracy. Consequently, the improvements of FEWSHOT and RAG could not be meaningfully assessed. Future work should evaluate the proposed metrics on harder benchmarks (e.g., TruthfulQA, MMLU) to test the generality of our theoretical predictions.

Summary

- **Quality:** High (solid theory, but empirical part weak).
- **Clarity:** Medium-high (dense abstract/intro, some notation duplication).
- **Significance:** High (addresses core limits of LLMs).
- **Originality:** Very high (procedural library framing, hallucination decomposition).

With the above refinements (clearer abstract, consistent references, stronger limitations section), the paper will be in good shape for *Agents4Science*.

Do you want me to also **suggest concrete review scores (1–5)** for each dimension (quality, clarity, significance, originality) as if I were a program committee reviewer?

   ...