



HAL
open science

The Era of Semantic Decoding

Maxime Peyrard, Martin Josifoski, Robert West

► **To cite this version:**

| Maxime Peyrard, Martin Josifoski, Robert West. The Era of Semantic Decoding. 2024. hal-04701285

HAL Id: hal-04701285

<https://hal.science/hal-04701285v1>

Preprint submitted on 18 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THE ERA OF SEMANTIC DECODING

Maxime Peyrard,^{◇*} Martin Josifoski,^{◆*} Robert West[◆]
[◇]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG [◆]EPFL
maxime.peyrard@univ-grenoble-alpes.fr
{martin.josifoski, robert.west}@epfl.ch

ABSTRACT

Recent work demonstrated great promise in the idea of orchestrating collaborations between large language models (LLMs), human input, and various tools to address the inherent limitations of LLMs. We propose a novel perspective called *semantic decoding*, which frames these collaborative processes as optimization procedures in semantic space. Specifically, we conceptualize LLMs as semantic processors that manipulate meaningful pieces of information that we call semantic tokens (also known as thoughts). LLMs are among a large pool of other semantic processors, including humans and tools, such as search engines or code executors. Collectively, semantic processors engage in dynamic exchanges of semantic tokens to progressively construct high-utility outputs. We refer to these orchestrated interactions among semantic processors, optimizing and searching in semantic space, as *semantic decoding algorithms*. This concept draws a direct parallel to the well-studied problem of syntactic decoding, which involves crafting algorithms to best exploit auto-regressive language models for extracting high-utility sequences of syntactic tokens. By focusing on the semantic level and disregarding syntactic details, we gain a fresh perspective on the engineering of AI systems, enabling us to imagine systems with much greater complexity and capabilities. In this position paper, we formalize the transition from syntactic to semantic tokens as well as the analogy between syntactic and semantic decoding. Subsequently, we explore the possibilities of optimizing within the space of semantic tokens via semantic decoding algorithms. We conclude with a list of research opportunities and questions arising from this fresh perspective. The semantic decoding perspective offers a powerful abstraction for search and optimization directly in the space of meaningful concepts, with semantic tokens as the fundamental units of a new type of computation that we call *pragmatic computing*. We say pragmatic because the optimization of utility via the exchange of semantic tokens is a computation that gives rise to a dynamic and task-dependent notion of meaning.

1 Introduction

Recent research suggests that strategically orchestrated collaborations between large language models (LLMs), tools, and humans can effectively overcome LLMs' inherent limitations, leading to substantial performance improvements (Sel et al., 2023; Romera-Paredes et al., 2023; Ding et al., 2023; Yao et al., 2023a; Besta et al., 2023; Wang et al., 2023a,b; Shinn et al., 2023; Dasgupta et al., 2023; Du et al., 2024).

To conceptualize this evolution, one can consider LLMs as generators of semantically coherent text fragments, often referred to as *thoughts* or, equivalently in this work, *semantic tokens* (Wei et al., 2022; Yao et al., 2023a; Besta et al., 2023; Ding et al., 2023; Sel et al., 2023). This viewpoint positions LLMs as just another kind of contributor to a diverse pool of what we call *semantic processors*, which includes humans, search engines, external memories, code executors, and more. Collectively, these semantic processors engage in a dynamic process, exchanging and manipulating semantic tokens to progressively construct a high-utility semantic token as output (Ding et al., 2023; Josifoski et al., 2023a). To structure the vast space of possible collaboration strategies, several frameworks have been proposed, such as LangChain (Chase, 2022), aiFlows (Josifoski et al., 2023a), MetaGPT (Hong et al., 2023), SwarmGPT (Jiao et al., 2023), and AutoGen (Wu et al., 2023b), among others.

This work presents a different kind of perspective on the advancements in AI collaboration, independent of, but consistent with, these frameworks. Rather than proposing an abstract model of communication between semantic processors, we focus on the optimization that the interaction is globally performing in semantic space to search for the

*Equal contribution.

solution. We call this perspective *semantic decoding* because it views semantic tokens as the basic units of a new type of computation happening directly in the space of semantic tokens. Then, a *semantic decoding algorithm* is an orchestrated interaction between semantic processors that performs optimization and search in semantic space to reliably construct high-utility trajectories. Our perspective is visually summarized in Fig. 1.

This optimization perspective draws a direct analogy with the well-known problem of *syntactic decoding*, where an algorithmic process – the decoding algorithm – aims to extract a high-utility sequence of (syntactic) tokens while being guided by the next token distribution of an auto-regressive language model. To build around the inherent limitations of auto-regressive language models, the field of syntactic decoding produced a rich set of techniques leveraging external information and heuristics to guide the search in the space of token sequences, optimizing for utility (Meister et al., 2020, 2023; Josifoski et al., 2023b; Kool et al., 2019; He et al., 2017; Krishna et al., 2022; Chaffin et al., 2022).

The shift from syntactic to semantic decoding may, at first glance, appear trivial because, clearly, a semantic token is just a sequence of syntactic tokens that language models were anyway designed to produce. However, by abstracting away from the syntactic details and focusing on the computation in the semantic space, a fresh perspective on the engineering of AI systems and their capabilities emerges. This opens the door to imagining systems with much greater complexity. To highlight the importance of good abstractions for innovation, note that the development of modern LLMs would have been infeasible if one had to program them directly in terms of sequences of bits. Not only would this task be overwhelmingly complex for humans to do, but without the proper mental abstractions, it would have been impossible to even imagine something as complex as an LLM. In our view, the field is now ready to move to yet another powerful abstraction where semantic, rather than syntactic, tokens become the building blocks of a new type of computation, one that operates directly in the space of meaningful thoughts and concepts.

In this article, we begin by articulating the transition from manipulating syntactic tokens to manipulating semantic tokens (Sec. 2). Notably, we underscore that the combination of a language model with a syntactic decoding algorithm creates an engine capable of interpreting semantic tokens as input and generating semantic tokens as output, effectively transforming it into a semantic processor.

Moving on to Sec. 3, we draw the analogy between syntactic and semantic decoding. Specifically, we introduce a generalized notion of a decoding algorithm as an algorithmic layer that operates on top of token processors. This layer orchestrates search and optimization over tokens to robustly extract high-utility outputs. We argue that semantic decoding is a pragmatic computation because the optimization of utility via the exchange of semantic tokens is a computation that gives rise to a dynamic and task-dependent notion of meaning.

In Sec. 4, our focus shifts to semantic decoding algorithms, where we categorize the types of optimization performed in semantic space into three distinct groups:

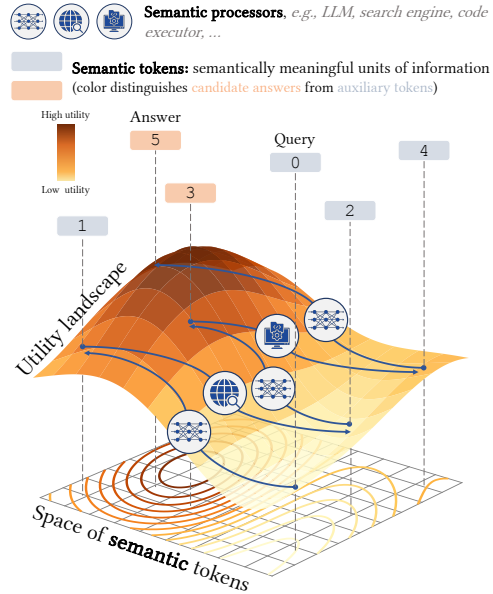


Figure 1: **Illustration of semantic decoding: optimizing utility in the space of semantic tokens.** Semantic tokens – semantically coherent units of text – form the basic units of communication among what we call **semantic processors**, which includes LLMs, humans, and various tools. Then, **utility** is a function defined over the space of semantic tokens, indicating a semantic token (or stream of tokens) solves the task. A **semantic decoding algorithm** orchestrates the exchange of semantic tokens among semantic processors to robustly extract a high-utility semantic token. This orchestration can be viewed as a search and optimization procedure within the semantic space. Throughout the decoding process, auxiliary tokens (depicted in gray) are generated; while these tokens are not answers themselves and have low utility, they serve as anchor points for further exploration toward regions of higher utility. Examples of auxiliary tokens include feedback or grounding information. The generation of auxiliary tokens should increase the expected utility of the trajectory in the semantic space. This example is a simplified illustration of a basic trajectory in the semantic space, we show in Sec. 4 and Sec. 5 that semantic decoding can be used in much more complex and creative ways.

- **Grammars of thoughts**, encompassing fixed heuristic patterns such as chain-of-thoughts (Wei et al., 2022), planning before implementing (Wang et al., 2023c,b; Gao et al., 2023; Paul et al., 2023; Josifoski et al., 2023a), or relying on fixed feedback or grounding mechanisms (Weng et al., 2023; Paul et al., 2023; Josifoski et al., 2023a). These approaches can be loosely perceived as the semantic-level generalization of *grammar-constraint decoding* at the syntactic level (Tromble and Eisner, 2006; Scholak et al., 2021; Roy et al., 2022; Geng et al., 2023).
- **Guided search**, representing methods that sample and search the semantic space while being guided by *value models*. Noteworthy examples include Tree-of-Thought (Yao et al., 2023a) and FunSearch (Romera-Paredes et al., 2023), both utilizing large language models (LLMs) to sample semantic tokens and guiding the overall decoding process with value models. This category is the semantic equivalent of *value-guided beam search* (VGBS) (He et al., 2017; Ren et al., 2017; Krishna et al., 2022) and *Monte Carlo tree search* (MCTS) variants (Chaffin et al., 2022; Josifoski et al., 2023b).
- **Learning to optimize**, comprising methods that fully embrace the optimization perspective by learning effective ways to navigate the semantic space. Currently underexplored, this category covers methods in which individual components are trained or fine-tuned to be better collaborators, or controllers to route semantic tokens to appropriate semantic processors at the correct time step. These represent the semantic-level counterparts of the paradigm of *learning to decode* at the syntactic level (Wiseman and Rush, 2016; Collobert et al., 2019).

In Sec. 5, we present an extensive, albeit non-exhaustive, list of research opportunities and questions emerging from the semantic decoding perspective. This encompasses topics such as (meta-)prompt engineering, learning to optimize in semantic space, synthetic data flows, human–computer interactions, evaluation, interpretability, control, ethics of semantic decoding algorithms, as well as the infrastructure necessary to support such developments.

To help the reader, we provide a glossary of key terms in Sec. 6. Additionally, each section concludes with a concise summary highlighting the key points for easy reference.

2 From Syntactic to Semantic Tokens

At its core, computation is a syntactic process, where fundamental information building blocks are manipulated. These building blocks might take the form of binary digits (bits), abstract symbols in some computational models, or tokens in language models. As Claude Shannon himself noted, computation is syntactic because the symbols lack inherent *meaning* (Shannon, 1948); instead, meaning arises externally through context, via the processes that manipulate them, and the outcomes they produce for external actors.

Since Shannon, many have tried to lift the theory of information processing from the syntactic to the semantic level. The prevailing idea is to shift to computational models wherein basic symbols inherently carry semantic content, readily understandable, with meaningful impact on the actors outside the computation (Bao et al., 2011; Carnap and Bar-Hillel, 1954; Peyrard, 2019). This idea is illustrated by the concept of *semantic units* in the semantic theory of information (Zhong, 2017; Floridi, 2009), or the proposal of a *language of thoughts* (Fodor, 1975), postulating that thinking operates on atomic units of meaningful content (Rescorla, 2023).

The shift discussed in this article is of a similar kind, moving from language models processing syntactic tokens to interactions between LLMs and tools whose basic units of computation are *semantic tokens*, i.e., concepts intelligible for users outside of the computation. These semantic tokens are not abstract symbols awaiting to be interpreted, but active carriers of meaning, embodying concepts and ideas directly. We now describe formally syntactic and semantic tokens.

2.1 Syntactic Tokens

Syntactic tokens serve as the fundamental computational building blocks in modern natural language processing systems. The finite collection of all possible tokens forms a *syntactic vocabulary* Σ . These syntactic tokens are designed to be assembled into sequences through concatenation. Let $\mathbf{x} \in \Sigma^*$ represent one such sequence, defined as $\mathbf{x} := \langle x_0, \dots, x_m \rangle$. Typically, syntactic tokens may consist of words or characters, but more commonly, they are sub-word units. The set of these units is often learned based on the frequency of character co-occurrences in the available training data with methods such as byte-pair encoding (BPE) (Gage, 1994; Sennrich et al., 2016). The algorithm that breaks down natural language into a set of tokens and therefore defines the vocabulary Σ is called the *tokenizer*.

2.2 Syntactic Token Processors: Language Models

Syntactic tokens are symbols, and language models are computational processes that manipulate them. In particular, a probabilistic language model (PLM) induces a probability distribution P over all possible strings, Σ^* , that can be constructed from the vocabulary of syntactic tokens Σ . The purpose of the language model is to read an input sequence \mathbf{x} of tokens and guide the assembling of an output sequence \mathbf{y} .

To efficiently represent a probability distribution over the large combinatorial space of all possible strings, language models use an auto-regressive decomposition, meaning that they model the probability of each subsequent token given a sequence of previous tokens: $P(y_t | \mathbf{y}_{<t})$, where $\mathbf{y}_{<t} := \langle y_0, \dots, y_{t-1} \rangle$. Most modern language models are parameterized by a Transformer architecture with trainable weights θ (Vaswani et al., 2017). Then, the probability distribution of an output sequence $\mathbf{y} := \langle y_0, \dots, y_n \rangle$, potentially conditioned on an input $\mathbf{x} := \langle x_0, \dots, x_m \rangle$, is given by

$$P(\mathbf{y} | \mathbf{x}) = \prod_{t=0}^{|\mathbf{y}|} p_{\theta}(y_t | \mathbf{y}_{<t}, \mathbf{x}). \quad (1)$$

Importantly, the language model alone does not directly tell us how to produce an output sequence; it only specifies a probability distribution over the next token given a partial sequence. As described in Sec. 3, the combination with a decoding algorithm is necessary to transform a language model into a system that can produce output sequences.

2.3 Semantic Tokens

We define a *semantic token*, also referred to as a *thought*, as a sequence of syntactic tokens that conveys *meaningful information*. A semantic token, denoted as $\mathbf{x}^{\sigma} \in \Gamma$, is an element of *semantic vocabulary* Γ , a subset of Σ^* , representing the potentially infinite set of semantic tokens. It is important to note that not all syntactically valid strings convey meaningful information; thus, $\Gamma \neq \Sigma^*$. To differentiate semantic tokens (e.g., \mathbf{x}^{σ}) from arbitrary strings (e.g., \mathbf{x}), we use the superscript σ .

What exactly defines a semantic token? When can we determine that a string is “semantically meaningful”? Semantic tokens are embedded within natural language, drawing their meaning directly from potential human interpretation. Similar to how words derive meaning through contextual usage within a language, observed by external actors, the meaning of a string is shaped by its interaction with other elements in the system and its relationship with other semantic tokens. Finally, they also carry *pragmatic meaning* due to the effects they induce on the computation. For instance, both the input and output sequences of an AI system are semantic tokens, with meaning assigned through their usage. The input sequence \mathbf{x}^{σ} is *intended to* prompt a specific response or behavior from the AI system, while the output sequence \mathbf{y}^{σ} is *interpreted as* a response to a query, serving a particular purpose. Semantic tokens extend beyond input and output sequences; they encompass any other text conveying meaningful information. This includes thoughts, as defined and utilized in various frameworks such as chain-of-thoughts (Wei et al., 2022), tree-of-thought (Yao et al., 2023a; Long, 2023; Xie et al., 2023), and graph-of-thoughts (Besta et al., 2023; Yao et al., 2023b).

2.4 Semantic Token Processors

Probabilistic semantic token model. A probabilistic language model is trivially also a *probabilistic semantic token model*. Due to the auto-regressive decomposition, the language model induces a probability over strings and therefore over thoughts: $p_{\theta}(\mathbf{y}^{\sigma} | \mathbf{x}^{\sigma})$. In general, if there is a history of previously generated semantic tokens $H = [\mathbf{x}_0^{\sigma}, \dots, \mathbf{x}_m^{\sigma}]$, then the language model induces a probability distribution on the next semantic token:

$$P(\mathbf{y}^{\sigma} | H) = P(\mathbf{y}^{\sigma} | \mathbf{x}_0^{\sigma}, \dots, \mathbf{x}_m^{\sigma}), \quad (2)$$

which is, under the hood, still decomposed via an auto-regressive model on syntactic tokens:

$$P(\mathbf{y}^{\sigma} | H) = \prod_{t=0}^{|\mathbf{y}^{\sigma}|} p_{\theta}(y_t | \mathbf{y}_{<t}^{\sigma}, \mathbf{x}_0^{\sigma}, \dots, \mathbf{x}_m^{\sigma}). \quad (3)$$

In fact, this property makes the perspective shift from syntactic to semantic token possible, the language model being the enabler of the shift.

From language models to semantic processors. A *semantic token processor*, or *semantic processor* in short, is a system designed to take a semantic token \mathbf{x}^{σ} as input and generate a corresponding output semantic token \mathbf{y}^{σ} . While numerous systems fall under this category, let us focus on a semantic processor based on a language model. The

language model, by itself, provides only a distribution over the next syntactic tokens. However, to generate a sequence of syntactic tokens, it needs to be paired with a *decoding algorithm*. Essentially, a decoding algorithm is responsible for utilizing the distribution of the next tokens and determining which syntactic tokens to assemble into the output string. In practice, language models are commonly paired with decoding algorithms like top-k (Fan et al., 2018), top-p (Holtzman et al., 2020), often in combination with beam search. Various decoding heuristics and algorithms have been extensively explored to address the challenges inherent in the auto-regressive nature of language models. The decoding problem is the primary focus of Sec. 3.

It is important to note that employing the same language model with different decoding algorithms yields distinct outputs, thereby constituting different semantic processors. Moreover, other factors such as the prompting scheme can further differentiate semantic processors. For example, the chain-of-thought (Wei et al., 2022) or least-to-most prompting (Zhou et al., 2023) schemes generally result in different outputs, even when applied to the same LLM, thus defining different semantic processors.

Other semantic processors. While syntactic tokens exist for technical purposes, to support the practical computation of language models, semantic tokens are ubiquitous as they are the preferred units of human thinking. Semantic tokens are manipulated by a great variety of semantic processors, with humans being the primary example. Additionally, the tools humans build serve a purpose: to transform a meaningful input into a meaningful and useful output. Naturally, their inputs and outputs correspond to semantic tokens. Humans leverage tools like search engines, code executors, databases, APIs, etc., directly on a daily basis. There are substantial research efforts focused on developing methods for augmenting LLMs with such tools (Mialon et al., 2023a). This perspective holds significance as AI systems, like any other tool, can now participate in a rich ecosystem of semantic processors that communicate via the exchange of semantic tokens.

Summary of the shift from syntactic to semantic tokens:

At the syntactic level, **syntactic tokens** are determined by the tokenizer and act as the basic symbols manipulated by language models, the **syntactic token processors**.

At the semantic level, **semantic tokens**, or **thoughts**, are meaningful units of information. **Semantic processors** are the processes that manipulate these semantic tokens. When combined with a decoding algorithm, language models become semantic processors and join the rich ecosystem of semantic processors that includes humans and tools.

3 Decoding: Extracting Utility from Token Processors

In the previous section, we introduce a more general definition of a token and the concept of a token processor – the components that make up modern AI systems. Now, our focus shifts to the algorithm layer responsible for orchestrating these components to solve practical tasks. We refer to this layer as the *decoding algorithm*. We draw an analogy between decoding operating on syntactic tokens and decoding operating on semantic tokens. Fig. 2 visually depicts this analogy.

3.1 General Formulation of the Decoding Problem

The objective: maximizing utility. To solve a given task, an AI system processes an input semantic token (a query) aiming to produce optimal output. This is where the concept of a *utility function*, denoted as $u_t(\mathbf{y}^\sigma | \mathbf{x}^\sigma)$, becomes critical. It scores candidate semantic token outputs, assessing how effectively they solve the task for the specific input \mathbf{x}^σ . For instance, in machine translation, the utility function might judge how well the translation \mathbf{y}^σ retains the original meaning conveyed by \mathbf{x}^σ . In an ideal scenario, when presented with input \mathbf{x}^σ , the system selects the output with the highest utility score: $\text{argmax}_{\mathbf{y}^\sigma \in \Gamma} u_t(\mathbf{y}^\sigma | \mathbf{x}^\sigma)$.

In practice, the AI system does not have access to the utility function during inference. However, it can rely on value models, which estimate the utility function for partial outputs, to guide the decoding process effectively.

The decoding problem. AI systems construct their answers by manipulating tokens, the relevant units of computation, using specialized computational tools, the token processors. A dedicated decoding algorithm orchestrates the execution of these basic components to solve the given task, taking into account the properties, capabilities, and limitations of the token processors and organizing the computation to robustly extract high-utility outputs. At the syntactic level, the syntactic decoding algorithm utilizes syntactic token processors (language models) to manipulate syntactic tokens. By analogy, we propose to conceptualize the orchestrated collaboration of AI, humans, and tools as the semantic equivalent,

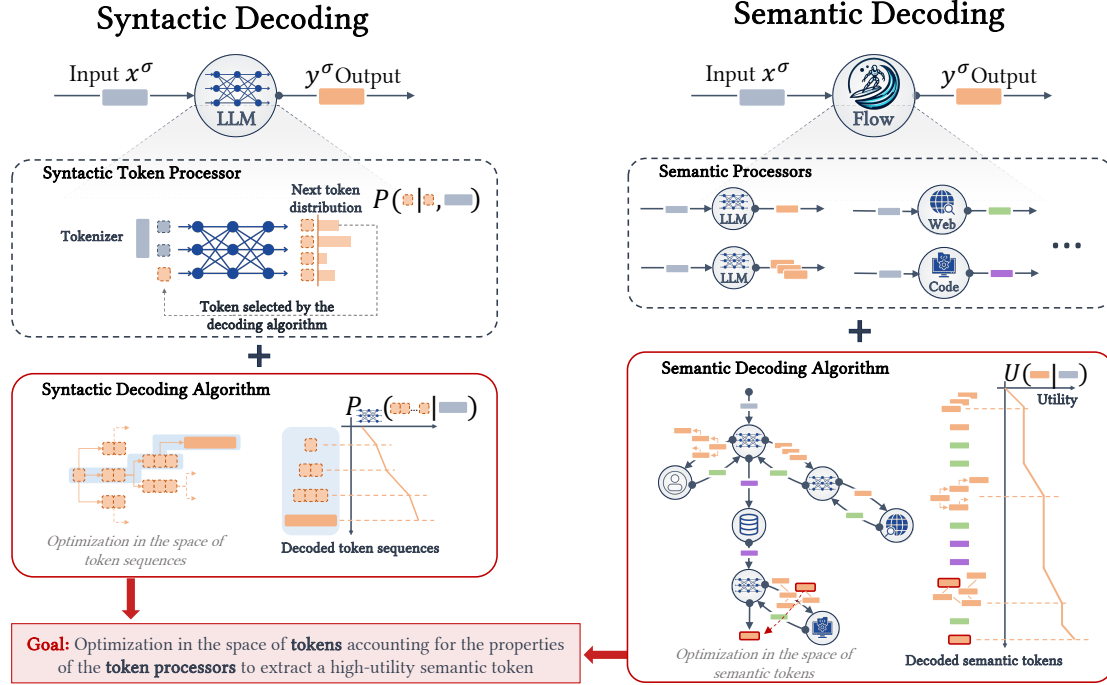


Figure 2: **Illustrating the analogy between syntactic and semantic decoding perspectives** On the left side, we depict the conventional (syntactic) decoding process of generating sequences of tokens from an auto-regressive language model. The decoding algorithm strategically uses the language model to produce a high-utility output sequence. Similarly, we frame recent progress in AI, human, and tool collaboration as the semantic-level analogy of this process. Here, the computational units are referred to as *semantic processors*, which manipulate semantic tokens, representing semantically coherent pieces of text. The semantic decoding algorithm harnesses the capabilities of the semantic processors, orchestrating their computation through semantic token exchange to extract a high-utility output. Both perspectives share a common objective: extracting high-utility output by leveraging the token processors available during inference.

wherein the semantic decoding algorithm utilizes semantic processors to manipulate semantic tokens. Both algorithms perform optimization within the token space with the objective of extracting high-utility semantic tokens.

3.2 Syntactic Decoding

The goal of syntactic decoding algorithms is to retrieve an element maximizing a given utility function from the space of the possible token sequences. As a primary source of signal, syntactic decoding algorithms rely on the probability distribution induced by a language model and aim to retrieve a sequence that maximizes the model’s likelihood. However, this approach faces two main challenges. Firstly, the optimization problem $\operatorname{argmax}_{y \in \mathcal{Y}} p\theta(y; |x)$ becomes intractable due to the exponentially large state space, necessitating approximation techniques. The commonly employed strategies to tackle this difficulty rely on greedy heuristics, such as beam search (Sutskever et al., 2014), which focuses on the most probable tokens, either deterministically selecting the top-k candidates or sampling from the top-k or top-p tokens from the distribution (Fan et al., 2018). Second, outputs associated with high likelihood are not necessarily of high utility, and effective decoding algorithms need to account for the potential misalignment (Josifoski et al., 2023b). Methods to address this problem include (i) value-guided beam search, which uses a greedy strategy similar to beam search but selects the next token using a linear combination of the model’s likelihood and the scores from a value model; and (ii) Monte Carlo tree search (MCTS) (Chaffin et al., 2022), which allocates a fixed computation budget for an informed exploration of multiple paths in the decoding tree before token selection. See Josifoski et al. (2023b) for a more comprehensive overview of prior work on syntactic decoding from an optimization perspective.

3.3 Semantic Decoding

We now discuss the proposal of *semantic decoding*. In this perspective, the objective remains the same as in syntactic decoding: solving a task by generating a high-utility semantic token. However, the fundamental unit of computation shifts from syntactic tokens to semantic ones, and the basic computational processes shift from syntactic processors to semantic ones.

A semantic decoding algorithm coordinates the exchange of semantic tokens among semantic processors to navigate through the semantic token space and identify a high-utility trajectory in semantic space.

Chain-of-thought (CoT) is a fundamental example, which generates a sequence of thoughts (semantic tokens) before arriving at an answer. In the initial version, the semantic tokens in the chain do not interact with any other semantic processors or algorithmic components. However, numerous variants have quickly emerged, including: (i) Sampling and combining multiple reasoning chains (Wang et al., 2023a), (ii) Incorporating feedback from other models, symbolic tools, or human inputs, (Wang et al., 2023c,b; Gao et al., 2023; Paul et al., 2023; Josifoski et al., 2023a), (iii) Exploring non-linear trajectories in semantic space, such as trees (Yao et al., 2023a; Long, 2023; Xie et al., 2023) or graphs (Besta et al., 2023; Yao et al., 2023b), and (iv) Utilizing evolutionary algorithms to select promising semantic tokens, e.g., FunSearch (Romera-Paredes et al., 2023) or PromptBreeder (Fernando et al., 2023).

The syntactic decoding setup often operates under the constraint to produce the full trajectory of syntactic tokens as its output. In contrast, semantic decoding has a more flexible ability to manipulate its trajectories and select only a subset of the trajectory as the final output, often selecting its last semantic token as the output. The constraint to output the entire trajectory makes every decision critical for the quality of the answer. This also renders value estimation more challenging because the value model has to estimate the expected utility for partial outputs that may not look like a candidate’s answer. However, at the semantic level, the semantic decoding algorithm has the flexibility to backtrack and remove parts of the trajectory. It’s worth noting that, in principle, nothing prohibits the syntactic decoding algorithm from producing subsets of the trajectory as the output. Recent examples include the usage of pause tokens, which are automatically removed from the final output but used during decoding to allow the model more computational steps.

In Sec. 4, we explore how adopting the semantic decoding perspective, which focuses on the optimization and search performed in the semantic space, enables us to broaden our understanding of what is possible for orchestrated interactions.

3.4 Connections with Pragmatics and Semiotics

Pragmatics, a subfield of linguistics, focuses on the context-dependent aspects of meaning and how language is used to achieve specific goals (Birner, 2012). Semantic decoding can then be seen as performing pragmatic computation as it relies on the goal-oriented usage of language. Pragmatic computing is an optimization process in semantic space happening via the orchestrated exchange of semantic tokens. The interaction between semantic processors reflects the pragmatic idea that meaning is constructed through the dynamic exchange of information between goal-driven participants in a conversation.

The semantic decoding perspective also intersects with semiotics (Chandler, 2022), the study of signs, symbols, and their interpretation. By bridging syntactic and semantic tokens, language models bring meaningless symbols into a space where semantic decoding algorithms can interpret and manipulate these concepts based on their pragmatic usage from various other semantic processors.

By focusing on the optimization of utility through the interaction of semantic processors, the semantic decoding perspective encompasses both the semiotic notion of meaning emergence and the pragmatic emphasis on the context-dependent usage of language in achieving specific goals.

3.5 Flows as Semantic Decoding Algorithms

The semantic decoding perspective we propose emerges from *Flows* (Josifoski et al., 2023a), an abstraction for modeling structured interactions among AI systems, tools, and humans. This abstraction revolves around *Flows* as compositional, self-contained, goal-driven entities that execute a semantically meaningful unit of work and communicate solely through semantic tokens. Flows represent semantically meaningful blocks of computation performed either by directly utilizing a “tool,” as in *Atomic Flows*, or emerges from purposeful interactions among other Flows, as in *Composite Flows*. Crucially, in the *Flows* abstraction, the notion of a *tool* is general and encompasses everything from a database, a compiler, or a search engine to powerful AI systems like LLaMA (Touvron et al., 2023), Stable Diffusion (Rombach et al., 2021), and GPT-4 (OpenAI, 2023); or a human. Fundamentally, the concept of an Atomic Flow lifts all tools into

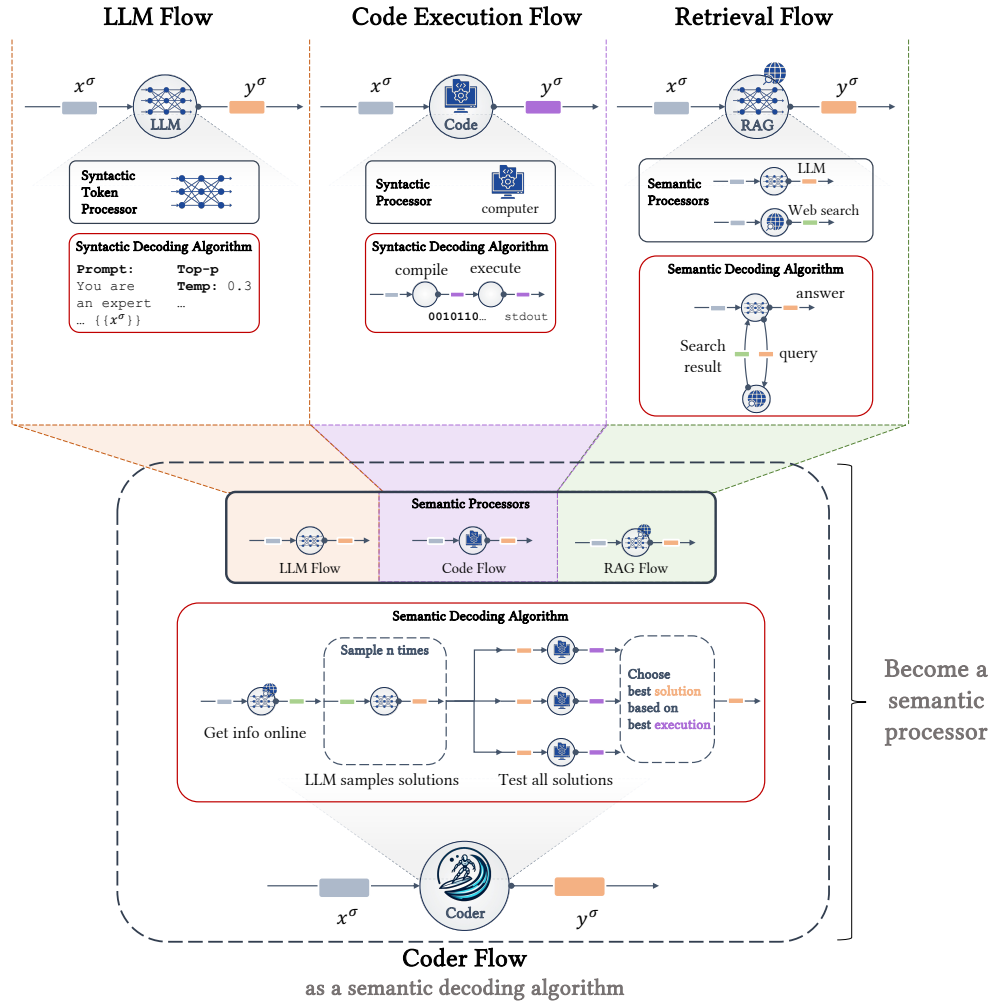


Figure 3: **Illustrating compositionality of Flows.** On the top, we see three semantic processors. The first one is implemented using a prompted language model, the second is a code executor wrapped in a Flow to enable communication, and the last one is a semantic decoding algorithm itself (Retrieval Augmented Generator), orchestrating both an LLM and a search engine. Together, these three semantic processors are orchestrated as part of a new Flow, producing a robust semantic decoding algorithm designed to solve coding problems. In this example, the Retrieval Flow initially extracts useful information from the web. Subsequently, the LLM Flow samples many candidate solutions, each of which is executed and tested by the code execution Flow. Finally, the answer to be returned is selected. This programmer Flow can then serve as a semantic processor for others to utilize. Flows, implementations of semantic decoding algorithms, become semantic processors themselves as they read and generate semantic tokens. They can then join the pool of semantic processors available for other Flows to utilize. This enables an open-ended complexity growth, with modular building blocks seamlessly interfaced through communication via semantic tokens. The Flows abstraction underscores that the distinction between semantic processors and semantic decoding algorithms is only in perspective.

a single semantic space shared with Composite Flows, in which they can all communicate via semantic tokens (i.e., messages).

Flows *are* semantic decoding algorithms as they orchestrate semantic processors to produce a useful semantic token as output. The semantic decoding perspective is concerned with the optimization that Flows performs in the semantic space (the *what*). Whereas the framework of *Flows* is the conceptual tool enabling the design and implementation of these compositional interactions (the *how*). Notably, a semantic decoding algorithm – implemented by a Flow – becomes a black-box engine that reads a semantic token as input and produces a semantic token as output, therefore becoming a semantic processor itself for other Flows to utilize. This compositional property of Flows enables open-ended complexity growth by constantly increasing the pool of existing semantic processors, blurring the frontier between

a semantic decoding algorithm and a semantic processor. A Flow can be both a semantic processor within a larger computation or a semantic decoding algorithm when it is itself the outermost algorithmic layer under scrutiny for understanding what type of optimization it is performing to construct its output semantic token. Fig. 3 illustrates the compositional properties of *Flows* and its close connection to semantic decoding through an example. Throughout the rest of the paper, we use the terms Flow and semantic decoding algorithm interchangeably.

Beyond the benefits at the conceptual level, the principled abstraction of *Flows* and its accompanying library aiFlows* enable seamless asynchronous and distributed execution, which are necessary for many promising directions, as discussed in Sec. 4 and Sec. 5. Furthermore, the library comes with *FlowVerse*, a repository of Flows that can be readily used, extended, or composed into more complex Flows. *FlowVerse* is open for anyone to contribute.

Summary of the decoding perspective:

A **decoding algorithm** is an algorithmic layer on top of token processors that orchestrates the computation over tokens to extract a high-utility output. A **semantic decoding algorithm**, in turn, is a decoding algorithm orchestrating semantic processors manipulating semantic tokens. It represents the semantic counterpart to the well-known problem of syntactic decoding from auto-regressive language models.

A semantic decoding algorithm is itself a semantic processor, ready to be employed by other, more complex, semantic decoding algorithms. This compositional property, emerging at the semantic level, enables **open-ended complexity growth**. The compositional *Flows* framework capitalizes on this property to enable and facilitate the modeling, implementation, and systematic study of arbitrarily complex structured interactions between tools, models, and humans. Then, certain equivalences come to light: $Flow \cong semantic\ decoding\ algorithm \cong semantic\ processor$. *Flows* offers a playground for engineering semantic decoding algorithms.

4 Semantic Decoding: Optimization in the Semantic Space

By adopting the semantic decoding perspective, we can analyze the orchestrated interactions of AI, tools, and humans based on the optimization processes they perform and the heuristics they use to navigate the semantic space. When compared to the syntactic level, the semantic level offers distinct advantages that semantic decoding algorithms can leverage. These advantages include:

Effectiveness in exploration: The structure of the semantic space induced by the semantic processors inherits the intrinsic meaning associated with the tools we build and provides us value. Thus, by definition, the semantic space comes with a natural structure that makes it more meaningful for the optimization of utility. Syntactic decoding, on the other hand, consistently encounters the core issue that a semantic concept can materialize in vastly different syntactic forms, making optimization in the syntactic space inefficient and disconnected from the semantic space. Furthermore, unaffected by the auto-regressive nature of syntactic decoding, semantic decoding algorithms can explore the output space in a non-linear fashion. This flexibility and structure create the potential for dramatically more effective decoding algorithms.

Human interpretability: Semantic tokens inherently carry meaning, allowing humans to fully understand the computations and optimization performed by the semantic decoding algorithm. As semantic processors themselves, humans can seamlessly engage in these algorithms' computations.

Open-endedness: Semantic decoding algorithms operate over semantic tokens. Reading and generating semantic tokens, like other semantic processors, makes them a valid semantic processor. With the internal optimization abstracted away, semantic decoding algorithms can be readily interfaced with the ever-growing pool of semantic processors, including other decoding algorithms. This compositional property is critical for enabling the design and implementation of processes with open-ended complexity.

The design space for semantic decoding algorithms is vast, and we're just scratching the surface of its full potential. As a community, we still haven't properly formalized the problem and lack methods for systematically discovering, crafting, and learning semantic decoding algorithms. In Sec. 5, we explore opportunities for future research. In this section, we categorize the types of search and optimization that can be performed in the semantic space, drawing connections to concepts from the world of syntactic decoding.

*<https://github.com/epfl-dlab/aiflows/>

4.1 Heuristic Decoding Patterns: Grammars of Thoughts

A simple strategy for navigating the semantic space is to rely on predefined workflows dictating which types of semantic tokens should be generated at which moments. This programmed exchange of semantic tokens among semantic processors is designed to ensure a robust progression towards a high-utility final output. This line of research has been initiated by the Chain-of-Thought (CoT) method (Wei et al., 2022), a prompting method enforcing the generation of intermediate semantic tokens before producing a final output.

Several works have studied more sophisticated reasoning patterns, incorporating additional semantic processors. For example, ReAct employs a two-step approach of switching between a reasoning and an acting step. This allows the model to leverage external tools to generate intermediary semantic tokens. Some research splits the problem-solving process into two phases implemented by dedicated planning and acting Flows (Wang et al., 2023c,b; Gao et al., 2023; Paul et al., 2023; Josifoski et al., 2023a), or leverage feedback-giving Flows (Weng et al., 2023; Paul et al., 2023; Josifoski et al., 2023a). Exploring and systematically comparing diverse patterns across tasks is necessary for identifying the general principles dictating which patterns are beneficial for different models and problems that are still lacking (Josifoski et al., 2023a).

An interesting analogy can be drawn with the concept of constrained decoding in the syntactic decoding literature. Constrained decoding refers to methods enforcing constraints during inference time (Tromble and Eisner, 2006; Geng et al., 2023). For instance, one may want to prevent repeated n-grams (Stahlberg and Byrne, 2019; Anderson et al., 2017), or enforce predefined structural constraints on the output (Hokamp and Liu, 2017; Hu et al., 2019; Post and Vilar, 2018; Josifoski et al., 2022; Deutsch et al., 2019; Shin et al., 2021), expressible with formal grammars (Scholask et al., 2021; Roy et al., 2022; Geng et al., 2023).

Then, it is worth contemplating what the semantic equivalent of a grammar constraint would be. Such a *grammar of thoughts* would represent formal constraints on trajectories’ structure in the semantic space, akin to generalized reasoning patterns. Planning before acting, validating with external tools before answering, and attempting to refute one’s own answer would be simple examples of such grammar.

4.2 Meta-Heuristics: Sampling and Value-Guided Search in Semantic Space

At the syntactic level, various decoding strategies exploit the probabilistic nature of language models, including top-p (Holtzman et al., 2020), top-k (Fan et al., 2018), and stochastic beams (Kool et al., 2019; Meister et al., 2021), using sampling of tokens to better explore the space of possible output sequences.

Given that language models are also probabilistic generators of semantic tokens, exploration of the semantic space through sampling naturally arises. Many existing semantic decoding algorithms are built on the foundation of sampling semantic tokens. For instance, self-consistency (Wang et al., 2023a) extends CoT by sampling multiple chains, thereby exploring different reasoning paths. This concept is further expanded to non-linear chains with approaches like Tree-of-Thought (Yao et al., 2023a; Long, 2023; Xie et al., 2023) and Graph-of-Thoughts (Besta et al., 2023; Yao et al., 2023b).

To ensure navigation towards high-utility regions of the output space, complementing sampling with an external signal, such as a value model – an estimator of the final utility given the current trajectory – can often be beneficial. At the syntactic level, this idea is embodied by Value-Guided Beam Search (VGBS) (He et al., 2017; Ren et al., 2017; Krishna et al., 2022), a variant of beam search utilizing a value model in addition to the model’s likelihood to decide which next token to sample. Monte-Carlo tree search (MCTS) (Chaffin et al., 2022; Josifoski et al., 2023b) extends this idea by performing simulations to explore the utility of future outcomes before making the next decision.

At the semantic level, the idea of a value model can be implemented very flexibly by leveraging the rich ecosystem of semantic processors. These processors offer extensive possibilities for incorporating external signals in the computation. Language models prompted to perform different roles or with different side information can provide feedback by estimating the value of a current stream of semantic tokens (Xue et al., 2023; Shinn et al., 2023; Paul et al., 2023). Tools or humans with access to external information, acting in the world, or executing code can further provide reliable grounded estimates of utility (Gao et al., 2023; Chen et al., 2023c; Josifoski et al., 2023a; Li et al., 2023a). Already, Ding et al. (2023) hint at the idea of performing MCTS-based planning in semantic space.

Semantic decoding algorithms like FunSearch (Romera-Paredes et al., 2023) and PromptBreeder (Fernando et al., 2023) utilize a language model as the sampler of candidate solutions (semantic tokens) within evolutionary algorithms optimizing a population of candidate solutions. The fitness function can be viewed as a simple value model guiding the decoding of the next semantic tokens, i.e., the next candidate solutions. These examples showcase the effectiveness of combining optimization through search based on sampling from a competent sampler, i.e., a language model.

Currently in its early stages, this research direction lacks proper formalization within the optimization domain. Nevertheless, it shows great potential to enable the emergence of new types of AI capabilities, as evidenced by systems such as FunSearch (Romera-Paredes et al., 2023).

4.3 Learning the Flow: Embracing Optimization in Semantic Space

Moving beyond heuristics and meta-heuristics, one could fully commit to the optimization perspective within the semantic space. At a micro level, optimization algorithms could focus on enhancing performance by training the semantic processors to collaborate more effectively. We term this approach as *learning to collaborate*. Alternatively, taking a more holistic view of the problem, optimization might involve training a controller that determines which semantic processor to invoke at each time step along with the appropriate parameters. We refer to this approach as *learning to orchestrate*. For example, one could utilize reinforcement learning (Bahdanau et al., 2016) or reward-based supervised learning (Norouzi et al., 2016; Rafailov et al., 2023; Gülçehre et al., 2023) to backpropagate a learning signal through semantic tokens, optimizing over sequences of syntactic tokens directly. Systems such as AutoGPT (Richards, 2023) or BabyAGI (Nakajima, 2023) already use a prompted language model as a heuristically optimized controller of a general-purpose semantic decoding algorithm. Moreover, learning to orchestrate can be seen as a planning problem, and therefore, ideas from the extensive literature on planning could be explored in this context.

This general notion of learning the semantic decoding algorithm builds upon previous work at the syntactic level. For instance, Wiseman and Rush (2016) introduced a differentiable relaxation of beam-search, and Collobert et al. (2019) developed a fully differentiable beam-search that can be optimized during training.

Summary of semantic decoding optimization:

Decoding directly in the semantic space offers many benefits, including flexible and effective exploration of the meaningfully structured semantic space, human-interpretable computation, and open-endedness enabled by the compositional nature of Flows.

The optimization performed in the semantic space can be broadly categorized into three main types: (i) **Heuristic decoding patterns**: programmed interactions such as CoT, ReAct, or meta-patterns like planning before acting and iterative, feedback-based refinement. (ii) **Sampling and value-guided search**: interactions defining optimization strategies that explore the semantic space by strategically sampling semantic tokens and leveraging a value function to guide the process. (iii) **Learning to optimize in the semantic space**: *learning to decode* by training the semantic decoding algorithms and their components. For instance, learning to collaborate, orchestrate, search, learn effective reasoning patterns, and more.

5 Research and Application Opportunities

In the previous section, we highlighted the distinctive benefits of the semantic decoding perspective and outlined what kind of optimization and search can be performed in semantic space. In this section, we shift towards applications and research questions opened up by the semantic decoding perspective. While not exhaustive, this section serves to illustrate the depth and potential of adopting the semantic decoding viewpoint.

5.1 Prompt Engineering and Meta-Prompt Engineering

The ability of LLMs to adapt to the information in their context has led to remarkable breakthroughs across fields (Brown et al., 2020; Kojima et al., 2022), with prompt engineering playing a central role in these advancements. For example, compared to traditional prompting methods, CoT (Wei et al., 2022) has demonstrated a threefold increase in performance on the GSM8K dataset. Likewise, for competitive coding, a fixed collaboration pattern between two GPT-4 instances – a coder and a critic with access to a code executor – increases the solve rate by 1.8 times (Josifoski et al., 2023a). These represent just a glimpse of the many examples where LLM performance has been significantly improved without fine-tuning. Researchers continue to push boundaries by crafting complex patterns to tackle increasingly complex tasks (Romera-Paredes et al., 2023) and exploring meta-prompting methods (Fernando et al., 2023). Viewing (meta-)prompt engineering as optimization in semantic space opens the door to novel ideas and provides a principled structure for the research efforts in this particularly active and promising direction.

5.2 Synthetic Data Flow

Data, itself composed of semantic tokens, can become the focal point for Flows dedicated to manipulating or synthesizing data. Similarly, specialized trainer Flows can be aimed at training semantic processors (e.g., models) or even entire (sub-)Flows. By combining trainer Flows with synthetic data generation Flows, opportunities emerge for creating sophisticated self-training loops. The synthetic data generation Flows can leverage domain knowledge (Tang et al., 2023), task properties (Lu et al., 2024; Veselovsky et al., 2023; Josifoski et al., 2023c), or collaboration (Abdullin et al., 2024), and synthesize data of notably higher quality than what a single model or simple heuristics can achieve. This sets the stage for effective self-improvement loops where a language model participates in a semantic decoding algorithm producing high-quality synthetic data. Then, the language model improves itself through fine-tuning, thereby improving the Flow’s capacity to generate even better synthetic data in a virtuous cycle (Silver et al., 2017; Burns et al., 2023; Singh et al., 2023; Chen et al., 2024b). An example of such a Flow is MAGDi (Chen et al., 2024a), a framework designed to distill reasoning interactions among multiple LLMs into smaller ones. This approach surpasses single-teacher distillation (Li et al., 2023c; Magister et al., 2023) and finetuning based on reasoning trajectories sampled from GPT-4 (Chen et al., 2023a).

5.3 Human in the Loop and Human-Computer Interaction

Humans operate within the semantic space and can seamlessly integrate into semantic decoding algorithms as just another type of semantic processor. This integration creates numerous opportunities to explore diverse human-computer interactions within a principled optimization-based framework. This framework can leverage human cognition as a computational input for semantic decoding algorithms aimed at maximizing a utility function. For instance, humans can provide nuanced low-level feedback, offer high-level guidance, or simply observe the ongoing exchange of semantic tokens (Josifoski et al., 2023a; Cai et al., 2023; Xiao and Wang, 2023; Kim et al., 2024; Li et al., 2023d). Moreover, humans can dynamically switch roles during the execution, intervening, pausing the process, and resuming as necessary. Flow engineering methods, such as searching for heuristic reasoning patterns (Sec. 4.1), crafting guided search methods (Sec. 4.2), or learning the Flow (Sec. 4.3), can be designed to optimize objectives that involve maximizing a utility function while minimizing the cognitive cost for the human in the loop (Horvitz, 1999).

5.4 General AI Assistants

Considerable research efforts are currently focused on constructing general-purpose assistants. Existing methodologies involve utilizing LLMs as controllers, together with selected tools to enhance *cognitive* abilities (Nakajima, 2023; Richards, 2023; Wang et al., 2023b; Wu et al., 2024). However, recent evaluations on the GAIA benchmark (Mialon et al., 2023b) have revealed a notable performance gap compared to humans, highlighting the need for improvements. For example, GPT-4, when equipped with tools, achieves only 15% accuracy, while humans attain 92% accuracy. Examining this challenge through the perspective of semantic decoding, it can be framed as the development or learning of a general-purpose semantic decoding algorithm.

5.5 Evaluation and Diagnostic

Evaluating semantic decoding algorithms presents a significant challenge because they evade standard controlled benchmarks for two primary reasons. Firstly, the issue of data contamination arises because language models are trained on vast amounts of internet data, potentially including benchmark samples. To address this issue, research is needed to untangle the impact of memorization and generalization across various contexts (Josifoski et al., 2023a; Shi et al., 2023; Golchin and Surdeanu, 2024).

Secondly, the dynamic nature of the environment leads to the evolution of AI systems over time. For instance, a search engine’s results can vary not only due to algorithmic changes but also due to updates in the real world. Overcoming this challenge necessitates advancing diachronic evaluation methodologies specifically tailored to assess dynamic AI systems in evolving environments, as exemplified by dynamically evolving benchmarks (Li et al., 2023e).

Despite these evaluation challenges, it remains imperative to develop methods that effectively discern between effective and ineffective semantic decoding algorithms. This understanding is crucial for informing users about the expected behavior of AI systems and enabling practitioners to enhance system performance. Tools from causality, such as causal mediation analysis, can be particularly useful in recognizing which components or messages were critical to the computation’s success or failure. These diagnostic considerations are tightly linked to the interpretability challenges discussed below.

5.6 Interpretability, Control, and Error Mitigation

Interpretability. A semantic decoding algorithm, as any AI system, is subject to questions of explainability to understand *why* some outputs were produced (Woodward, 2003; Potochnik, 2017; Lipton, 2018). One approach is to perform behavioral analyses by manipulating inputs and observing the resulting effects on the outputs. Additionally, model-agnostic feature importance methods, such as LIME (Ribeiro et al., 2016) or SHAP (Yeh et al., 2020), can be expanded to include semantic tokens rather than just syntactic features. Opening the black box to inspect the network of semantic tokens exchange leads the path toward mechanistic interpretability of semantic decoding algorithms. This involves manipulating intermediate semantic tokens, placing the system in a counterfactual state, and then resuming the computation to precisely measure the impact of each computational step (Pearl and Mackenzie, 2018; Geiger et al., 2022). In general, interpreting semantic decoding algorithms is somewhat simpler than language models because semantic tokens serve as discrete, semantically meaningful bottlenecks in the computation that can be easily inspected and intervened upon.

Control and ethics. Closely linked to the issue of explainability is the notion of control (Woodward, 2003; Pearl and Mackenzie, 2018). By breaking down computation into discrete, human-interpretable chunks, we obtain many levers for controlling the computation. This can be done manually by humans or automatically by appending dedicated semantic processors to correct, adjust, or prohibit some predefined undesirable intermediate steps, thus preventing undesired outcomes. For example, specialized *ethics semantic processors* (possibly complex flows themselves) can be inserted at critical steps to exclude unwanted semantic tokens and steer the computation toward areas of the output space that align with predefined ethical objectives (Gallegos et al., 2024). Such ethics components might leverage expert debiasing methods that scrutinize semantic tokens for potential societal biases and either remove them by reformulating or providing informed feedback.

5.7 Richer Semantic Spaces

From syntactic to semantic and back. One advantage of semantic tokens is that they are readily understood by humans, rendering semantic decoding algorithms inspectable, interpretable, and conducive to human participation. Yet, the decoding perspective does not inherently require semantic tokens to be human-interpretable. One could imagine a collaboration between LLMs and tools that, for the sake of communication efficiency or effectiveness, relies on new made-up languages. These new semantic spaces can be learned as part of the *learning the flow* pipelines to optimize the amount of useful information exchanged per message.

Multimodal semantic tokens. Another way in which the concept of semantic token can be stretched is by considering other modalities. The semantic decoding perspective easily accommodates more general informational units (Reed et al., 2022). Developing more competent multimodal language models is an active area of research (Wu et al., 2023a). Multimodal models can easily participate in semantic decoding algorithms, significantly enlarging the semantic space and the variety of signals available to guide exploration toward high-utility outputs.

5.8 Infrastructure

In order to support the innovations discussed above, it is imperative to have robust infrastructures. Significant efforts have already been invested in creating efficient abstractions (Lu et al., 2023; Li et al., 2023b; Shen et al., 2023; Chase, 2022; Hong et al., 2023; Wu et al., 2023b). However, the *Flows* framework (Josifoski et al., 2023a) stands out by introducing the modularity and compositionality essential for systematically constructing intricate systems (Sec. 3). Its support for concurrent and distributed execution unlocks creative applications like FunSearch (Romera-Paredes et al., 2023), PromptBreeder (Fernando et al., 2023), meta-reasoning Flows (Josifoski et al., 2023a), and arbitrary peer-to-peer collaborations between Flows, which could define anything from a stateless tool to a general assistant or autonomous agent.

Beyond proper abstractions, *Flows* and semantic decoding algorithms enable a level of complexity and flexibility that comes with additional technical challenges. For instance, *Flows* supports open-ended complexity and allows practitioners to freely develop and publicly deploy Flows, thereby necessitating systematic and contextual Flow indexing and retrieval (i.e., a search engine over Flows). Additionally, executing Flows may depend on resource-intensive semantic processors, such as commercial LLMs, which necessitates efficiency optimizations to minimize resource usage or latency while maintaining utility.

Flow indexing and retrieval. Similar to how a web page provides information or access to some services, a Flow provides a meaningful computation. Given the ability to publicly deploy Flows and the infrastructure for peer-to-peer

communication between Flows, the need for a search engine over the space of Flows is going to grow as the number of publicly accessible Flows increases. Developing scalable yet effective methods for indexing Flows (i.e., semantic computation), akin to PageRank for web pages, is bound to become an important research direction in the near future.

Efficiency optimization. Improving the efficiency of a Flow can be achieved by enhancing the efficiency of its semantic processors. Thankfully, a rich body of work is dedicated to improving the efficiency of language models’ inference. Techniques like batching and key-value caching can mitigate the cost and latency of decoding long sequences autoregressively (Pope et al., 2022). Speculative decoding methods, which involve generating candidate samples using smaller models and then refining or filtering these samples with larger models, represent a different approach to optimization (Schuster et al., 2022; Chen et al., 2023b; Kim et al., 2023; Stern et al., 2018). Further optimizations include memory usage reduction through weight quantization (Yao et al., 2023c; Dettmers et al., 2022; Frantar et al., 2023; Dettmers et al., 2023; Xiao et al., 2023; Ashkboos et al., 2023), and leveraging non-homogeneous computational costs in the transformer graph, as exemplified by SkipDecode (Corro et al., 2023) and PowerInfer (Song et al., 2023).

At the semantic level, caching Flow calls with soft, approximate caches (Ramírez et al., 2023) and dynamically routing queries to different models based on their properties (Hu et al., 2023; Liu et al., 2023; Šakota et al., 2024; Yue et al., 2023; Lee et al., 2023) are feasible strategies. Concurrent execution of semantic processors, such as Skeleton-of-Thoughts (Ning et al., 2023), also contributes to latency gains. Despite these promising approaches, the concept of *speculative semantic decoding* remains relatively unexplored. Systematic studies could explore replacing expensive sub-flow components with faster, more economical modules trained to emulate or cache computations from costly sub-Flows.

6 Conclusion

This paper proposes semantic decoding, which formalizes LLMs, humans, and other tools as semantic processors that read and generate semantic tokens, and the collaborations between them as optimization processes in the semantic space. These optimization processes correspond to semantic decoding algorithms and aim to find high-utility semantic tokens as answers to queries. This perspective allows us to systematically study the design space of semantic decoding algorithms based on the optimization they perform.

The transition from syntactic to semantic tokens represents a pivotal moment in the development of AI systems. By abstracting the syntactic details, we can conceptualize LLMs, humans, and tools as a form of computation that operates directly within the space of meaningful concepts. The potential offered by orchestrated collaborations between them is vast, yet we are only scratching the surface of what is possible.

We believe that the concept of Flows (Josifoski et al., 2023a), supported by the aiFlows library[†], serves as a practical framework for engineering robust, modular, and compositional semantic decoding algorithms.

Acknowledgments

West’s lab is partly supported by grants from Swiss National Science Foundation (200021_185043, TMSGI2_211379), Swiss Data Science Center (P22_08), H2020 (952215), Microsoft Swiss Joint Research Center, and Google, and by generous gifts from Facebook, Google, and Microsoft.

[†]<https://github.com/epfl-dlab/aiflows>

Glossary of Important Terms

Syntactic Token Basic representational units of language, typically defined by the tokenizer in modern NLP systems.

Semantic Token Semantically coherent units of text, also known as *thoughts*, representing meaningful concepts or ideas.

Syntactic Processor Entities that process and manipulate syntactic tokens, such as parsers or language models.

Semantic Processor Entities that process and manipulate semantic tokens, including LLM-based systems, humans, and various tools such as search engines and code executors. Notably, a semantic processor can be implemented by a semantic decoding algorithm itself, showcasing the recursive compositional property that emerges at the semantic level.

Utility Function A function that scores candidate solutions, assessing their effectiveness in solving a task for a given input query. The goal of an AI system is to produce an output with high utility.

Value Model A model that estimates the expected utility of partial trajectories during a decoding process, guiding decoding strategies towards regions of expected high utility.

Decoding The process of extracting high-utility sequences or trajectories, which can be performed at both the syntactic and semantic levels.

Syntactic Decoding The process of extracting high-utility sequences of syntactic tokens, often relying on sampling tokens and guided by value models or heuristics. Examples include top-p, value-guided beam search, and Monte Carlo tree search decoding.

Semantic Decoding The process of extracting high-utility trajectories in semantic space, leveraging sampling, semantic-level value models, or heuristics to optimize the output. Examples include methods that orchestrate interactions between semantic processors, such as chain-of-thought, tree-of-thoughts, ReAct, and FunSearch. When ignoring computational details, a semantic decoding algorithm becomes a semantic processor, showcasing the recursive compositional property that emerges at the semantic level.

Flows An abstract model of communication between LLMs, humans, and tools that supports the implementation of modular and compositional semantic decoding algorithms. Flows *are* semantic decoding algorithms in that they specify a general framework to implement them. We use the term *semantic decoding algorithm* to emphasize the optimization aspect in semantic space, while we use the term *Flows* to focus on the engineering and implementation aspects.

References

- Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. [Synthetic dialogue dataset generation using llm agents](#).
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. [Guided open vocabulary image captioning with constrained beam search](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.
- Saleh Ashkboos, Ilia Markov, Elias Frantar, Tingxuan Zhong, Xincheng Wang, Jie Ren, Torsten Hoefler, and Dan Alistarh. 2023. [Quik: Towards end-to-end 4-bit inference on generative large language models](#).
- Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2016. [An actor-critic algorithm for sequence prediction](#). *CoRR*, abs/1607.07086.
- Jie Bao, Prithwish Basu, Mike Dean, Craig Partridge, Ananthram Swami, Will Leland, and James A. Hendler. 2011. [Towards a theory of semantic communication](#). In *2011 IEEE Network Science Workshop*, pages 110–117.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. [Graph of thoughts: Solving elaborate problems with large language models](#).
- Betty J Birner. 2012. *Introduction to pragmatics*. John Wiley & Sons.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. [Weak-to-strong generalization: Eliciting strong capabilities with weak supervision](#).
- Zefan Cai, Baobao Chang, and Wenjuan Han. 2023. [Human-in-the-loop through chain-of-thought](#).
- Rudolf Carnap and Yehoshua Bar-Hillel. 1954. [An outline of a theory of semantic information](#). *Journal of Symbolic Logic*, 19(3):230–232.
- Antoine Chaffin, Vincent Claveau, and Ewa Kijak. 2022. [PPL-MCTS: Constrained textual generation through discriminator-guided MCTS decoding](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2953–2967, Seattle, United States. Association for Computational Linguistics.
- Daniel Chandler. 2022. *Semiotics: the basics*. Routledge.
- Harrison Chase. 2022. Langchain. <https://github.com/hwchase17/langchain>.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023a. [Fireact: Toward language agent fine-tuning](#).
- Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023b. [Accelerating large language model decoding with speculative sampling](#).
- Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024a. [Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models](#).
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023c. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#).
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. [Self-play fine-tuning converts weak language models to strong language models](#).

- Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. 2019. [A fully differentiable beam search decoder](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1341–1350. PMLR.
- Luciano Del Corro, Allie Del Giorno, Sahaj Agarwal, Bin Yu, Ahmed Awadallah, and Subhabrata Mukherjee. 2023. [Skipdecode: Autoregressive skip decoding with batching and caching for efficient llm inference](#).
- Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. 2023. [Collaborating with language models for embodied reasoning](#).
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [Llm.int8\(\): 8-bit matrix multiplication for transformers at scale](#).
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#).
- Daniel Deutsch, Shyam Upadhyay, and Dan Roth. 2019. [A general-purpose algorithm for constrained sequential inference](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 482–492, Hong Kong, China. Association for Computational Linguistics.
- Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. [Everything of thoughts: Defying the law of penrose triangle for thought generation](#).
- Yu Du, Fangyun Wei, and Hongyang Zhang. 2024. [Anytool: Self-reflective, hierarchical agents for large-scale api calls](#).
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. [Prompt-breeder: Self-referential self-improvement via prompt evolution](#).
- Luciano Floridi. 2009. [Philosophical conceptions of information](#). *Lecture Notes in Computer Science - LNCS*, 5363:13–53.
- Jerry A. Fodor. 1975. *The Language of Thought*. Harvard University Press.
- Elias Frantar, Saleh Ashkboos, Torsten Hoeffler, and Dan Alistarh. 2023. [Gptq: Accurate post-training quantization for generative pre-trained transformers](#).
- Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. [Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes](#).
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Pal: Program-aided language models](#).
- Atticus Geiger, Zhengxuan Wu, Karel D’Oosterlinck, Elisa Kreiss, Noah D. Goodman, Thomas Icard, and Christopher Potts. 2022. [Faithful, interpretable model explanations via causal abstraction](#). Stanford AI Lab Blog.
- Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. [Grammar-constrained decoding for structured NLP tasks without finetuning](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.
- Shahriar Golchin and Mihai Surdeanu. 2024. [Time travel in llms: Tracing data contamination in large language models](#).
- Çağlar Gülçehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. [Reinforced self-training \(rest\) for language modeling](#). *CoRR*, abs/2308.08998.

- Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2017. [Decoding with value networks for neural machine translation](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. [Metagpt: Meta programming for multi-agent collaborative framework](#). *CoRR*, abs/2308.00352.
- Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.
- Bin Hu, Chenyang Zhao, Pu Zhang, Zihao Zhou, Yuanhang Yang, Zenglin Xu, and Bin Liu. 2023. [Enabling intelligent interactions between an agent and an llm: A reinforcement learning approach](#).
- J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. [Improved lexically constrained decoding for translation and monolingual rewriting](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aoran Jiao, Tanmay P. Patel, Sanjmi Khurana, Anna-Mariya Korol, Lukas Brunke, Vivek K. Adajania, Utku Culha, Siqi Zhou, and Angela P. Schoellig. 2023. [Swarm-gpt: Combining large language models with safe motion planning for robot choreography design](#).
- Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. [GenIE: Generative information extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643, Seattle, United States. Association for Computational Linguistics.
- Martin Josifoski, Lars Klein, Maxime Peyrard, Yifei Li, Saibo Geng, Julian Paul Schnitzler, Yuxing Yao, Jiheng Wei, Debjit Paul, and Robert West. 2023a. [Flows: Building blocks of reasoning and collaborating ai](#).
- Martin Josifoski, Maxime Peyrard, Frano Rajič, Jiheng Wei, Debjit Paul, Valentin Hartmann, Barun Patra, Vishrav Chaudhary, Emre Kiciman, and Boi Faltings. 2023b. [Language model decoding as likelihood–utility alignment](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1455–1470, Dubrovnik, Croatia. Association for Computational Linguistics.
- Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023c. [Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.
- Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. [Meganno+: A human-llm collaborative annotation system](#).
- Sehoon Kim, Karttikeya Mangalam, Suhong Moon, Jitendra Malik, Michael W. Mahoney, Amir Gholami, and Kurt Keutzer. 2023. [Speculative decoding with big little decoder](#).
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Wouter Kool, Herke Van Hoof, and Max Welling. 2019. [Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3499–3508. PMLR.

- Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. [RankGen: Improving text generation with large ranking models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2023. [Orchestrallm: Efficient orchestration of language models for dialogue state tracking](#).
- Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. 2023a. [Chain of code: Reasoning with a language model-augmented code emulator](#).
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.
- Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023c. [Symbolic chain-of-thought distillation: Small models can also "think" step-by-step](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.
- Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023d. [Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation](#).
- Yucheng Li, Frank Guerin, and Chenghua Lin. 2023e. [Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction](#).
- Zachary C. Lipton. 2018. [The myths of model interpretability](#). *Commun. ACM*, 61(10):36–43.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. [Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization](#).
- Jieyi Long. 2023. [Large language model guided tree-of-thought](#).
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. [Chameleon: Plug-and-play compositional reasoning with large language models](#). *ArXiv*, abs/2304.09842.
- Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. [Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms](#).
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.
- Clara Meister, Afra Amini, Tim Vieira, and Ryan Cotterell. 2021. [Conditional Poisson stochastic beams](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 664–681, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. [If beam search is the answer, what was the question?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.
- Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox, and Ryan Cotterell. 2023. [On the efficacy of sampling adapters](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1455, Toronto, Canada. Association for Computational Linguistics.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023a. [Augmented language models: a survey](#). *CoRR*, abs/2302.07842.
- Grégoire Mialon, Clémentine Fourier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023b. [Gaia: a benchmark for general ai assistants](#).
- Yohei Nakajima. 2023. [Babyagi](https://github.com/yoheinakajima/babyagi). <https://github.com/yoheinakajima/babyagi>.

- Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2023. [Skeleton-of-thought: Large language models can do parallel decoding](#).
- Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. [Reward augmented maximum likelihood for neural structured prediction](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. [Refiner: Reasoning feedback on intermediate representations](#).
- Judea Pearl and Dana Mackenzie. 2018. *The Book of Why*. Basic Books, New York.
- Maxime Peyrard. 2019. [A simple theoretical model of importance for summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.
- Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2022. [Efficiently scaling transformer inference](#).
- Matt Post and David Vilar. 2018. [Fast lexically constrained decoding with dynamic beam allocation for neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.
- Angela Potochnik. 2017. *Idealization and the Aims of Science*. University of Chicago Press, Chicago.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#).
- Guillem Ramírez, Matthias Lindemann, Alexandra Birch, and Ivan Titov. 2023. [Cache & Distil: Optimising API Calls to Large Language Models](#).
- Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. [A generalist agent](#).
- Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. [Deep reinforcement learning-based image captioning with embedding reward](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1151–1159. IEEE Computer Society.
- Michael Rescorla. 2023. The Language of Thought Hypothesis. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should i trust you?": Explaining the predictions of any classifier](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 1135–1144, New York, NY, USA. Association for Computing Machinery.
- Toran Bruce Richards. 2023. Autogpt. <https://github.com/Significant-Gravitas/Auto-GPT>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. [High-resolution image synthesis with latent diffusion models](#). *CoRR*, abs/2112.10752.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2023. [Mathematical discoveries from program search with large language models](#). *Nature*.
- Subhro Roy, Sam Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2022. [Benchclamp: A benchmark for evaluating language models on semantic parsing](#).

- Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. [PICARD: Parsing incrementally for constrained auto-regressive decoding from language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. [Confident adaptive language modeling](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17456–17472. Curran Associates, Inc.
- Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. 2023. [Algorithm of thoughts: Enhancing exploration of ideas in large language models](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Claude Elwood Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27:379–423.
- Yongliang Shen, Kaitao Song, Xu Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. 2023. [Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface](#). *ArXiv*, abs/2303.17580.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. [Detecting pretraining data from large language models](#).
- Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. [Constrained language models yield few-shot semantic parsers](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#).
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. [Mastering chess and shogi by self-play with a general reinforcement learning algorithm](#).
- Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2023. [Beyond human data: Scaling self-training for problem-solving with language models](#).
- Yixin Song, Zeyu Mi, Haotong Xie, and Haibo Chen. 2023. [Powerinfer: Fast large language model serving with a consumer-grade gpu](#).
- Felix Stahlberg and Bill Byrne. 2019. [On NMT search errors and model errors: Cat got your tongue?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.
- Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. [Blockwise parallel decoding for deep autoregressive models](#). *CoRR*, abs/1811.03115.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. [Does synthetic data generation of llms help clinical text mining?](#)

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.
- Roy Tromble and Jason Eisner. 2006. [A fast finite-state relaxation method for enforcing global constraints on sequence decoding](#). In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 423–430, New York City, USA. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. [Generating faithful synthetic data with large language models: A case study in computational social science](#).
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. [Self-consistency improves chain of thought reasoning in language models](#).
- Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. 2023b. [Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models](#). *arXiv preprint arXiv: 2311.05997*.
- Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. [Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents](#). *arXiv preprint arXiv:2302.01560*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *CoRR*, abs/2201.11903.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. [Large language models are better reasoners with self-verification](#).
- Sam Wiseman and Alexander M. Rush. 2016. [Sequence-to-sequence learning as beam-search optimization](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.
- James F. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.
- Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023a. [Multimodal large language models: A survey](#).
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023b. [Autogen: Enabling next-gen LLM applications via multi-agent conversation framework](#). *CoRR*, abs/2308.08155.
- Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. 2024. [Os-copilot: Towards generalist computer agents with self-improvement](#).
- Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. [Smoothquant: Accurate and efficient post-training quantization for large language models](#).
- Hengjia Xiao and Peng Wang. 2023. [Llm a*: Human in the loop large language models enabled a* search for robotics](#).
- Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. [Self-evaluation guided beam search for reasoning](#).
- Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. 2023. [Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought](#).
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models](#).

- Yao Yao, Zuchao Li, and Hai Zhao. 2023b. [Beyond chain-of-thought, effective graph-of-thought reasoning in large language models.](#)
- Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2023c. [Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation.](#)
- Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. [On completeness-aware concept-based explanations in deep neural networks.](#) In *Advances in Neural Information Processing Systems*, volume 33, pages 20554–20565. Curran Associates, Inc.
- Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2023. [Large language model cascades with mixture of thoughts representations for cost-efficient reasoning.](#)
- Yixin Zhong. 2017. [A theory of semantic information.](#) *Proceedings*, 1(3).
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. [Least-to-most prompting enables complex reasoning in large language models.](#)
- Marija Šakota, Maxime Peyrard, and Robert West. 2024. [Fly-swat or cannon? cost-effective language model choice via meta-modeling.](#) In *Proceedings of The International ACM Conference on Web Search and Data Mining (WSDM)*.