Controllable Text Generation for Large Language Models: Methods, Challenges, and Future Directions

SurveyForge

Abstract— Controllable text generation represents a rapidly evolving field that enables large language models to produce text conditioned on desired attributes such as sentiment, style, factuality, or tone. This survey provides a comprehensive overview of the core methodologies, challenges, and applications of controllable text generation for large language models (LLMs). It outlines foundational techniques, including prompt engineering, prefix tuning, energy-based decoding, and retrieval-augmented mechanisms, for achieving nuanced control during text generation. Taxonomies of control dimensions, such as style, structure, content, and task-specific outputs, are discussed alongside advanced multi-attribute frameworks and emerging trends in multimodal systems. The paper examines key challenges, including trade-offs between adherence to constraints and linguistic quality, hallucination, biases, and computational efficiency, while proposing solutions such as modular architectures, causal frameworks, and hybrid evaluation metrics. Emerging applications span creative writing, personalized conversational agents, cross-cultural adaptation, and domain-specific content generation. Concluding with future directions, the survey emphasizes scalable control mechanisms, real-time adaptability, universal benchmarking, and ethical safeguards to ensure robust and human-aligned implementations. This synthesis aims to advance the field by bridging methodology, usability, and research opportunities to address open challenges and maximize the transformative potential of controllable text generation systems.

Index Terms—controllable text generation, multi-attribute frameworks, retrieval augmentations

1 Introduction

Controllable text generation stands at the intersection of creativity and precision within the field of natural language processing (NLP), enabling the production of text conditioned on specific attributes or constraints. At its core, controllable text generation aims to adapt the outputs of language models to desired characteristics such as sentiment, tone, factuality, style, or structural format. This capability is pivotal for advancing both functionality and safety in AI-driven text generation, as it enables personalized, domain-specific, and ethically aligned applications.

Traditional NLP models exhibited limited capacity for controllable generation, often requiring fine-tuned statistical systems or manually encoded rules governing specific attributes [1]. The rise of neural network-based language models, particularly those pre-trained on enormous corpora such as GPT-2, GPT-3, and T5, has transformed this landscape. These large language models (LLMs) possess unparalleled generative fluency, offering a foundation to enable nuanced and scalable control [2]. The introduction of specific mechanisms—such as control tokens, prompt engineering, fine-tuning procedures, and decoding-time interventions—has further expanded the practicality of controllable text generation [3], [4].

Despite these advancements, significant trade-offs underscore the challenges involved. Pre-trained models like GPT-3 and PaLM often achieve control through fine-tuning or reinforcement learning with human feedback (RLHF), processes that demand immense computational resources and carefully curated datasets [5]. Techniques such as energy-based decoding frameworks and attribute classifiers offer lightweight alternatives, enabling control without retraining the base model [4], [6]. These approaches capitalize on the modular separation of generation and control mechanisms, allowing domain adaptation at minimal cost. However, such methods can introduce challenges in balancing control and fluency, often requiring post-hoc validation or adjustment to ensure fidelity and coherence [7].

The integration of explicit external knowledge into LLMs marks another transformative milestone in this space. Frameworks such as MEGATRON-CNTRL utilize dynamic knowledge retrieval and ranking to imbue generation systems with domain awareness, improving both coherence and adherence to constraints [8]. Although such advancements hold promise for controllable generation tasks in high-stakes contexts (e.g., legal or medical text), they underscore the persistent issue of hallucination—where models introduce plausible yet unfounded information [9]. Techniques such as retrieval-augmented generation and counterfactual reasoning have shown promise in reducing these errors [10].

Emerging trends reflect increasing sophistication in decomposition and composition mechanisms for multi-attribute controls. Methods such as DExperts leverage expert and anti-expert models during decoding to mitigate undesired attributes, such as toxicity while reinforcing compliance with desired qualities [11]. Dynamic attribute graphs, on the other hand, represent a promising avenue for achieving nuanced, context-specific control across multiple constraints [12]. These architectures align control objectives with probabilistic distributions, offering scalable solutions for balancing conflicting constraints without overfitting to any individual attribute.

While controllable text generation has demonstrated vast potential, it is constrained by evaluation limitations. Existing metrics often fall short in capturing multifaceted criteria such as fluency, constraint adherence, and semantic coherence simultaneously [13]. Standardized benchmarks and robust human-machine hybrid evaluation pipelines, such as those proposed by Texygen, are critical for advancing the reliability and interpretability of these systems [14]. Furthermore, ethical considerations, such as preventing bias amplification and ensuring inclusivity, must be embedded within both training and deployment pipelines [15].

In the ever-evolving field of NLP, controllable text generation exemplifies a paradigm shift from static, task-specific models to more adaptable, multifaceted systems. As techniques mature, integrating multimodal control, cross-lingual performance extensions, and real-time adaptability are promising directions for further research. These developments aim to produce text generation systems capable of delivering tailored, transparent, and ethically robust outputs, paving the way for safe and transformative AI applications [5].

2 Taxonomy and Dimensions of Control in Text Generation

2.1 Content Control

Content control in text generation focuses on steering the model's outputs to match predefined content-specific requirements, such as factuality, domain adherence, or thematic accuracy. As large language models (LLMs) underpin most text generation systems today, achieving robust content control is vital for deploying these systems in high-stakes applications like news reporting, technical writing, and fact-sensitive communication. The complexity of aligning unconstrained, generative systems with externally defined content constraints presents unique technical challenges, which have spurred the development of diverse methodologies. This subsection examines key approaches to content control, evaluates their strengths and limitations, and identifies future research directions in this dimension of controllability.

One foundational mechanism for content control is **retrieval-augmented generation**, where a language model incorporates contextually relevant data retrieved from external knowledge sources into its output. This approach ensures that generated content aligns with up-to-date information, reducing hallucinations while improving factual accuracy. Systems like MEGATRON-CNTRL integrate external knowledge through keyword predictors, knowledge retrieval, and ranking modules to generate content that remains consistent with specified constraints [8]. While retrieval-augmented generation offers enhanced accuracy, its reliance on retrieval pipelines introduces latency and computational overhead, making scalability a challenge for complex tasks.

Another methodology involves **constrained decoding strategies** tailored to enforce content-related constraints during text generation. Techniques such as NeuroLogic A*esque leverage future-sequence heuristic estimates to satisfy lexical and semantic constraints, enabling guided generation that adheres to predefined content elements [16]. These decoding-time methods are particularly effective in tasks requiring strict keyword inclusion or semantic consistency, but they often struggle with tasks necessitating broad contextual understanding or dynamic adaptation. Similarly, NeuroLogic Decoding demonstrates the potency of predicate logic in managing hard lexical constraints, achieving high flexibility across tasks [10].

Fine-tuning LLMs on **domain-specific corpora** is another widely explored approach. By curating and embedding specialized datasets into the training pipeline, domain relevance and thematic consistency can be explicitly reinforced. However, traditional fine-tuning methods demand significant computational resources and risk overfitting, limiting their ability to generalize across diverse content-generation scenarios. Parameter-efficient tuning techniques, such as low-rank adaptation (LoRA), mitigate these challenges by only modifying a subset of model parameters while maintaining strong adherence to control objectives [17]. This balance of efficiency and performance makes LoRA particularly attractive for domain-specific use cases like technical documentation or legal drafting.

Plug-and-play strategies also represent a novel paradigm in content control by eschewing retraining altogether. PPLM allows pretrained models to interact with lightweight attribute classifiers that enforce content parameters through iterative gradient adjustments [4]. While these methods are highly flexible and computationally efficient, certain constraints requiring nuanced global coherence or large-scale structural adjustments often exceed their capabilities.

Custom knowledge integration provides another promising avenue, enhancing content alignment by embedding dynamic resources such as attribute graphs and structured knowledge bases. Approaches like Dynamic Attribute Graphs evaluate attribute alignment via scoring mechanisms and modulate key attribute-word distributions to maintain coherence while modifying content based on external constraints [12]. Such mechanisms introduce interpretability and adaptability but necessitate improved scalability and robustness under real-world input variability.

Critical challenges persist in the quest for reliable content control. Balancing fluency, creativity, and strict adherence to content constraints is inherently conflicting, particularly in multinodal control tasks that require simultaneous enforcement of multiple constraints (e.g., factuality and sentiment alignment). Emergent methods like energy-based modeling, which parameterize content attributes as differentiable energy functions during decoding, have demonstrated potential for fine-grained control without retraining [6].

Future directions in content control include refining reinforcement learning techniques via reward signals specific to content fidelity and exploring collaborative feedback systems between users and models to iteratively refine adherence to requirements in real time. Additionally, leveraging multimodal inputs, such as combining textual and visual data, could further enrich content alignment, particularly in domains like news reporting and educational content generation. Beyond technical innovations, standardizing evaluation metrics, such as integrating tools like AIS for source attribution [18], will be essential for benchmarking performance consistently across diverse applications. As LLMs continue to scale, addressing these challenges will be pivotal to ensuring that content control frameworks deliver both precision and adaptability.

2.2 Style and Tone Control

Style and tone control in text generation focuses on finely steering linguistic attributes such as sentiment, emotional expression, formality, politeness, and creativity. This capability complements content and structural control by ensuring that generated outputs are not only accurate and well-organized but also contextually appropriate in their stylistic presentation. As large language models (LLMs) continue to permeate applications like creative writing, conversational agents, and personalized communication, the ability to customize style and tone has gained increasing prominence.

One of the most extensively studied areas in style and tone control is sentiment and emotional modulation. Techniques leveraging fine-tuned discriminators have shown effectiveness in guiding outputs to reflect specific sentiments, such as positivity, neutrality, or negativity [11], [19]. Models like GeDi employ a smaller generative discriminator to regulate token probabilities under Bayes' rule, ensuring both stylistic alignment and linguistic coherence [19]. Similarly, DExperts integrates "expert" and "anti-expert" models during decoding to dynamically balance desired attributes while preserving fluency and without altering base model parameters [11]. These approaches, while effective, face trade-offs, particularly when strong constraints compromise output diversity, leading to stylistic redundancy.

Formality and politeness control represent another crucial frontier within this domain. Techniques such as style embeddings and prefix-tuning provide means to encode stylistic attributes like "formal" versus "informal" tones, enabling models to generate outputs tailored to distinct social or professional contexts [17], [20]. Prefix-tuning, for instance, introduces small attribute-specific vectors into the input prompt, achieving computational efficiency compared to full fine-tuning [17]. However, fine-grained adjustments to degrees of formality or politeness (e.g., mildly formal vs. highly formal) often require more sophisticated mechanisms, such as scoring-based adjustments or reinforcement learning, as highlighted in multi-attribute benchmarks [21].

Creative writing and poetic text generation offer distinct challenges requiring highly adaptive style modulation. Frameworks like hierarchical conditioning and attribute graphs have emerged as powerful tools to integrate content and stylistic goals. Dynamic attribute graphs, for example, map stylistic attributes to specific tokens, enabling flexible integration of creative elements during the text generation process [12]. Hierarchical conditioning further decomposes global stylistic traits into granular sentence-level attributes for better alignment between creative objectives and linguistic structure [22]. Nonetheless, these methods often involve substantial computational complexity and are particularly resource-intensive in real-time applications, posing scalability challenges.

Maintaining stylistic consistency across domains—such as conversational, professional, and creative outputs—emerges as a pivotal challenge. Latent space manipulation techniques, such as those employed by the CoCon framework, offer promising advancements by isolating style-specific vectors and adjusting them for zero-shot stylistic adaptation [23]. However, manipulating high-dimensional latent spaces can lead to content drift or reduced fidelity, especially when managing complex interactions between multiple stylistic attributes.

Emerging approaches target task-agnostic, interpretable mechanisms for improved adaptability across diverse style and tone requirements. For instance, energy-based models (EBMs) optimize token probabilities to balance stylistic fidelity against fluency, providing a flexible pathway for multi-attribute control. NeuroLogic Decoding demonstrates how predicate logic constraints can guide stylistic relevance without sacrificing linguistic authenticity [10]. Additionally, reinforcement learning-based frameworks, such as those leveraging actor-critic models, have been explored to handle complex stylistic objectives in creative and formal contexts [24].

The rise of user-centric applications underscores the need for interpretability and adaptability in style control techniques. Adaptive plug-and-play architectures, such as those enabled by dynamic attribute graphs, allow compatibility with black-box models while maintaining efficiency and modularity [12]. Causal modeling frameworks have also begun addressing stylistic biases in generated text, offering tools to couple style control with fairness and debiasing initiatives [25].

Future directions in style and tone control will likely emphasize methods that accommodate increasingly nuanced and context-sensitive stylistic demands. The integration of user feedback to enable iterative refinements during real-time interactions represents a promising avenue. Furthermore, hybrid approaches, such as combining retrieval-augmented generation with stylistic controls, could enable a more robust balance between stylistic variety and semantic fidelity. These advancements are poised to unlock greater personalization and adaptability across diverse applications, spanning creative, conversational, and professional domains.

2.3 Structural Control

Structural control in text generation encompasses techniques aimed at regulating organizational elements of text, including sentence complexity, paragraph structuring, and narrative sequencing. This dimension is fundamental to tasks such as hierarchical summarization, technical reporting, and long-form narrative construction, where maintaining coherence and adhering to structural constraints are critical. Unlike content or style control, structural control prioritizes text arrangement and logical flow, ensuring outputs conform to predefined frameworks while preserving natural linguistic quality.

One prominent approach to structural control is **syntax-driven generation**, where models explicitly incorporate syntactic features or parse trees as constraints during decoding. For instance, syntax-informed methods leverage dependency parses or constituency trees to guide sentence structuring, enabling fine-grained control over sentence complexity or grammatical composition. Techniques using syntax parsers integrated with transformers, such as Syntax Guided Controlled Paraphraser [26], exhibit strong adherence to structural constraints while maintaining semantic fidelity. However, these methods often inherit the limitations of syntactic parsing, including error propagation when parse trees are inaccurate. Additionally, decoding complexity escalates with more intricate syntactic constraints, posing computational challenges in real-time applications.

Another critical focus of structural control lies in **paragraph-level organization**. Modern methods aim to ensure logical progression across sentences and paragraphs by incorporating discourse-aware mechanisms. For example, transformer-based models augmented with hierarchical latent spaces can control inter-sentence coherence by leveraging embeddings at the discourse level [27]. These models often employ objectives like distinguishing between shuffled and coherent sentence orders, thus learning to enforce structural continuity. While effective for coherence, these approaches are constrained by their reliance on labeled discourse corpora, which are limited in availability and scope.

Narrative sequencing in long-form generation represents another frontier for structural control, particularly for creative and argumentative content. Iterative planning frameworks iteratively refine narrative direction by integrating high-level plans, such as event chains or discourse-level keywords, into decoding [28], [29]. These planning techniques overcome the limitations of local token-level decisions, enabling models to maintain thematic alignment over extended texts. Such methods, however, introduce challenges of balancing rigidity with creativity—overly constrained plans may reduce fluency or lead to dull output, while insufficient planning risks incoherence.

Emerging trends in structural control leverage cutting-edge frameworks like **latent-space optimization** and **dynamic attribute graphs**, which disentangle structural attributes from content semantics. Latent-variable models use probabilistic formulations to encode structure, enabling finer manipulation of attributes such as sentence order or hierarchical discourse without requiring explicit supervision [30]. Similarly, dynamic attribute graphs enable text structuring by dynamically adjusting the prominence of structural features in generated text [12]. These methods provide considerable flexibility but require high-dimensional optimization techniques that can increase model complexity and computational overhead.

Despite advances, challenges persist. Structural control often suffers from tension between global coherence and local flexibility, especially in tasks requiring simultaneous control over multiple levels of detail (sentence, paragraph, and discourse). Additionally, the lack of robust benchmarks for structural adherence complicates systematic evaluation. Future research should explore hybrid techniques combining explicit syntactic rules with neural optimization frameworks, enabling more scalable and robust control. Integrating causal modeling frameworks could further enhance our understanding of structural dependencies and mitigate spurious correlations often seen in training data [25].

In conclusion, structural control facilitates domain-specific and hierarchical text generation, but achieving a balance between rigidity and diversity demands innovative approaches, such as latent variable modeling and attribute-driven optimization. The development of universal benchmarks and scalable frameworks will be indispensable in advancing this critical dimension of controllable text generation.

2.4 Task-Specific Control

Task-specific control in text generation represents a cornerstone of aligning output text with specific operational objectives, spanning use cases like translation, summarization, question answering, and structured data-to-text tasks. Unlike generalized controls that manage cross-domain attributes such as style or sentiment, this form of control focuses on task-level fidelity, ensuring that generated outputs adhere closely to the functional requirements of a given application. Achieving task-specific control hinges on an intricate interplay of model architecture, fine-tuning methods, retrieval-augmented strategies, and decoding techniques.

A foundational approach involves fine-tuning LLMs on task-specific datasets. By using annotated task data, supervised fine-tuning optimizes models for precise objectives like generating summaries or translating between languages. For instance, frameworks in structured data-to-text systems combine content selection and logical sequencing to tailor generated text to domain-specific objectives [31], [32]. An innovative advancement in this paradigm is decoupling planning from surface realization [33], where the generation process is divided into two distinct stages—a content plan is first created as an intermediate representation, followed by the realization of natural language output. This two-step structure enhances reliability and ensures tighter compliance with task-specific constraints.

Another key strategy leverages control codes or prompt-based conditioning for dynamic task alignment at inference time. The CTRL model [3] is a notable example, embedding task-specific control codes within the model architecture to adjust the generation process on-the-fly, bypassing the need for exhaustive fine-tuning. This method significantly broadens adaptability across a range of generative tasks. However, control code-based systems often encounter difficulty in managing nuanced demands, such as producing semantically abstract summaries or sustaining contextual coherence in open-ended dialogue, where domain-specific adaptability is essential.

Decoding-time optimization techniques have also expanded the landscape of task-specific control. Constrained decoding algorithms, such as those implementing task-specific beam search with penalties, allow models to integrate lexical and structural constraints while maintaining fluency [16]. Such methods are indispensable for tasks like keyword-guided text generation and grammatically precise machine translation. Similarly, retrieval-augmented generation approaches [8] employ external knowledge bases to ground outputs in domain-relevant information, a capability crucial for applications like producing context-aware narratives or responses. These techniques dynamically retrieve and rank evidence, enabling greater factual consistency and contextual accuracy across tasks.

Reinforcement learning (RL)-based methods offer further advancements by optimizing task-focused metrics to enhance generative fidelity. For example, RL-driven techniques used for alignment with question-answering objectives or consistent length in summarization rely on gradient-based fine-tuning to iteratively refine outputs [24]. However, RL approaches face challenges such as sample inefficiencies and susceptibility to adversarial drift, particularly in multi-objective tasks requiring simultaneous optimization across divergent performance measures.

Emerging trends in task-specific control are increasingly incorporating multimodal interactions and adaptive constraint management. Tasks involving cross-modal integration, such as image captioning, highlight the importance of aligning text generation with visual information [34]. These systems ensure that generated captions maintain semantic alignment with input images. Additionally, dynamic attribute graphs [12] have shown promise in reasoning-intensive tasks, where interdependencies among variables and constraints evolve over multiple steps of generation.

Despite significant progress, challenges persist in optimizing task-specific control. Balancing fine-grained control across diverse tasks, managing fluency against functional accuracy, and minimizing computational demands during model adaptation remain open areas for improvement. Hybrid frameworks that combine symbolic reasoning with neural deep learning architectures could provide more robust generalization across complex task settings. Furthermore, advancements in retrieval-augmented systems—for example, integrating mechanisms for domain-specific memory retrieval [35]—could enhance the scalability and adaptability of task-specific solutions in resource-intensive applications.

In conclusion, task-specific control serves as a critical enabler for tailoring outputs to sophisticated application requirements, embodying both immediate functional precision and a growing sophistication in scalable generative approaches. Continued innovations in multi-step planning, optimization, and cross-modal alignment will be essential in addressing new challenges and extending the utility of controllable text generation across diverse domains.

2.5 Granular Attribute Combinations and Multimodal Control

Granular attribute combinations and multimodal control represent an intersection of advanced controllable text generation techniques, where complex attribute combinations and cross-modal interactions underpin the generation process. These approaches are pivotal to enabling nuanced outputs that adhere to multiple constraints simultaneously, while also ensuring coherence when integrating diverse input modalities such as text, images, or audio.

Controlling multiple attributes simultaneously has been a long-standing challenge due to the potential for attribute interference and trade-offs in fidelity. For instance, models tasked with enforcing sentiment alignment and maintaining formal tone may encounter conflicts in style generation. Techniques such as dynamic weighting and reward balancing address this by dynamically adjusting the influence of individual attributes during decoding or optimization phases. Methods like MuCoCO [36] directly frame multi-attribute optimization as a relaxed continuous optimization problem, leveraging Lagrangian multipliers to achieve balance between constraints. This approach mitigates the direct conflicts between attributes while preserving the underlying generative distribution of the language model. Similarly, distribution-matching paradigms [37] minimize the divergence between the model's output distribution and desired constraints, allowing for smoother interactions between complex combinations of attributes.

A major limitation of non-adaptive multi-attribute systems is their lack of responsiveness to dynamic scenarios. Real-time systems, such as those deployed in conversational agents or user-guided applications, necessitate instant adjustments to attributes based on evolving inputs or user feedback. Self-Refine [38] introduces iterative refinement, where large language models (LLMs) self-critique and iteratively update outputs, dynamically improving adherence to multiple attribute dimensions without retraining. This mechanism is especially potent for fine-grained control, as it uses feedback-based refinement to correct outputs that diverge from constraints.

Multimodal control extends these principles by integrating signals from diverse modalities. For example, in text-to-image or text-to-video systems, attributes such as tone, style, or semantic alignment must be harmonized with image-driven details. MAGIC [39] achieves this by scoring generated text against image relevance scores derived from pre-trained models like CLIP. This scoring dynamically influences the most appropriate text tokens, ensuring semantic and contextual alignment with images. Similarly, in systems like ControlVideo [40], structural consistency across frames is maintained by hierarchical sampling and interleaved-frame smoothing, ensuring that temporally evolving modalities such as video generation maintain coherence while adhering to textual controls.

A key enabler of multimodal systems is the use of shared latent spaces, as seen in UniControl [41]. By aligning textual, visual, and other modalities into a unified representation, such systems enable precise control over multimodal outputs, balancing attribute-specific constraints without prioritizing one modality over another. This establishes a cohesive framework for combining modalities and attributes seamlessly.

Despite these advances, several challenges persist. The inherent complexity of aligning granular attributes across modalities can lead to conflicting signals that degrade the overall quality of outputs. Balancing attributes dynamically requires robust scoring mechanisms, yet current techniques often lack explainability or transparency in how trade-offs are moderated. Moreover, multimodal control relies on extensive pretraining across modalities, which necessitates computationally intensive pipelines. Future advancements could focus on scalable, lightweight interventions such as pluggable virtual tokens [42], which integrate multimodal inputs or multiple attributes without retraining the foundational model, enhancing adaptability in resource-constrained environments.

In conclusion, granular attribute combinations and multimodal control represent the frontier of controllable text generation, promising unprecedented flexibility and applicability. Fine-grained optimization frameworks, combined with cross-modal integration capabilities, offer dynamic and contextually adaptive generation. However, addressing conflicts between attributes and modalities while ensuring computational efficiency remains critical for scalability and broader adoption. The interplay of iterative refinement mechanisms, unified multimodal representations, and distributional approaches will likely define the next phase of innovation in this domain.

3 Foundational Approaches and Core Techniques for Controllability

3.1 Prompt-based Techniques

Prompt-based techniques have emerged as foundational approaches for controllability in large language models (LLMs), leveraging the input prompt to guide text generation. These techniques eschew altering the underlying weights of the model, establishing themselves as lightweight and flexible alternatives to fine-tuning or reinforcement learning methods. They are particularly attractive because of their dynamic adaptability, allowing users to steer generation outputs by carefully engineering instructions, examples, or demonstrations provided to the model.

At their core, prompt-based techniques exploit the pre-trained capabilities of LLMs by leveraging the contextual information encapsulated in the prompt. For instance, "controlled text generation with natural language instructions" [43] has demonstrated that verbalizing constraints as instructions can effectively steer generation outputs, especially when combined with pre-trained language models like GPT or T5. These instructions align the generation context with user-defined attributes, breaking from the rigid templates of earlier rule-based systems and benefiting from the inherently flexible capabilities of transformer-based architectures [5]. By verbalizing constraints, prompt-based methods enable models to generalize to unseen tasks and adapt to varied generation requirements without requiring prior fine-tuning.

A significant branch of these techniques revolves around **prompt engineering, where static and dynamic prompts are crafted manually or programmatically to influence generation. Static prompt templates, as used in "Plug and Play Language Models: A Simple Approach to Controlled Text Generation" [4], integrate fixed phrases or keywords, whereas dynamic prompts adapt iteratively based on real-time model outputs, enhancing their effectiveness in multi-turn or feedback-rich applications.** For example, techniques like "inverse prompting," where the generated output is continuously evaluated and used to refine the underlying prompt, have been highlighted as effective strategies for alignment to control objectives [7].

However, the efficacy of prompt engineering is not universal and depends heavily on the nuances and context of the target task. One of the key challenges is **prompt sensitivity**, where minor variations in the phrasing or order of the input can lead to significant differences in the generated output. As demonstrated in "ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer" [44], task fidelity can sometimes be reduced due to variability in how prompts interact with pre-trained LLMs, especially when working across diverse linguistic or stylistic demands. Addressing such sensitivity has been a critical focus of recent studies, with innovative approaches moving toward learned or optimized prompts.

Techniques like **Prompt Tuning**, which trains continuous embeddings directly linked to task-specific prompts, offer an automated alternative to manual engineering. Although computationally lighter than full fine-tuning, this approach retains some level of task-optimization capability [12]. Continuous prompt embeddings, wherein the prompt is represented in latent space rather than human-readable text, have also demonstrated advantages in reducing prompt-order sensitivity and improving generalization for complex tasks such as multi-attribute text generation [17].

Moreover, **composite and concatenated prompts have gained attention as viable approaches for joint control of multiple constraints or attributes. These are crafted by combining independent prompt conditions into a single encoded input, enabling the simultaneous enforcement of stylistic, structural, and content constraints [45]. Yet, challenges of composite prompt engineering include managing conflicts between overlapping constraints, where trade-offs between attributes must be codified explicitly within the prompt structure.**

One of the most promising advancements to address these limitations lies in **optimization frameworks for prompt refinement**, such as differentiation through latent embeddings or leveraging reinforcement signals. For instance, Low-Rank Adaptation (LoRA) and other gradient-based fine-tuning techniques on continuous prompts show potential for high-fidelity control incorporating user feedback loops [4]. These feedback-oriented advancements reduce the brittleness of prompt-based systems while improving alignment with user expectations through adaptive iteration.

However, the reliance on pre-trained representations emphasizes inherent constraints in prompt-based control. Models like GPT-3 or PaLM are limited by the breadth of their pre-training data, often reflecting biases or a lack of granularity in nuanced scenarios [15]. Despite these challenges, prompt-based approaches remain a cornerstone for achieving controllability in LLMs due to their accessibility, flexibility, and computational efficiency when compared to alternatives like fine-tuning or reward-model-driven optimization.

Future research is likely to focus on hybrid approaches that integrate prompt engineering with other mechanisms, such as knowledge retrieval or latent-variable activity, to enhance control [45]. Combining structured optimization, reinforcement-based feedback, and emergent few-shot learning phenomena promises more robust, interpretable, and scalable models of control. The interplay between manual and automatic prompt construction also presents opportunities to synthesize the strengths of both paradigms, creating a harmonized framework that aligns LLMs with human intent dynamically and effectively.

3.2 Fine-Tuning for Specific Controls

Fine-tuning for specific controls is a foundational approach to empowering large language models (LLMs) with the ability to consistently align with predefined, explicit objectives across diverse tasks and domains. In contrast to prompt-based techniques, which rely on external specifications, or reinforcement learning methods, which focus on dynamic optimization, fine-tuning ensures intrinsic control by embedding task-specific signals directly into the model's parameters. This method enables robust, interpretable behavior modifications tailored to control attributes such as sentiment, style, factuality, or domain-specific expertise.

Central to this approach is task-specific fine-tuning, which adjusts a model's weights to optimize for task-aligned objectives using curated datasets representative of the desired controls. For example, fine-tuned models for toxicity reduction rely on datasets annotated with toxic and non-toxic classifications, adapting the probability distributions over next-token predictions to enable safer conversational dynamics. Techniques such as GeDi [19] build on this principle by employing auxiliary models trained on labeled data to influence and refine generative outputs, demonstrating the efficacy of directly fine-tuning classifiers to produce detoxified yet fluent text. Similarly, methods like the Detoxification Generator [46] leverage contrastive learning by integrating auxiliary detoxifiers with the primary model, effectively guiding token sampling within constrained decoding frameworks.

A significant advancement in fine-tuning lies in parameter-efficient techniques that reduce computational overhead without sacrificing task performance. Methods like adapters and Low-Rank Adaptation (LoRA) selectively update small subsets of the model's parameters, leaving the majority frozen, thus facilitating efficient and cost-effective tuning [47]. Such approaches are particularly beneficial for multi-domain use cases, where lightweight parameters tuned independently for each domain or attribute allow modular integration of control objectives. Empirical studies further highlight the consistency of performance across diverse tasks while preserving the linguistic richness of the pre-trained model [12].

Expanding beyond single-attribute control, multi-domain and multi-attribute fine-tuning aim to encode diverse conditions into a unified model. However, optimizing across multiple attributes—such as sentiment, topic, and toxicity—raises challenges due to competing objectives within shared feature representations. Recent techniques address this by dynamically modeling relationships between attributes during fine-tuning. For instance, methods utilizing dynamic attribute graphs [12] or causal formulations [25] disentangle conflicting objectives, improving fidelity and specificity in generated outputs.

Curriculum learning strategies further refine fine-tuning processes by structuring training phases in a progressive manner. Beginning with simpler control objectives and gradually introducing more complex constraints, these methods enhance robustness against nuanced challenges. For instance, factual generation studies demonstrate that introducing counterfactual data early in training cycles enhances factual accuracy [48], while incremental supervision in domain-specific contexts has shown improved stylistic adherence [17].

Despite its strengths, fine-tuning poses challenges in scalability and adaptability, particularly for resource-constrained or emergent tasks. The need for extensive task-specific annotations often limits its applicability, while overfitting to these datasets can reduce generality, as evidenced by diminished adaptability to out-of-distribution inputs. To address these constraints, data augmentation strategies like Rule-based Data Recycling [49] generate synthetic datasets through rule-driven transformations, amplifying training signals while reducing dependency on manual annotation.

Looking ahead, the evolution of fine-tuning will likely involve integration with emerging paradigms such as retrieval-augmented generation [50], bridging controlled generation with external knowledge to tackle hallucination issues. Parameter-efficient fine-tuning can further complement reinforcement learning techniques, enabling scalable adaptation to real-world, dynamic feedback. Cross-lingual and multimodal extensions also present new opportunities for broader applicability and enhanced control capabilities.

In summary, fine-tuning remains central to advancing controllable text generation by embedding precise, interpretable, and robust control within LLMs. With its expanding repertoire of techniques and strategic innovations, fine-tuning continues to balance fluency with specificity, offering a critical tool for addressing the nuanced demands of both general-purpose flexibility and domain-specific utility.

3.3 Reinforcement Learning for Control Optimization

Reinforcement learning (RL) has emerged as a pivotal framework for optimizing controllable text generation by aligning language models (LLMs) with predefined objectives through reward-driven exploration. By employing reward signals—originating from human feedback, surrogate models, or both—RL-based methods enable fine-grained control of text attributes such as style, sentiment, factuality, or structural coherence, especially when such objectives are difficult to encode explicitly through supervised learning.

Central to RL approaches in controllability is the notion of a reward model that quantifies the quality of generated outputs relative to desired attributes. Methods like Proximal Policy Optimization (PPO) have been influential in balancing exploration and exploitation during policy updates. For example, OpenAI's use of PPO for fine-tuning their language models with human feedback has demonstrated significant improvements in aligning outputs with user-defined prompts while minimizing unintended consequences [24]. RL frameworks typically model text generation as a sequential decision-making problem, where tokens represent actions, and their cumulative reward is based on control fidelity and textual quality. For instance, a reward function for sentiment control might combine scores from sentiment classifiers with penalties for fluency degradation to ensure coherent yet sentiment-aligned outputs [36].

A prominent challenge in RL-based controllability arises from the reward sparsity inherent in tasks requiring complex multi-attribute optimization. Methods such as Future Discriminators for Generation (FUDGE) address this by estimating the likelihood of satisfying desired constraints earlier in generation, leveraging partial sequences to propagate attribute-specific rewards dynamically during decoding [7]. Such modular predictors allow RL systems to effectively guide generation trajectories, particularly in tasks involving multi-aspect controllability (e.g., balancing tone, factuality, and style).

The design of reward models also plays a significant role in RL's success. Traditional classification-based reward models, while effective for single-attribute control, struggle with compositionality and scalability when managing conflicting constraints [36]. To address this, hybrid approaches integrating energy-based models (EBMs) have demonstrated flexibility in encoding both soft and hard constraints by decomposing generation objectives into differentiable components. These methods iteratively optimize the latent space distributions to resolve conflicts between competing attributes [36].

Human-in-the-loop feedback remains a cornerstone of RL methods, particularly for highly subjective attributes such as tone, humor, or creativity. Techniques like preference modeling allow reward models to calibrate subjective judgments across diverse contexts. For example, ChatGPT has leveraged reinforcement learning with human feedback to refine stylistic consistency in outputs catered to specific user interactions [44]. However, reliance on human annotations poses challenges in scalability and biases—addressed partially by automated surrogates employing pre-trained attribute classifiers for reward estimation [51].

Critic-based RL methods offer further refinements by leveraging actor-critic frameworks. These approaches decouple the generation model (actor) from the reward estimation model (critic), enabling more stable and interpretable optimization. For instance, Critic-Guided Decoding manipulates token probabilities using learned critics to enforce attribute alignment with fewer training iterations, outperforming more compute-intensive fine-tuning-based approaches [24]. Similarly, contrastive reinforcement learning strategies, which bias sampling away from undesirable outcomes, provide an alternative avenue for detoxification and content improvement [21].

Despite these advancements, RL-based control optimization faces several challenges. First, the trade-off between fluency and adherence to constraints persists, notably when constraints deviate significantly from the training distribution. Second, computational overhead during fine-tuning or sampling remains a bottleneck for real-time applications. Emerging solutions, such as lightweight prefix-tuning and gradient-free optimization methods, aim to reduce this overhead while maintaining control fidelity [52]. Finally, the semantics of control often interact nonlinearly, requiring innovative approaches to disentangle conflicting attribute dependencies.

Future directions in RL-based controllability emphasize modularity and adaptability. Hierarchical RL frameworks could offer more interpretable and flexible control by structuring reward learning at multiple levels of granularity (e.g., token-level and document-level). Additionally, advances in self-supervised learning may complement RL by reducing dependence on annotated reward signals, enabling more generalized and robust control across languages and domains [53]. By integrating these innovations, RL methods stand to further expand their role in controllable language modeling, driving applications ranging from personalized content generation to safe and ethical text generation.

3.4 Architectural Innovations for Intrinsic Controllability

Advancing intrinsic controllability in large language models (LLMs) necessitates architectural innovations that embed guidance mechanisms directly into the model's structural design. Unlike external control approaches such as prompt engineering, reinforcement learning, or fine-tuning, which operate at the periphery, intrinsic methods restructure the model's internal mechanisms to seamlessly align generation with predefined attributes. This subsection examines the principal architectural innovations designed to enable intrinsic controllability, offering a complementary perspective to reinforcement learning and forming a foundational layer that hybrid methods often leverage.

A pivotal innovation in this domain is the integration of control codes within Transformer-based architectures. By associating discrete tokens or embeddings with desired attributes, models like CTRL have demonstrated how specialized control tokens can regulate stylistic or task-specific outputs without compromising fluency [3]. This approach offers scalability through its ability to encode multiple constraints within a unified framework, enabling smooth application across varying domains during training. However, the reliance on sufficiently diverse and labeled data for each control dimension represents a bottleneck, particularly for complex or less-common attributes.

Complementing control codes, latent variable models like Conditional Variational Autoencoders (CVAEs) extend controllable generation by disentangling content and control attributes in latent spaces [54]. CVAEs enable fine-grained manipulation over attributes such as sentiment or tone while maintaining coherence in generation. However, challenges in these models often include computational expense and difficulties in ensuring both interpretability and disentanglement of latent variables. Promising extensions, such as those employing distributional optimization via energy-based models, aim to address these concerns while ensuring robust generalization [37].

Further refinements arise from modular and adapter-based extensions to Transformer architectures. Structural adapters, like those developed in StructAdapt, represent a lightweight solution for integrating new control objectives without retraining the entire model [55]. By injecting task-specific modules into pre-trained systems, these architectures achieve high scalability while maintaining the integrity of the base model. Similarly, residual memory frameworks dynamically encode and refine control signals across the generation process, enabling models like RecurrentGPT to sustain contextual alignment over extended text spans [56]. These modular designs emphasize scalability and adaptability, often bridging gaps between intrinsic and hybrid control paradigms.

Emerging dynamic masking and gating mechanisms within Transformer layers further enhance intrinsic controllability. These techniques conditionally regulate attention paths and token probabilities to align outputs with target constraints, achieving fine-grained contextual control during generation [57]. Gating mechanisms offer soft constraint enforcement, balancing fidelity to user-defined objectives with generation diversity. Nonetheless, achieving computational efficiency for real-time adaptability remains a persistent challenge, particularly for interactive systems.

Attribute-aware models leveraging structured representations have demonstrated unique potential for multi-attribute tasks. Dynamic Attribute Graphs (DATGs) construct and refine graph-based structures representing diverse target constraints, enabling systematic control over token selection and latent dynamics during generation [12]. While particularly effective for high-dimensional control problems, the overhead of dynamically managing these graphs often proves computationally intensive, limiting scalability for broader applications.

Neural techniques inspired by energy-based models (EBMs) provide an evolving avenue for intrinsic control. By formulating text generation as a probabilistic optimization problem, EBMs assign energy scores to candidate sequences based on fluency and attribute fidelity, optimizing outputs through token-level calibration during inference [58]. Unlike approaches requiring extensive retraining, EBMs rely on sophisticated heuristics for real-time adaptability, presenting opportunities for modular and flexible control mechanisms.

Despite these advancements, intrinsic methods face several bottlenecks. Interpretability remains an ongoing challenge, particularly for architectures employing latent spaces. Moreover, balancing the computational complexity introduced by internal control mechanisms with the need for scalable, real-time applications is essential. Looking forward, modular innovations such as adapters and energy-based systems hold promise when integrated within hybrid frameworks leveraging feedback mechanisms and dynamic post-hoc adjustments. Combining intrinsic architectural control with extrinsic flexibility can advance LLMs capable of managing highly nuanced and diverse constraints while maintaining fluency, coherence, and domain adaptability.

3.5 Hybrid Approaches for Enhanced Control

Hybrid approaches for enhanced control in controllable text generation synthesize multiple methodologies to overcome individual limitations, striking a balance between control precision and generation quality. These methods often integrate techniques such as retrieval-augmented generation, symbolic reasoning, constraint-based optimization, and neural-guided control. By leveraging the complementary strengths of diverse strategies, hybrid methods facilitate nuanced text generation that adheres to multiple overlapping constraints while maintaining naturalness, coherence, and fluency.

Retrieval-augmented generation (RAG) is a cornerstone of hybrid paradigms, especially for knowledge-intensive tasks. These approaches augment the generative capacity of large language models (LLMs) with external retrieval systems, ensuring that outputs are both controllable and factually grounded. Models like RAG [50] retrieve relevant information from databases such as Wikipedia to condition generation, effectively mapping retrieved content to non-parametric memory for fine-grained control. While this method achieves enhanced factual alignment, challenges persist in harmonizing retrieved content with parametric knowledge embedded in LLM weights, often resulting in over-reliance on one knowledge source [50]. To address this, advanced approaches use adaptive mechanisms like dynamic retrieval per token to match context granularity [50].

Constraint-based decoding mechanisms complement RAG by imposing hard or soft constraints during inference. Constrained optimization techniques, such as those utilized in MuCoCO [36], redefine the decoding process as a differentiable optimization problem. This approach seamlessly integrates multiple constraints—e.g., stylistic consistency, factual correctness, or sentiment alignment—via Lagrangian multipliers or weighted objectives. The advantage of this method lies in its modularity and scalability across diverse tasks, though challenges remain in tuning weights for competing constraints in multi-objective settings.

Symbolic reasoning acts as another critical pillar in hybrid approaches, enabling rule-based guidance for controllability. Techniques like dynamic attribute graphs (e.g., DATG) reinforce attribute control by dynamically constructing graphs to modulate the presence of specific features such as toxicity or sentiment [12]. By leveraging symbolic logic and structured representations, these methods allow deterministic control without heavily modifying the underlying model. However, their challenges lie in scaling to complex, higher-dimensional tasks or implicitly contradictory constraints.

Hybrid systems also benefit from feedback-driven iterative refinement. Methods like Self-Refine [38] and LLM-Augmenter [48] introduce iterative loops that combine retrieval-based grounding with post-hoc feedback correction, fine-tuning outputs over multiple refinement steps. These systems demonstrate strengths in reducing factual hallucinations and improving control robustness by drawing on feedback from both external knowledge sources and task-specific utility functions, all while preserving fluency. However, avoiding compounding errors across refinement cycles constitutes an enduring challenge.

Emerging trends in hybrid approaches are increasingly focused on user-guided and mixed-initiative systems, where user feedback modulates model behaviors during generation in real time. These techniques integrate on-the-fly adjustments, such as directional prompts [59], and dynamic control through plug-and-play modules like prefix tuning [60]. While promising for personalization and real-time adaptability, such systems face computational concerns and user experience challenges at scale.

Despite their potential, hybrid strategies are not without trade-offs. Integrating diverse techniques often necessitates additional computational resources, introducing latency due to multi-step reasoning, retrieval, or refinement processes. Moreover, the interplay of implicit (neural) and explicit (rule-based or symbolic) mechanisms can lead to inconsistencies when constraints are ill-defined or competing. Nonetheless, their versatility and extensibility suggest they will remain instrumental in achieving multi-objective controllability. Future directions may prioritize optimizing hybrid systems for scalability, exploring multimodal extensions [61] and refining task-specific user engagement to enhance alignment with real-world demands.

4 Techniques for Enhanced and Granular Control in Decoding

4.1 Constrained Sampling Strategies for Decoding

Constrained sampling strategies serve as foundational tools for dynamically steering text generation during the decoding phase, ensuring adherence to predefined attributes such as fluency, diversity, and alignment with specific constraints. These methods operate on the probabilistic distributions predicted by large language models (LLMs) and aim to selectively guide token sampling while optimizing trade-offs among quality, controllability, and computational efficiency. Beam search, top-k sampling, and top-p (nucleus) sampling are among the most prominent methods in this domain, each offering distinct advantages and challenges.

Beam search is a deterministic decoding method that systematically explores a fixed number of candidate sequences (beams) at each time step, retaining the most probable ones based on cumulative log-probabilities. While it excels at generating fluent and coherent outputs, its greedy search mechanism often leads to limited diversity. To address this, constrained adaptations of beam search have been proposed, such as incorporating hard constraints (e.g., mandatory inclusion of specific keywords) or soft constraints (e.g., ensuring diversity by penalizing repetitive sequences). NeuroLogic Decoding [10] exemplifies one such extension, introducing predicate logic to enforce complex lexical constraints while preserving fluency. However, beam search variants can be computationally expensive due to their exhaustive scoring across all beams at every step, limiting scalability for large LLMs.

Top-k sampling, on the other hand, introduces stochasticity by restricting token selection to the k most probable tokens at each decoding step. This ensures manageable computational overhead while promoting diversity in the generated text. By truncating the tail of the probability distribution, top-k sampling mitigates the risk of low-probability (and often nonsensical) tokens influencing generation, as shown in studies like Plug and Play Language Models [4]. However, such truncation can inadvertently exclude valid but less likely tokens critical for complex or creative tasks, leading to overly conservative outputs.

Top-p (nucleus) sampling improves upon top-k by dynamically adjusting the set of eligible tokens. Rather than fixing the number k, it selects tokens whose cumulative probability exceeds a threshold p (e.g., 0.9), thus capturing the natural variance in token likelihood distributions. This method has demonstrated superior adaptability across tasks, balancing coherence and diversity, as outlined in studies like COLD Decoding [6]. Nevertheless, its reliance on a single threshold for all tokens may introduce limitations for fine-grained controls where specific sequences exhibit highly skewed probability distributions.

Emerging hybrid methods aim to leverage the strengths of multiple strategies. For instance, NeuroLogic A*esque Decoding [16] integrates the structured exploration of beam search with probabilistic heuristic adjustments inspired by A* search algorithms, outperforming standard approaches in lexically-constrained applications. Additionally, controlled decoding frameworks like DExperts [11] refine token probabilities by combining outputs from base LMs with "expert" and "anti-expert" models to enhance nuanced constraints such as lexical diversity or detoxification.

A critical challenge with constrained sampling remains achieving balance between rigid enforcement of constraints and natural text quality. Excessive application of constraints often results in syntactically strained or semantically inconsistent outputs, as evidenced in work by Syntax-guided Controlled Generation of Paraphrases [26], which highlights the risks of conflicts between syntactic structures and stylistic controls. Moreover, computational efficiency remains a pressing concern, particularly for real-time or interactive applications. Methods incorporating lookahead heuristics or energy-based projections, such as COLD decoding, have shown promise in accelerating inference while maintaining effective constraint satisfaction, though further progress is needed to scale these solutions for large-scale deployments.

Looking ahead, future research should focus on dynamic, context-aware adaptations of these strategies, which progressively refine token selection based on evolving conditions during decoding. Integrating reinforcement learning frameworks for real-time feedback or combining stochastic and deterministic methods with differentiable optimization techniques could further enhance the granularity of control. Lastly, comprehensive benchmarks, such as Texygen [14], should incorporate multidimensional evaluations of control fidelity, computational efficiency, and output quality to better quantify the trade-offs inherent in constrained decoding strategies.

4.2 Attribute-Specific Decoding Mechanisms

Attribute-specific decoding mechanisms are pivotal for aligning generated outputs with predefined linguistic attributes such as sentiment, style, or formality, thereby offering granular control over text generation without altering a model's parameters. Positioned at the intersection of constrained sampling strategies and knowledge-augmented frameworks, these mechanisms empower Large Language Models (LLMs) to produce attribute-conforming outputs by strategically guiding the decoding process. This subsection delves into key technical methodologies, evaluates their strengths and limitations, and highlights future trajectories for improving these techniques.

A foundational approach in this domain involves **dynamic adjustment of token probabilities** during generation. Here, token probabilities predicted by the base model are reweighted using external attribute-scoring functions, effectively biasing the output towards desired attributes. A notable example is GeDi, which uses Bayes' rule to normalize posterior probabilities over attribute-conditioned and anti-attribute-conditioned distributions, thus steering generation towards predefined linguistic goals [19]. This approach demonstrates efficacy in scenarios with binary attributes, such as sentiment or toxicity classification, but often struggles to accommodate more nuanced or multi-valued attributes, where finer control is required.

Building on this, contrastive decoding introduces penalties to deter undesirable outputs while preserving fluency and coherence [21]. By minimizing probabilities for negative sequences or attributes, this method achieves efficient control during decoding without necessitating retraining. While suitable for scenarios involving isolated attribute control, its dependence on high-quality negative samples limits versatility, particularly in managing scenarios with complex, multi-attribute interactions.

Energy-based models (EBMs) offer an alternative by formalizing attribute-specific decoding as a constrained optimization problem. Here, target distributions are parameterized through energy functions that combine fluency metrics with attribute-alignment terms [37]. Sampling from these distributions ensures explicit enforcement of linguistic constraints, although issues related to computational overhead and sampling convergence remain significant challenges. These methods excel in scenarios where high precision is required but can be prohibitively expensive in real-time applications or interactive systems.

Repetition and degeneration suppression techniques, extensively studied within constrained sampling, also serve a critical role in attribute-specific decoding. Degeneration, a frequent issue in neural text generation, is mitigated through tailored modifications to logits, token penalty matrices, or hard constraints [10]. When integrated with attribute-focused strategies, these techniques can ensure stylistic or semantic alignment while maintaining coherence and naturalness in outputs.

Advanced mechanisms enable **multi-aspect control**, addressing the challenges of integrating multiple overlapping attributes. For example, distributional intersection techniques estimate subspaces corresponding to different attributes in latent representations and optimize along their intersections, effectively isolating attribute overlaps while minimizing tradeoff conflicts [62]. While these methods hold promise for synthesizing multi-attribute outputs, they remain computationally intensive and demand further refinement to improve scalability.

Emerging frameworks seek to expand the scope of attribute-specific control by introducing fine-grained constraints at the span or location level. Locationally constrained decoding frameworks, for instance, employ auxiliary energy functions to selectively adjust token probabilities for specific text spans [58]. Such approaches show potential in use cases demanding intricate control, such as content moderation or stylistic editing, though their reliance on auxiliary models introduces additional complexities in terms of compatibility and computational feasibility.

A persistent challenge across attribute-specific decoding methods is generalizability. Many techniques optimize for narrowly defined attributes or specific datasets, limiting their adaptability to diverse tasks or domains. Additionally, the tradeoff between fluency and adherence to constraints remains a critical concern, particularly when constraints are rigid or computationally demanding. Promising solutions, such as prefix-based conditioning and dynamic soft latent adjustments, aim to enhance flexibility while minimizing disruptions to fluency, but these remain areas of active exploration [17].

Looking forward, key advancements in this field will likely center around improving scalability and robustness in multi-attribute control. Techniques integrating dynamic control graphs or reinforcement-based feedback mechanisms offer substantial opportunities for increased flexibility, particularly in interactive applications across domains such as creative writing or policy-centric systems. Furthermore, combining retrieval-augmented modules with attribute-focused decoding holds promise for enhancing factual consistency alongside stylistic alignment [48]. Finally, addressing computational costs and extending methods to multi-modal settings, where multiple input streams influence decoding, represent critical milestones for future research.

In summary, attribute-specific decoding mechanisms serve as a vital component of the broader controllable text generation pipeline, enabling precise and adaptive modifications to generated text. While existing methods provide powerful control capabilities, addressing challenges related to scalability, multi-attribute conflicts, and domain adaptability is essential for realizing their full potential in diverse real-world applications. By prioritizing these advancements, the field can unlock new possibilities for producing linguistically aligned, contextually relevant, and computationally efficient text outputs.

4.3 Knowledge-Augmented Decoding

Knowledge-augmented decoding—an essential paradigm for controllable text generation—aims to integrate external knowledge sources, contextual retrieval, or structured semantic frameworks into the decoding phase of Large Language Models (LLMs). This approach ensures enhanced factuality, coherence, and domain-specific relevance, addressing one of the fundamental limitations of pre-trained models: their static nature and susceptibility to generating hallucinated or irrelevant content. By dynamically incorporating external knowledge during decoding, knowledge-augmented methodologies provide a finer level of control without requiring extensive retraining or modification of the underlying model architecture. This subsection explores diverse methodologies for knowledge-enhanced decoding, analyzing their technical implementations, strengths, and challenges.

A predominant knowledge-augmented decoding framework is retrieval-augmented generation (RAG), which combines pre-trained models with external databases or knowledge graphs to provide contextually relevant information during decoding. RAG involves retrieving pertinent documents or knowledge snippets from an indexed corpus, which are then fused into the decoding process—typically by conditioning the model's token probabilities on both the input prompt and retrieved content [3]. By leveraging factual references from external sources, retrieval-augmented methods significantly reduce hallucination rates in high-stakes applications such as medical or legal report generation. However, the choice of retrieval mechanisms (e.g., BM25, dense retrieval) and their integration with the generation pipeline introduces computational overhead and dependency bottlenecks. Advances in hybrid RAG systems, which blend symbolic and neural retrieval, aim to mitigate this tradeoff, enhancing retrieval accuracy while managing response latency [63].

Context-aware decoding represents another vital strategy within knowledge-augmented paradigms. These methods use task-specific metadata or extended contextual inputs such as document structure, semantic relationships, or user-provided annotations to guide generation. For instance, grounding generation in highly specific task-driven metadata, such as section headings or domain-specific terminologies, ensures logical consistency across multiple levels of abstraction. Such approaches often leverage pre-trained attribute-conditioned embeddings or hierarchical attention mechanisms to maintain fidelity to external constraints while preserving fluency [29]. One promising extension is adaptive contextualization, where the model dynamically refines its understanding of long-range semantic cues during autoregressive decoding. However, ensuring temporal coherence or synthesizing multifaceted metadata into a unified representation remains challenging, particularly in long-form or multi-task generation contexts.

More structured implementations of knowledge-augmented decoding utilize trie-based constraints or scoring systems derived from knowledge graphs to ensure logical consistency and entity fidelity in generated outputs. Trie structures, for example, limit the token search space during decoding to ensure that selected paths align with predefined factual or syntactic constraints. Additionally, constraint-aware scoring functions—whether derived from neural classifiers or symbolic systems—allow for token probability re-ranking to enforce relevance and alignment. Dynamic attribute graphs (DAGs) further enhance this capability by representing decoding objectives as evolving graphs that adapt to domain-specific contexts, enabling simultaneous balancing of multiple constraints such as factuality, sentiment, and stylistic coherence [12]. While effective, these methods introduce scalability concerns, as maintaining the graphs or trie constraints in real time demands substantial computational resources.

One particularly novel frontier in knowledge-augmented decoding is probabilistic constraint modeling, often implemented through energy-based models or latent-space optimization during token selection [36]. These methods frame decoding as solving a constrained optimization problem, where external knowledge constraints form part of an energy function that assigns scores to candidate sequences. By iteratively minimizing the energy function during generation, these approaches maintain both fluency and compliance with control objectives. Such optimization techniques are highly flexible, accommodating both "soft" (e.g., stylistic relevance) and "hard" (e.g., factual correctness) control criteria. However, the non-convexity of these constrained optimization problems increases computational complexity, particularly for tasks requiring fine-grained control across multiple attributes.

Despite substantial progress, knowledge-augmented decoding systems face critical challenges. First, the seamless integration of external knowledge sources without overwhelming the generation process remains an unresolved issue. Effective filtering and prioritization mechanisms to balance external inputs with model-internal representations continue to be areas of active exploration [7]. Further, the reliance on static knowledge bases limits adaptability to evolving domains or real-time contexts. Emerging techniques that incorporate retrieval-augmented updates alongside memory constructs aim to address this issue, as seen in models using residual memory transformers to encode dynamic factual updates. Additionally, the trade-off between fluency, factuality, and computational efficiency introduces quantifiable yet competing demands—a balance most methodologies struggle to achieve.

Future directions for this domain include developing modular plug-and-play frameworks capable of integrating multiple external sources seamlessly, advancing low-latency retrieval techniques for real-time knowledge incorporation, and exploring hybrid methods that unify symbolic and latent decoding constraints. Adaptive hybrid configurations, where energy-based guidance intersects with retrieval-augmented pipelines or DAG-based dynamic constraints, hold particular promise. Finally, explainable knowledge-augmented decoding—where models provide insight into decisions about knowledge incorporation during generation—represents a critical step for research into accountability and robustness.

In summary, integrating external knowledge into decoding phases provides powerful tools for enhancing the factuality, context relevance, and logical coherence of LLM generations. However, achieving a scalable, interpretable, and adaptive framework remains a tantalizing challenge at the forefront of controllable text generation research.

4.4 Energy-Based and Constraint Optimization Frameworks

Energy-based and constraint optimization frameworks represent cutting-edge methodologies in decoding for controlled text generation, leveraging probabilistic constraints and score-based objectives to enforce fine-grained control over outputs. These approaches are particularly effective in scenarios requiring adherence to complex and multifaceted constraints, offering a probabilistic alternative to hard-rule enforcement. At their core, they optimize decoding processes to align with desired outputs while trading off fluency, diversity, and attribute-specific requirements in a mathematically principled manner.

Building upon the foundational principles of knowledge-augmented decoding, energy-based models (EBMs) reframe controlled text generation as an optimization problem over an energy function that evaluates the quality of a sequence. The energy function encapsulates various desired properties, including fluency, attribute alignment, and logical consistency. These frameworks have gained prominence due to their robustness in incorporating both pointwise and distributional constraints. For instance, a distributional framework optimizing constraints based on KL divergence from a pretrained model's prior distribution has shown efficiency in balancing constraint satisfaction with minimal disruption to linguistic fluency [37]. EBMs offer the flexibility to integrate complex attribute requirements and composite constraints by defining multi-dimensional energy landscapes. Despite their adaptability, their computational cost can be significant, as energy minimization requires iterative optimization during decoding, often using techniques such as Metropolis-Hastings sampling [45] or gradient-based updates.

Dynamic constraint optimization frameworks extend these principles by formalizing decoding as a constrained optimization problem, where outputs must satisfy one or multiple predefined constraints. These methods often utilize Lagrangian multipliers or penalty-based approaches to dynamically enforce soft constraints. For example, MuCoCO models the decoding process as a relaxed continuous optimization problem where differentiable constraints are incorporated, enabling seamless alignment of outputs with attributes such as sentiment or syntactic structure [36]. This methodology is particularly effective in tasks like content detoxification, where balancing fluency and non-toxicity is critical. While these frameworks reduce reliance on fine-tuning large models, real-time applications face significant challenges due to computational intensity.

Graph-based representations and sequential constraint satisfaction mechanisms further expand the scope of controlled generation by enabling adaptable and evolving constraints within dynamic contexts. For instance, dynamic constraint graphs steer token probabilities at each step to meet multivariate control objectives, outperforming static execution frameworks in flexibility and adaptability [12]. Similarly, sequential constraint satisfaction has demonstrated efficacy in iteratively optimizing token distributions to align with overlapping constraints such as factuality, coherence, and style [16]. These mechanisms benefit from logical foresight and iterative correction, ensuring global consistency across attributes. However, scalability remains a pressing challenge, especially when satisfying tightly interdependent constraints across diverse attribute dimensions.

EBMs and dynamic constraint optimization methods mark a paradigm shift toward achieving fine-grained control in high-dimensional spaces, especially in scenarios where attribute interdependencies complicate decoding. For example, penalty-based dynamic optimization frameworks have demonstrated improved multi-aspect control across dimensions such as sentiment, diversity, and alignment to user-defined styles [24]. While these strategies enhance controllability and flexibility, they often face trade-offs between decoding efficiency and computational overhead.

Emerging trends suggest promising directions to address these trade-offs. Incorporating pre-trained knowledge models or external attribute classifiers into energy components could improve both accuracy and constraint adherence while minimizing computational expenses [58]. Additionally, hybrid strategies combining EBMs with reinforcement learning could further refine energy functions for domain-specific control objectives, offering a path toward more targeted optimizations. Sampling-free approaches that leverage differentiable surrogates for energy functions present another intriguing avenue for reducing computational latency. As text generation scenarios grow in complexity, these frameworks will continue to evolve, addressing challenges in scalability, adaptability, and efficiency while balancing quality with user-driven requirements.

By bridging the granular constraint mechanisms of knowledge-augmented decoding with the adaptability of lightweight plug-and-play techniques explored in the next subsection, EBMs and constraint optimization frameworks provide a critical middle ground. They balance fluency and flexibility, pushing the boundaries of controllable text generation while inviting further innovation in reconciling computational efficiency with dynamic, multi-attribute control.

4.5 Lightweight Plug-and-Play Decoding Techniques

Lightweight plug-and-play decoding techniques enable dynamic and flexible control in text generation at inference time without requiring retraining or fine-tuning of the underlying large language models (LLMs). These methods are particularly appealing due to their resource efficiency, allowing practitioners to tailor outputs based on various constraints while preserving the base model's general capabilities. This subsection examines key approaches, their technical strengths, trade-offs, and emerging trends.

Plug-and-play methods often rely on external mechanisms to adjust the token probabilities during decoding, circumventing the need to modify the model's parameters. A foundational example is the use of external classifiers to align output distributions with desired attributes, such as tone or sentiment [64]. By injecting attribute scores into the decoding process dynamically, external classifiers provide models with external control signals that constrain generation in a lightweight and modular manner. For instance, external sentiment classifiers can penalize token probabilities deviating from the target sentiment. Although effective, reliance on pre-trained classifiers introduces trade-offs: these methods are sensitive to inaccuracies or biases within the external classifier, posing risks of unintended control artifacts.

Another prominent approach is prefix or bias tuning, where lightweight modifications are applied to the token distribution without intervening in the generation architecture. Bias is introduced either by priming the decoding process with special prefix embeddings or by modifying logits of target-specific tokens during inference. Unlike retraining-heavy methods, prefix tuning operates effectively as a low-resource solution for steering outputs to align with desired attributes [60]. These techniques are particularly advantageous in low-data and multi-attribute scenarios. For example, a prefix tuned on style-specific prompts can adapt a generative model to switch fluidly between varying stylistic constraints. Despite their efficiency, prefix tuning's performance can degrade when multiple, competing attributes are involved, as the engineered distribution shifts may fail to balance conflicting requirements empirically.

Low-Rank Adaptation techniques (LoRA) have also been extended for decoding control, where pre-trained frozen layers are combined with task-specific low-rank matrices applied only during inference [12]. In these setups, LoRA introduces soft, attribute-specific constraints to token selection without modifying the encoded knowledge of the LLM. For instance, LoRA has been shown to efficiently adapt to dynamic attributes such as formality and length constraints without compromising computational scalability. However, LoRA often relies on careful pre-specification of low-dimensional control embeddings, which can be less optimal when the target constraints are non-stationary or emergent during decoding.

Contrastive decoding techniques present another lightweight framework for refining outputs based on dynamic constraints. By penalizing tokens associated with attribute distributions outside the target scope, contrastive methods can steer outputs away from undesired behaviors, such as repetition or hallucination [65]. Unlike sampling-based methods such as nucleus or top-k sampling, contrastive decoding balances fluency preservation with attribute alignment by optimizing dynamic logits directly during token generation. However, this design, while effective in reducing degeneration, may introduce inefficiencies in decoding latency, as the iterative penalties require real-time evaluation.

Recent advancements are exploring methods that leverage energy-based formulations to define explicit constraints as part of the decoding objective. For instance, Locate&Edit enables targeted span modifications via energy minimization techniques, selectively revising constraint-violating spans while preserving the semantic coherence of untouched sections [58]. This aligns output quality with multiple overlapping constraints, offering a practical mechanism for modular plug-and-play control. A complementary framework, Activation Addition (ActAdd), further optimizes output bias by directly modifying forward-pass activations at specific inner layers. Calculated by contrasting the activation patterns of exemplar input-target pairs, this technique offers nuanced steering capabilities with minimal overhead [66].

The flexibility and efficiency of these plug-and-play techniques raise important considerations. While they circumvent the computational burdens of retraining, their dependency on auxiliary modules and heuristic-guided constraints can introduce instability when applied under diverse and high-dimensional attributes. Moreover, their scalability across multimodal contexts (e.g., text paired with images or audio) remains a pertinent avenue for future work. Emerging trends propose combining lightweight methods with retrieval-augmented generation (RAG) systems, where retrieval modules provide external knowledge constraints during decoding [50]. This hybrid approach holds promise in enhancing attribute fidelity across dynamic contexts without overloading the decoding pipeline.

In summary, lightweight plug-and-play decoding strategies exemplify the promise of minimally invasive, resource-optimal methods for enhancing generation fidelity and control. However, their effectiveness hinges on balancing computational efficiency against the fidelity and robustness of constraints. Future research should explore adaptive frameworks capable of addressing simultaneous, multi-attribute constraints while extending these techniques into dynamic, multimodal, and interactive domains.

4.6 Multi-Aspect and Interactive Control at Decoding

Achieving multi-aspect and interactive control during decoding represents an advanced frontier in controllable text generation and serves as a natural extension of lightweight plug-and-play approaches. This paradigm focuses on satisfying multiple constraints—such as sentiment, topic, style, and structure—while accommodating dynamic user feedback to refine outputs in real time. By leveraging the flexibility of the decoding phase, multi-aspect control enables efficient solutions without requiring retraining or fine-tuning, making it particularly relevant for resource-constrained or rapidly evolving scenarios. This subsection delves into contemporary methodologies, evaluates their strengths and limitations, and highlights emergent trends in achieving sophisticated, interactive control capabilities.

A foundational strategy for multi-aspect control during decoding involves incorporating multiple attribute-specific guidance mechanisms directly into the token selection process. Techniques such as FUDGE [7] exemplify this approach, predicting the likelihood of future tokens satisfying desired attributes and iteratively modifying the probability distribution during decoding. This modularity enables FUDGE to handle multi-aspect tasks by composing multiple discriminators to balance competing constraints. Nevertheless, its dependence on future token predictions can introduce latency, posing bottlenecks for real-time applications—a trade-off that underscores the challenge of achieving simultaneous control and efficiency.

Complementing this, distributional approaches provide a mathematically rigorous framework for achieving compositionality across multiple control aspects. For example, the Distributional Lens [37] formulates text generation as a latent space optimization problem, minimizing the Kullback-Leibler (KL) divergence between controlled and original distributions while incorporating multiple constraints. This ensures a balance between fluency and adherence to attributes. However, as the number of entangled constraints increases, distributional methods can face scalability issues, highlighting the need for innovations in latent space structuring and efficient sampling mechanisms to alleviate computational bottlenecks.

Interactive control frameworks have also gained prominence, building on lightweight plug-and-play methodologies to introduce user guidance during decoding. Plug-and-Play Language Models (PPLM) [4] exemplify such approaches by steering pretrained language models toward desired attributes using external classifiers, without altering the core architecture. The plug-and-play design provides significant flexibility, supporting iterative user input to dynamically modify decoding trajectories. However, while adaptive and lightweight, methods like PPLM often struggle to balance attribute alignment with text fluency, particularly in multi-aspect settings demanding fine-grained control.

Reinforcement-based decoding methods offer a promising avenue for managing competing constraints by dynamically weighting attributes according to contextual needs. Reinforcement Learning with Dynamic Multi-Reward Weighting [67] incorporates multiple style-specific rewards to optimize the generation process, ensuring that dominant attributes do not overshadow weaker ones. This framework is particularly relevant for addressing attribute suppression issues common in multi-aspect control. However, its computational demands render real-time application a formidable challenge, especially when dealing with overlapping constraints in dynamic settings.

Additionally, novel methods such as MAGIC [68] have advanced interactive control by modeling the evolution of attributes during decoding. MAGIC employs successor features to disentangle attribute spaces, enabling both preemptive adjustments and real-time user guidance. Counterfactual augmentations further enhance attribute adherence by providing flexibility to refine outputs on-the-fly. Nevertheless, scalability to high-dimensional feature spaces remains an open problem, particularly when handling diverse or heterogeneous datasets—a recurring issue in interactive systems.

Despite these advancements, achieving robust and scalable multi-aspect control during decoding remains challenging. A significant complexity arises from ensuring compatibility among conflicting control dimensions, especially when balancing fluency with diverse constraints. Methods grounded in compositional latent dynamics [69] offer potential solutions by resolving these conflicts via Ordinary Differential Equation (ODE) solvers. However, such approaches are computationally intensive, limiting their practicality for high-speed, real-time applications. Moreover, methods that emphasize control precision often compromise adaptability, while those prioritizing flexibility may degrade text quality—highlighting the delicate trade-offs inherent in multi-aspect control.

Interactive control further complicates the landscape by requiring systems to adapt fluidly to evolving constraints while maintaining fluency and coherence. Lightweight plug-and-play extensions [70] and sampling-based refinement mechanisms [58] represent strides toward bridging this gap. However, optimizing for both computational efficiency and user satisfaction under dynamically shifting constraints remains an ongoing research priority.

Future research must converge on hybrid frameworks that synthesize the interpretability of causal models [25] with the performance scalability of distributional optimization techniques. Additionally, advancements in multimodal integration, inspired by text-to-image generation [71], can inform novel methods to harmonize diverse constraints in textual generation. Addressing these challenges will pave the way for robust, responsive, and adaptable multi-aspect control paradigms, ultimately extending the utility of controllable text generation to diverse and interactive applications.

5 Evaluation and Metrics for Controllable Text Generation

5.1 Controllability-specific Evaluation Metrics

Evaluating the efficacy of controllable text generation (CTG) systems necessitates specialized metrics that quantify adherence to predefined attributes, styles, and structural or task-specific constraints. This subsection critically reviews existing controllability-specific evaluation metrics, their applications, limitations, and emerging challenges, providing a foundation for their refinement and broader adoption.

A cornerstone of controllability evaluation lies in **attribute fidelity metrics**, which measure the system's capacity to generate outputs conforming to specified attributes (e.g., sentiment, tone) or constraints (e.g., factuality or keyword inclusion). For binary or categorical attributes like sentiment, classifiers are often used to validate alignment, as demonstrated in methods such as Plug and Play Language Models [4]. These classifiers evaluate whether generated outputs exhibit the desired attribute while remaining semantically coherent. However, classifier-based metrics often struggle with nuanced attributes like emotional complexity or hybrid attributes, where manual evaluation or task-specific human feedback remain indispensable [44]. Moreover, classifier performance is limited by its dependence on training data, often leading to embedded biases that cascade into evaluation results, a critical challenge observed in studies like [72]. Future methodologies must prioritize robust attribute classifiers, particularly for multi-attribute control tasks.

Stylistic adherence assessments—focused on evaluating nuanced stylistic dimensions such as formality, politeness, or creativity—often leverage pre-trained style classifiers or embedding-based scoring models [3]. These metrics quantify the model's alignment with the stylistic spectrum defined during training or prompt construction. Examples include cosine similarity within embedding spaces for stylistic evaluation or Fine-tuned LMs (FLMs) for scoring [17]. While these approaches are effective for static and clearly defined styles, they often struggle with dynamic or emergent styles, particularly in creative settings such as poetry or narrative generation [26]. Interestingly, contrastive learning techniques have proven valuable for style transfer scenarios by penalizing undesirable artifacts and rewarding adherence to stylistic targets during evaluation [7]. These approaches, while effective, are typically resource-intensive, necessitating further optimization for real-time applications.

Systems requiring compliance with **constraint-specific metrics**—like factual accuracy or structural adherence—have benefited from targeted measures such as BLEU-style metrics for structural constraints or retrieval-augmented benchmarks for factuality [18]. For instance, factuality-focused metrics employ retrieval-based tools to check whether output content aligns with trusted external knowledge, including AIS (Attributable to Identified Sources) evaluation frameworks [18]. Similarly, keyword-specific control has been effectively quantified using constrained beam-search inventories and metrics that penalize omissions of specified entities [10]. Nevertheless, limitations persist in balancing such hard constraints with linguistic fluency; rigid adherence to keywords or structure often results in stilted or unnatural outputs, a common challenge highlighted in [16].

Another facet is the trade-off between adherence to control parameters and overall text quality. Metrics such as CTRLEval, an unsupervised and reference-free evaluation scheme, assess multiple dimensions of controlled generation, including fluency, adherence, and semantic overlap [13]. These multi-dimensional evaluation techniques offer a more holistic view of controllability but lack explainable mechanisms for failure cases, inhibiting their utility in dissecting errors.

As the field advances, emerging challenges include **evaluation inconsistencies in multi-attribute generation tasks**, where metrics must capture interactions between control objectives. For instance, systems tasked with integrating sentiment, tone, and thematic constraints simultaneously often face trade-offs where improvements in one dimension degrade others [45]. Addressing these interdependencies requires developing metrics capable of balancing diverse objectives, such as those relying on energy-based modeling or dynamic scoring approaches [6].

In conclusion, while significant progress has been made in evaluating controllability, key challenges remain. Diverse metrics rooted in machine-learning classifiers and retrieval-augmented models have demonstrated potential for single-attribute tasks. However, as tasks grow complex—encompassing multiple dimensions, emergent styles, and real-time constraints—future work must prioritize scalable, low-latency, and interpretable metrics. Further, robust frameworks incorporating human-machine hybrid evaluation, such as CoEval-style systems [73], offer promising directions for refining controllability assessments. Expanding benchmarks and aligning metrics with user-centric objectives will be pivotal in bridging current gaps and enabling more reliable and expressive evaluations in controllable text generation.

5.2 Output Quality Evaluation

Output quality evaluation in controllable text generation focuses on assessing the naturalness, fluency, coherence, and grammatical correctness of generated text, while balancing adherence to control constraints. Building on the evaluation challenges discussed earlier, this subsection delves into state-of-the-art approaches and metrics used to evaluate the output quality of controllable text generation systems, emphasizing their strengths, limitations, and emerging trends. Ensuring high-quality outputs is not only vital for the functionality of these systems but also key to their acceptance in real-world applications.

A cornerstone of output quality evaluation is the assessment of **fluency and grammaticality**, which involves measuring the syntactic correctness and linguistic naturalness of generated outputs. Traditional metrics, such as perplexity—evaluating the likelihood of text under a language model—remain integral for fluency analysis. However, perplexity often proves inadequate at capturing nuanced aspects of human-like naturalness or domain-specific language standards. While BLEU-style adaptations for fluency evaluations have been proposed, they similarly struggle in assessing the fluidity of unreferenced, open-ended text [13]. Specialized tools like Grammarly or linguistically-informed discriminative models can provide deeper syntactic and grammatical insights, but their computational overhead and scalability limitations hinder their broader applicability. Advancements, such as unsupervised and reference-free metrics exemplified by CTRLEval, have been developed to improve efficiency and adaptability, leveraging text infilling tasks with pre-trained language models to evaluate fluency without relying on task-specific training data [13].

Another critical dimension in output quality evaluation is **semantic coherence**, especially in long-form or multi-turn text generation. Maintaining logical consistency across sentences, paragraphs, or conversational turns continues to challenge many models, particularly those constrained by multiple control dimensions. Embedding-based coherence metrics, such as sentence-level cosine similarity or graph-based coherence frameworks, have been explored to address this. Hierarchical language models and methods like MEGATRON-CNTRL, which utilize keyword prediction and contextual knowledge ranking, demonstrate enhanced local and global coherence in generative outputs [74]. However, subtle thematic or rhetorical coherence, like maintaining metaphors or evolving narratives across generations, remains difficult to evaluate quantitatively, necessitating further research into context-sensitive coherence evaluation strategies.

To expand beyond correctness and coherence, **diversity and novelty** evaluations aim to measure the originality of generated outputs while safeguarding alignment with control goals. Established techniques, such as Self-BLEU, assess intra-sample redundancies, helping to balance creativity with constraint satisfaction. Innovative approaches like Composition Sampling employ intermediate representations (e.g., entity chains or structured plans) to foster diversity while preserving naturalness, proving effective in conditional generation tasks like summarization and question generation [22]. Similarly, strategies for mitigating repetition and degeneration during decoding, such as those in NeuroLogic Decoding, have been integral to improving linguistic variation without impairing text fluidity [10].

Recent advancements suggest that **large language models (LLMs)** themselves can play a dual role as both generators and evaluators of output quality. GPT-Eval-style frameworks, for example, leverage the interpretive capacities of models like GPT-4 to assess text quality across multiple dimensions. These frameworks have demonstrated strong correlations with human judgment, yet they remain prone to biases embedded in the LLMs' training data and often lack transparency in how evaluations are conducted [48]. Hybrid strategies that combine human input with LLM-based assessments show promise for capturing intricate quality dimensions such as creativity and contextual relevance, thus making evaluations more reliable in user-centric scenarios [5].

Nonetheless, significant challenges persist, particularly in reconciling **high output quality with strict adherence to control constraints**. Techniques emphasizing rigid content fidelity, such as retrieval-augmented frameworks or constraint-specific decoding strategies, often compromise linguistic fluidity by prioritizing factual accuracy or structural adherence over naturalness [47], [50]. Furthermore, subjective qualities, such as creativity or emotional resonance, remain difficult to quantify using existing metrics, particularly for domains where user preferences or cultural factors dictate notions of quality [75].

To address these challenges, future directions in output quality evaluation must focus on developing **universal, task-agnostic benchmarks** capable of accommodating diverse evaluation criteria, combining fluency, coherence, diversity, and constraint satisfaction into multi-dimensional metrics. Emerging approaches, such as multi-aspect evaluation pipelines enriched with causal explainability and user feedback loops, hold promise for creating interpretable and robust evaluation ecosystems [62]. Efforts to quantify and align subjective human preferences, especially for creative and domain-specific use cases, will be instrumental in bridging the gap between technical metrics and real-world expectations. These developments will pave the way for scalable and nuanced assessment frameworks that harmonize automation, interpretability, and user experience in controllable text generation systems.

5.3 Human-centric Evaluation Techniques

Human-centric evaluation techniques occupy a critical role in assessing the quality and effectiveness of controllable text generation systems, as they provide insights that purely automated metrics often overlook. Unlike algorithmic evaluations, which focus on attributes like fluency, diversity, or adherence to constraints, human-centric methods anchor evaluation on subjective, contextual, and task-specific assessments encompassing user satisfaction, interpretability, and real-world usability. This subsection explores the breadth and depth of human evaluations, their methodologies, and their integration with automated metrics, while highlighting strengths, limitations, and emerging challenges.

Human evaluations are essential for capturing nuanced aspects of text generation, including the appropriateness of tone, contextual relevance, and emotional resonance, which automated systems struggle to quantify reliably [76]. A common approach involves Likert-scale ratings or ranking-based methods, where human annotators score outputs on attributes such as readability, naturalness, and alignment with control objectives [44]. While effective for granular assessments, such methods are labor-intensive, subject to variability among raters, and challenging to scale. Hybrid approaches that combine human scores with algorithmic evaluations, such as GPT-Eval frameworks, have emerged as promising solutions to balance cost and scalability [77]. These hybrid systems often achieve high inter-rater agreement by calibrating human judgments against LLM-based evaluators, but they risk inheriting biases from the underlying language models.

One notable dimension of human evaluations lies in task-specificity. For example, creative writing systems require assessments of originality, emotional depth, and thematic coherence, whereas domain-specific tasks like legal or medical content generation require strict evaluations for factual accuracy and terminological precision [78]. The variance in evaluation criteria across tasks necessitates a flexible, user-centric framework for scoring, which some studies have addressed through specialized benchmarks such as StylePTB [79]. In such settings, combining task-relevant metrics with subjective human ratings enables comprehensive evaluation pipelines.

Another dimension is the temporal adaptability of evaluations in iterative systems. Real-time feedback mechanisms, where users actively participate in refining outputs, have gained attention for enhancing alignment with immediate user needs. These systems, often leveraging plug-and-play methods [80], allow human evaluators to interact with text as it evolves, offering feedback that refines controllable attributes dynamically. However, the trade-offs include increased evaluation latency and the cognitive burden on evaluators tasked with continuously monitoring and revising outputs.

Emerging trends in human-centric evaluation address the complexity of multi-attribute and multi-task text generation systems. General evaluative challenges arise when assessing trade-offs among competing control dimensions, such as balancing sentiment transformation with content preservation [47]. Tools that assist human evaluators, such as LLM-assisted annotation interfaces or hierarchical scoring mechanisms, show promise in tackling such complexity, but their efficacy depends heavily on reducing evaluator biases and ensuring interpretability [45].

Integration with automated metrics represents another frontier in human-centric evaluation. Techniques like contrastive conditioning [21] or control-theoretic frameworks [53] can refine the relationship between human-provided feedback and machine-learned evaluations, optimizing alignment in controllable text generation models. Additionally, research into model-guided evaluations, where LLMs themselves act as proxies for human judgment [72], offers scalable prospects but raises concerns regarding transparency and reliability.

An enduring challenge lies in the subjectivity inherent to human-centric evaluations. Variability among annotators, contextual dependencies, and cultural biases can skew results. Recent studies have explored causally-guided evaluation frameworks [25] to mitigate these issues, employing causal inference mechanisms to identify latent confounders in annotation patterns. Further work must also explore cross-lingual frameworks for human evaluation, given the growing prevalence of controllable generation tasks in multilingual domains [81].

In conclusion, human-centric evaluation techniques are indispensable for advancing controllable text generation systems, providing the interpretative richness and contextual sensitivity that automated metrics lack. While hybrid and iterative approaches embody transformative potential, future research must emphasize scaling these evaluations while mitigating subjectivity and bias. Emerging methods, such as causality-driven assessment frameworks and LLM-facilitated annotation, suggest scalable paths forward but require rigorous empirical validation. Creating universal, flexible benchmarks that fuse human intuition with machine precision will underpin the next generation of robust and user-aligned controllability systems.

5.4 LLM-based Evaluation Techniques

Large Language Models (LLMs) have emerged not only as generative tools but also as pivotal evaluators for controllable text generation (CTG) tasks. Leveraging their sophisticated natural language understanding, these models serve as adaptable frameworks for assessing control adherence, fluency, and generation quality. This subsection examines the methodologies, merits, limitations, and potential advancements in using LLMs as evaluators, situating them within the broader context of hybrid and human-centric evaluation paradigms.

As evaluators, LLMs bring unmatched versatility, enabling evaluation across multiple dimensions of controllable text generation. A notable application lies in their ability to assess alignment with control attributes like sentiment polarity, formality, or stylistic constraints. For instance, frameworks such as GPT-Eval employ pretrained LLMs to measure fluency, style adherence, and fidelity to control objectives, often leveraging embeddings or attention-based similarity measures [12]. Prompt-based evaluation, wherein specific task instructions guide the LLM to produce scores, rankings, or qualitative feedback, exemplifies their flexibility. However, this very adaptability introduces a sensitivity to prompt phrasing that can inadvertently skew results, highlighting the need for cautious application [44].

Crucially, LLM-based evaluators excel in scenarios demanding multi-faceted assessments without requiring task-specific retraining. For example, weakly supervised, instruction-tuned frameworks like InstructCTG demonstrate how LLMs can evaluate attributes such as sentiment accuracy and content fidelity through encoded constraints [43]. Similarly, role-player prompting, wherein models are guided to simulate diverse evaluative perspectives, enables rich judgments across subjective and objective criteria, encompassing coherence, creativity, and contextual appropriateness [82]. These flexible methodologies illustrate the scalability of LLM evaluation, particularly in contexts where traditional metrics such as BLEU or ROUGE fall short in capturing nuanced dimensions of control.

Despite their potential, LLM-based evaluation approaches face significant challenges rooted in their black-box nature, biases, and inconsistent reliability. The lack of explainability concerning how LLMs generate scores or make evaluative judgments often undermines transparency. Biases embedded in pretraining datasets can also influence evaluative performance, particularly for sensitive control attributes involving fairness, toxicity, or cultural context [25]. Such biases could unintentionally propagate into stylistic assessments, especially when control parameters are underrepresented or misrepresented in the datasets. Additionally, the trustworthiness of LLM evaluators can be compromised by their susceptibility to subtle variations in prompt phrasing, calibration, or task framing, often leading to inconsistent outcomes [83].

To mitigate these limitations, hybrid evaluation frameworks blending LLM-driven feedback with human judgment have been developed. For instance, methods such as CoEval prioritize achieving scalability through LLM evaluations while maintaining fidelity by incorporating human oversight for subjective dimensions [44]. Strategies like role-player prompting enrich the evaluation process by enabling dynamic, context-sensitive profiles that assess outputs across multiple criteria [82]. Moreover, post-hoc calibration techniques—such as anchored scoring and consistency regularization—offer potential solutions to improve LLM-based evaluations' robustness and reproducibility [83].

Emerging research directions highlight a growing emphasis on enhancing the transparency, task-specific adaptability, and standardization of LLMs in evaluation roles. Explainable evaluation protocols grounded in causal modeling represent a promising avenue for demystifying how LLMs reach their decisions, fostering trust through auditable processes [25]. Similarly, adaptive calibration frameworks that align LLM scoring mechanisms with evolving task-specific constraints or dynamic dataset attributes seek to improve their utility in real-world applications [12]. Concurrently, efforts to standardize benchmarks, such as CoDI-Eval, are pivotal in formalizing evaluation practices for fine-grained control and addressing the complexities of diverse CTG scenarios [84].

In summary, large language models offer significant promise as scalable, adaptable evaluators for controllable text generation, complementing both automated and human-centric evaluation paradigms. However, challenges surrounding consistency, transparency, and bias must be critically addressed for these techniques to reach their full potential. Future research should prioritize integrating LLMs within hybrid, interpretable evaluation pipelines while advancing task-specific calibration methods. By leveraging innovative benchmarks and fostering human-machine collaboration, LLM-based evaluation frameworks can achieve greater reliability, fairness, and standardization, driving progress in benchmarking CTG systems responsibly and comprehensively.

5.5 Holistic Benchmarking and Evaluation Challenges

Holistic benchmarking and evaluation of controllable text generation (CTG) present intricate challenges due to the diverse nature of control attributes, contextual dependencies, and the dynamic application-specific requirements. While existing benchmarks and metrics have made significant strides in quantifying controllability and generation quality, several critical gaps persist that hinder comprehensive evaluations across various dimensions of control, application domains, and subjective criteria.

A central issue in benchmarking CTG lies in accommodating **multi-dimensional evaluation**. Traditional evaluation pipelines often focus on isolated metrics such as fluency, coherence, and attribute fidelity, but fail to integrate multi-aspect evaluations in a systematic fashion. For example, tasks such as style transfer or factuality control demand simultaneous optimization across dimensions like stylistic adherence, semantic preservation, and factual accuracy. However, existing benchmarks lack the granularity required to rigorously assess trade-offs between these objectives. Recent efforts like ConGenBench [85], which orchestrates a testbed spanning 17 CTG tasks, highlight the potential of consolidated evaluations but reveal stark limitations in addressing structural tasks, as prompting-based approaches diverge from human-level performance on these dimensions.

Moreover, challenges in **task-specific adaptability** arise from the inherent subjectivity of many control attributes. For instance, evaluating creativity, humor, or emotional intensity introduces ambiguities not easily captured by rule-defined metrics. While hybrid evaluation models combining large language models (LLMs) as evaluators and human assessments have grown in popularity [82], questions about the objectivity, bias, and reproducibility of LLM-based evaluations remain pervasive. Studies have shown that these models are susceptible to prompt-induced variability and context-specific biases, raising concerns about their consistency compared to human evaluations [86].

The contextual and **interactive nature of controllable tasks** adds another layer of complexity to benchmarking frameworks. For tasks requiring dynamic control, such as real-time sentiment adaptation or user-adjustable narrative generation, static benchmark datasets and one-shot metrics are insufficient. Incorporating **dynamic evaluation pipelines**, as explored in frameworks like Self-Refine [38]—which iteratively improve outputs through feedback—might address the fluidity required in such scenarios. However, scaling these evaluations across diverse CTG dimensions and tasks demands modular yet contextually adaptive benchmarks.

Comparative studies further underscore the tension between **automatic and human-centric evaluations**. Metrics such as BLEU or perplexity inadequately capture nuanced stylistic controls or relational dynamics in conversational AI, often leading to discrepancies with human judgments [83]. Advances in contrastive decoding techniques [87] and retrieval-augmented setups [50] underline the importance of grounding evaluations in meaningful contexts. Nevertheless, the absence of standardized benchmarks that weigh factual grounding alongside stylistic, structural, and task-specific fidelity complicates efforts to produce generalizable and scalable assessment frameworks.

Emerging trends also expose **gaps in cross-lingual and cultural adaptability** of CTG benchmarks. Language models often exhibit biases in multilingual settings, where controllability suffers due to limited resources for low-resource languages. Benchmarks tailored to diverse linguistic and cultural contexts, such as attribution-driven evaluations or fine-grained cultural alignment tasks, remain underdeveloped [88]. Similarly, the integration of multimodal control further accentuates the need for benchmarks that holistically evaluate alignment across text, image, or video-based modalities [41].

To address these challenges, **comprehensive benchmark frameworks** must evolve, focusing on integrating multifaceted objectives, mitigating evaluator inconsistency, and adapting to emergent tasks, such as multimodal or low-resource generating contexts. Task-specific benchmarks like ConGenBench are important stepping stones, but their limitations, particularly in dynamic and structural evaluations, call for advancements that incorporate modular and real-time adaptability. Moreover, LLM-based evaluators should be augmented with interpretable and causality-driven frameworks [25], ensuring both transparency and scalability. Future directions should prioritize **universal benchmarks** with context-aware, multidimensional metrics that standardize evaluation processes while maintaining the flexibility to accommodate the rapidly expanding boundaries of CTG applications.

5.6 Towards Standardization and Explainability in Evaluation

The standardization and explainability of evaluation protocols are pivotal to driving advancements in controllable text generation (CTG). While considerable progress has been made in defining metrics and evaluation methods tailored to specific generation tasks, the fragmentation of these approaches poses challenges for reproducibility, comparability, and interpretability across the field. This subsection delves into ongoing efforts to establish standardized practices and explainable evaluation frameworks, highlighting current methodologies, emerging trends, and unresolved challenges.

The inherent diversity and multifaceted nature of CTG tasks complicate evaluation, as they require assessing both overall text quality (e.g., fluency, coherence) and compliance with localized attributes (e.g., sentiment, style, factuality). Historically, evaluation efforts have relied on task-specific metrics, such as BLEU for content fidelity and perplexity for fluency, which inadequately capture more nuanced control dimensions. For instance, metrics assessing attribute fidelity often depend on classifier-based evaluations, as seen in approaches using external attribute scorers [4], [12]. While these methods streamline automated assessment, they face significant interpretability challenges due to the opacity of decision boundaries and feature importance within classifiers. Moreover, the reliance on potentially biased training data in classifiers risks perpetuating unethical trade-offs between attribute fidelity and fairness, as identified in [25]. This has led to growing calls for explainability techniques to be embedded into classifier-based evaluations [25].

Hybrid evaluation frameworks, blending automated metrics with human feedback, have emerged as promising alternatives to capture complex quality indicators like creativity, humor, or stylistic nuances. For instance, [43] demonstrates the value of pairing instruction-tuned language models with human raters to assess stylistic and compositional controls. However, scalability and cost remain limiting factors for human evaluations, prompting the development of hybrid models like GPT-Eval. These models leverage large language models as scalable stand-ins for human evaluators, approximating human preferences with reasonable fidelity, as shown in [89]. Still, their dependency on pre-trained priors raises issues around calibration, trustworthiness, and robustness, especially in tasks requiring precise factual adherence.

Central to tackling the evaluation challenges in CTG is the pursuit of standardization—both in benchmark datasets and evaluation protocols. Emerging datasets aim to reflect a wide spectrum of constraints, exemplified by resources like StylePTB [79], which enable fine-grained assessments for compositional style transfer. Similarly, [85] advocates for task-agnostic testbeds that encompass a broader array of control attributes. Comprehensive platforms like ConGenBench offer essential tools for cross-model comparisons and performance consistency across diverse tasks. However, the proliferation of benchmarks raises concerns about overfitting to narrowly defined scenarios, limiting generalizability.

Explainability in evaluation integrates interdisciplinary innovations to enhance interpretability and mitigate bias. Techniques such as post-hoc model analyses, including attention visualizations and counterfactual reasoning, are becoming valuable tools for interpreting evaluation metrics specific to CTG outputs [25]. Structural causal models have also emerged as robust evaluative frameworks for disentangling the relationships between inputs, control attributes, and outputs, exposing latent biases and improving metric reliability [25]. Furthermore, energy-based models, as applied in [58], directly map performance on constraint satisfaction to interpretable energy functions, offering fine-grained insights into attributes like fluency and style deformations.

Despite these developments, pressing challenges persist for achieving standardized and explainable CTG evaluation. Firstly, there is a clear need for modular, task-agnostic evaluation pipelines adaptable to the rapid evolution of model capabilities and multi-modal contexts, as highlighted in [61]. Additionally, ethical and fairness considerations remain underexplored yet critically important, with methods inspired by counterfactual fairness [25] offering guidance for embedding accountability in controlled text outputs.

To ensure future readiness in evaluation practices, the field must prioritize collaborative efforts to define universal benchmarks and protocols alongside interpretabilityfocused methodologies. Modular toolkits, such as those proposed in [90], provide essential building blocks for flexible and unified evaluation. Likewise, advances in explainable reinforcement learning, including multi-reward weighting for style diversity [67], signal a promising shift toward optimizing controllability evaluations with transparency and adaptability in mind. Collectively, these directions underscore the importance of harmonizing robust, interpretable metrics with universally accepted evaluation frameworks, setting the stage for the next phase of progress in CTG.

6 Applications of Controllable Text Generation

6.1 Creative Content Generation

Controllable text generation (CTG) is reshaping the landscape of creative industries by offering tools to produce compelling and contextually appropriate content tailored to artistic or stylistic demands. From storytelling and poetic compositions to interactive gaming dialogues, CTG enables large language models (LLMs) to dynamically adapt to specific creative constraints such as tone, style, and narrative progression. This capability has far-reaching implications for creative professionals and industries aiming to balance efficiency with artistic vision.

One of the most prominent applications of CTG in creative content generation is adaptive storytelling. Here, models are trained to generate plots and dialogues that align with predefined narrative arcs, thematic elements, or user preferences. For instance, techniques such as progressive generation, which breaks the task into stages by first generating thematic keywords and gradually refining them into full narrative passages, have shown enhanced coherence and adaptability [91]. This refinement mimics the hierarchical thought processes of human authors, ensuring logical plot progression and engaging story development. Moreover, innovations like dynamic attribute graphs allow systems to adjust outputs based on key narrative attributes such as character personas or overarching plot objectives [12].

In the domain of poetic and figurative language production, which depends heavily on creativity and nuanced stylistic control, CTG techniques excel by leveraging fine-tuning on style-specific datasets or incorporating constraints at decoding time. Algorithms like NeuroLogic Decoding model predicate logic constraints, enabling the generation of text that adheres to strict stylistic or syntactic patterns while maintaining semantic relevance [10]. Similarly, contrastive prefixes, a lightweight method of steering generation by conditioning with attribute-specific vectors, have been effective at producing diverse creative styles with minimal computational cost [17]. These approaches highlight the potential to generate outputs that are not only syntactically rich but also aligned with users' artistic expectations.

CTG is further redefining artistic exploration and ideation processes by transforming the brainstorming stage of creative projects. Models fine-tuned with specific domain instruction sets or guided through attribute-driven probability adjustments deliver high-quality, context-aware suggestions for writers, artists, and content creators. For example, storytelling tools like MEGATRON-CNTRL integrate external knowledge bases to dynamically shape narratives, enabling artists to experiment with expansive, interconnected storylines while maintaining control over the thematic essence [8]. These tools help address common creative challenges such as writer's block or lack of thematic continuity.

Game content creation is another area in which CTG shows exceptional promise. Dynamic text generation for interactive dialogues, procedurally generated quests, and scenario-specific texts can enhance immersive experiences for players by adapting content to their choices and play styles. Decoding algorithms such as DExperts, which blend expert and anti-expert LMs to refine outputs, have been particularly successful in applications requiring consistent tone, character adherence, and thematic alignment [11]. Additionally, methods like speculative decoding and reinforcement-guided generation offer avenues to maintain fluency while ensuring the adaptability needed for complex, responsive gaming environments [92].

Despite its transformative potential, CTG in creative content generation also faces significant challenges. Balancing fluency with rigorous control remains a persistent trade-off, with over-constrained systems sometimes producing rigid or monotonous outputs [45]. Furthermore, ensuring originality and preventing overfitting to training data are critical when generating creative content, as excessive reliance on memorized patterns can stifle innovation [75]. Models should also grapple with addressing inherent stylistic biases, which may limit their applicability in diverse cultural and artistic contexts [72].

Emerging trends, such as multimodal generation combining textual outputs with other media like images or audio, offer exciting new directions for creative applications. Systems like ControlGAN extend the paradigm of CTG by enabling fine-grained manipulation of visual attributes based on linguistic constraints, inspiring cross-modal storytelling opportunities [71]. Additionally, real-time interactive systems employing user-feedback loops reveal opportunities for iterative refinement and co-creation processes with human artists [56].

Looking ahead, future research in CTG must address scalability and personalization challenges to better serve creative industries. Enhancing multimodal integration, improving stylistic diversity, and developing robust evaluation frameworks for creative quality will be crucial for advancing CTG's role as a partner in human-centered artistic exploration and storytelling. By fostering innovation in algorithm design and addressing ethical considerations, CTG can become a cornerstone of the creative economy.

6.2 Personalized End-User Applications

Personalization is a pivotal dimension of controllable text generation (CTG), enabling systems to tailor outputs to individual user preferences and transforming human-computer interactions across domains such as virtual assistants, chatbots, and content delivery platforms. By dynamically adapting to tone, sentiment, linguistic style, and user-specific contextual constraints, personalized CTG fosters enhanced usability, engagement, and satisfaction. This subsection explores the methodologies, trade-offs, and challenges of achieving personalized text generation, setting the stage for its potential to redefine user interactions through adaptive systems.

A foundational capability in personalized CTG lies in the dynamic control of sentiment and tone to align outputs with users' emotional preferences and communication expectations. Techniques like GeDi [19] and DExperts [11] exemplify this by enabling precise sentiment steering during text generation. These methods mitigate undesirable properties, such as toxicity, through discriminative probability adjustments applied to token distributions, allowing conversational agents to shift seamlessly between empathetic, professional, or casual tones based on user needs. However, this fine-grained control can sometimes compromise diversity and fluency, particularly in multi-turn interactions where adaptability is paramount.

Dynamic user adaptation further advances personalized CTG by optimizing outputs to meet specific linguistic or contextual preferences. Lightweight approaches such as prefix-tuning [17] encode personalized user descriptors or linguistic styles through compact embeddings, facilitating efficient adaptation without requiring full model retraining. Similarly, frameworks like CoCon [23] enable fine-grained personalization by conditioning word-level attributes to preserve syntactic coherence while aligning with user preferences. Yet, balancing multidimensional constraints—such as achieving both formality and brevity—poses persistent challenges, necessitating innovative solutions for dynamic optimization.

Accessibility-focused personalization represents another transformative application, especially in fostering inclusivity for users with varied linguistic or cognitive needs. Approaches that integrate fine-grained lexical and syntactic controls [20] have successfully simplified language for non-native speakers or individuals with cognitive impairments. Leveraging modular attribute controllers and residual hierarchies, these methods ensure the simultaneous satisfaction of multiple constraints, though computational cost and scalability barriers remain significant obstacles for broader adoption in real-world systems.

Emerging trends spotlight the potential of real-time adaptation to implicit user feedback, which bridges the gap between static prompts and evolving user dynamics. Reinforcement learning frameworks like CriticControl [24] leverage feedback signals during interactions, enabling systems to intuitively refine text outputs in response to subtle shifts in user contexts. However, optimizing trade-offs between latency and the granularity of learned feedback is crucial for the practical deployment of such adaptive systems, especially in scenarios requiring high responsiveness.

Despite these advancements, personalized CTG still contends with scalability and generalizability challenges. For example, while dynamic attribute graphs [12] offer sophisticated multi-attribute modeling, their reliance on comprehensive user profiling introduces risks of bias and overfitting, particularly for underrepresented demographic groups. Ethical concerns, including privacy in user-specific customization, further underscore the need for responsible frameworks that balance innovation with user trust and data security.

Future research should prioritize the integration of user-centric design principles, emphasizing multimodal inputs—such as combining sentiment analysis, contextual metadata, and interaction histories—to refine personalization dynamically. Lifelong learning paradigms, capable of continuously adapting to evolving user preferences, offer promising pathways to enhance the effectiveness of personalized CTG systems [93]. Robust evaluation protocols are equally necessary to quantify personalization outcomes, including subjective variables such as user satisfaction and long-term engagement. As CTG evolves, its ability to enable rich, user-tailored interactions will be instrumental in advancing the next generation of intelligent, context-aware systems.

6.3 Domain-Specific Content Generation

Domain-specific content generation represents a crucial application of controllable text generation (CTG), enabling precise, accurate, and efficient output tailored to highly specialized fields such as law, healthcare, and technical writing. These fields demand stringent adherence to factual correctness, strict syntactic conventions, and domain-specific knowledge, presenting unique challenges that CTG systems aim to address.

In the legal domain, CTG facilitates the drafting of complex legal documents, contracts, or case summaries, where the precision of language and compliance with jurisdiction-specific standards are non-negotiable. Approaches such as pretraining models on domain-focused corpora or fine-tuning for legal syntax and vocabulary have demonstrated effectiveness in maintaining legal coherence and accuracy. For instance, CTRL [3] achieves controllability by conditioning generation outputs on domain-specific codes, ensuring alignment with legal jargon and reasoning structures. Techniques emphasizing lexical constraints, such as plug-and-play methods [80], leverage external classifiers to adjust token probabilities, ensuring compliance with legally binding syntactic and semantic standards without retraining foundational models.

In healthcare, CTG underpins critical applications, including the drafting of medical reports, clinical summaries, and diagnostic content. The high stakes associated with this domain necessitate factual accuracy and terminological consistency. Systems like retrieval-augmented generation (RAG), which integrates external medical knowledge bases during inference, have shown promise in grounding text generation in up-to-date clinical guidelines and evidence-based knowledge [29]. For example, frameworks incorporating normalizing flow techniques for latent control [30] ensure that generated medical text adheres to both linguistic fluency and pathological coherence across multiple input attributes, such as patient-specific conditions and symptoms. However, these systems often face challenges in balancing the dynamic incorporation of evolving medical terminology with real-time adaptability—a problem partially addressed through parameter-efficient fine-tuning methods [89].

Technical domains such as software documentation and engineering reports also benefit significantly from CTG systems that enable the generation of structured, domain-specific textual content. A notable example involves leveraging hierarchical generation approaches [29] to impose sentence- and discourse-level planning, ensuring coherence across technical explanations. Frameworks like Syntax-Guided Controlled Paraphraser [26] adapt syntactic structures to meet constraints such as instructional clarity and modularity, both essential in technical documentation. Furthermore, multi-aspect control systems [62] are increasingly integrated to simultaneously address multiple demands, such as correctness, style, and brevity—key attributes in engineering texts.

Nonetheless, training CTG systems for domain-specific applications presents several challenges. First, domain corpora are typically limited in size, leading to insufficient training resources for robust generalization, especially when handling rare or low-resource domains. Techniques focusing on few-shot or zero-shot transfer learning, as demonstrated by advanced prompting systems [78], have emerged as potential solutions to address this bottleneck. Second, domain-specific generation often encounters a trade-off between maintaining creativity and adhering to factual correctness, particularly where generation systems are applied to generate summaries or extrapolate novel insights. Recent mathematical frameworks [94] propose optimizing trade-offs between creativity (exploration) and factual precision (exploitation) by encoding generation constraints within low-dimensional latent spaces.

Emerging trends in domain-specific CTG research underscore modular and hybrid architectures that combine retrieval-based systems with latent control mechanisms to achieve higher levels of factual fidelity and adaptability. Advances in explainability mechanisms, such as enabling models to justify generated assertions through structured causal reasoning [25], also provide a robust pathway to ensure reliability in highly sensitive fields like healthcare or law. Future efforts must further prioritize the development of scalable solutions for integrating dynamic domain knowledge while balancing efficiency and fluency, ensuring that CTG systems remain efficient and ethically aligned across all domain-critical applications.

6.4 Educational and Accessibility Applications

Controllable text generation (CTG) is transforming educational and accessibility applications by enabling systems to produce personalized, inclusive, and adaptive content suited to diverse linguistic and cognitive needs. Harnessing the strengths of large language models (LLMs), CTG offers solutions for overcoming barriers in education and accessibility, such as supporting learners with disabilities, adapting content for different educational levels, and generating multilingual or multimodal learning resources. Positioned alongside advancements in domain-specific and multilingual CTG, these applications underscore the wide-ranging societal impact of this technology.

A major contribution of CTG in education is its capacity to generate simplified and accessible content, providing engagement opportunities for individuals with varying levels of literacy or cognitive abilities. Techniques like contrastive conditioning and prefix optimization [17], [95] enable systems to control the complexity and style of output, ensuring readability without compromising semantic integrity. For instance, technical documents can be simplified for non-expert readers while maintaining accuracy, fostering inclusive education. These approaches also enhance accessibility for learners with disabilities, such as those with dyslexia, by generating texts with improved clarity, logical flow, and reduced syntactic complexity.

Interactive question generation for assessments and personalized learning paths exemplifies another promising use of CTG in education. Models such as InstructCTG [43] guide generation based on structured input or control codes, producing relevant, pedagogically sound questions. These systems adapt to learners' knowledge levels, offering differentiated assessments that support individualized learning. Furthermore, planning-based frameworks like PAIR [96] align educational content with predefined instructional goals, generating domain-specific explanations and problem sets tailored to diverse learner needs. These innovations parallel the structured approaches seen in domain-specific CTG, addressing the demands for precision and adaptability.

Beyond simplifying text and generating dynamic assessments, CTG facilitates multi-level summarization and creation of instructional content for varied audiences. Summarization methods that incorporate control constraints, such as diffusion-based approaches [97], enable the generation of multi-tiered summaries adapted to differing levels of expertise. In classroom settings, this capability allows students with varying proficiencies to receive appropriately detailed explanations, promoting equitable and inclusive education practices. Such multi-level adaptability reflects the nuanced control mechanisms seen in CTG for multilingual applications, highlighting its broader relevance across domains.

CTG further supports the creation of multimodal educational tools, integrating textual generation with image or audio inputs. Frameworks like Show, Control and Tell [34] enable the development of holistic learning materials, such as audio-described videos or annotated diagrams, benefiting visually or hearing-impaired learners and encouraging cross-modal learning experiences. Procedural generation frameworks like SceneX [98] exemplify how CTG can create interactive environments or educational simulations, making abstract concepts more accessible and engaging. These developments align with efforts in multilingual and cultural CTG to provide contextually relevant, highly adaptive outputs.

Despite these advances, challenges remain in ensuring the factual accuracy, ethical grounding, and cultural relevance of generated content. Techniques such as retrieval-augmented generation [16] enhance factual reliability by grounding output in external knowledge bases, yet maintaining control across diverse educational scenarios—including multilingual and cross-cultural contexts—poses persistent technical hurdles. Cross-lingual transfer learning [77] offers promising solutions for under-resourced languages, paralleling issues of linguistic equity in multilingual CTG, but requires further innovation to guarantee robustness and inclusivity at scale.

Emerging trends in educational CTG emphasize real-time user feedback to refine generative outputs interactively, as illustrated in frameworks like RecurrentGPT [56]. Such technologies could dynamically adapt learning materials based on learner interactions, fostering more personalized and engaging educational experiences. Advances in explainability [25], critical in educational applications, enhance trust in CTG systems by clarifying how content is controlled and aligned with pedagogical objectives. These developments echo trends in other CTG applications that prioritize transparency and user-centered customization.

In conclusion, CTG is reshaping education and accessibility by delivering adaptive, inclusive, and engaging content tailored to diverse needs. Future research should focus on refining multimodal, cross-lingual, and culturally sensitive approaches, along with robust evaluation frameworks to ensure alignment with pedagogical goals across settings. By addressing these challenges and building upon techniques shared across domain-specific and multilingual CTG applications, this technology holds the potential to democratize education and bridge accessibility gaps, fostering more equitable learning opportunities worldwide.

6.5 Multilingual and Cross-Cultural Communication

Controllable Text Generation (CTG) has become a pivotal technology in tackling the linguistic and cultural diversity challenges inherent in multilingual and cross-cultural communication. With the global prominence of large language models (LLMs), CTG aims to produce text that not only fulfills linguistic precision but also adheres to sociolinguistic and cultural norms across different regions. This subsection critically examines the advancements, trade-offs, and emerging trends in CTG for multilingual and cross-cultural scenarios while aligning discussions with relevant work.

At the core of multilingual CTG lies the challenge of generating text in multiple languages tailored to diverse cultural expectations. Models like CTRL [3] and multilingual fine-tuned LLMs have demonstrated the ability to encode linguistic variation within a unified framework. These models utilize control codes or embedding spaces to align language generation with specific cultural or linguistic contexts, enabling regionally appropriate translations. However, their reliance on extensive pretraining datasets introduces biases and uneven coverage, particularly against low-resource languages, which often lack the requisite corpus size for robust fine-tuning [5]. Addressing this, techniques combining retrieval-augmented generation (RAG) [50] with dynamic prompts tailored for low-resource languages have emerged as avenues for improving inclusivity while maintaining text fluency and factual consistency.

Automatic translation remains a cornerstone application of CTG in multilingual communication. Beyond direct linguistic translation, nuanced tasks such as sentiment retention or idiomatic expression preservation require specialized control strategies. Chain-editing frameworks like Locate&Edit [58] integrate energy-based constraints to preserve contextual and idiomatic integrity during translation while efficiently adapting semantic nuances. Such advancements outperform purely statistical models but often require fine-grained, language-specific attribute tuning.

Generating culturally sensitive text is a more intricate challenge, as sociolinguistic norms vary significantly across regions. Advanced CTG models integrate mechanisms like dynamic attribute graphs [12], offering controllability over variables such as politeness and formality. By mapping sociolinguistic norms to control parameters, these models allow flexible modulation of tone to align with expectations across cultures. For instance, apologetic phrasing in Japanese often demands high levels of politeness distinct from English, which stresses directness. However, deploying these solutions at scale remains computationally expensive, especially across multi-dimensional cultural contexts.

An emerging trend involves leveraging instruction-tuned language models (e.g., InstructGPT) for handling multilingual text generation tasks by rephrasing cultural nuances into explicit, instruction-based frameworks [43]. This approach minimizes the need for retraining models for each linguistic constraint, enhancing scalability. Additionally, contrastive decoding mechanisms [87] have shown promise in resolving the trade-off between local contextual fidelity and global fluency in multilingual outputs. By penalizing irrelevant or geographically biased completions, these mechanisms robustly ground generative processes in culturally relevant contexts.

Low-resource languages present another significant frontier. Cross-lingual transfer learning strategies, such as multilingual extensions to RAG [50], emphasize combining prompts and retrieval capabilities to dynamically enhance low-resource language support. Techniques such as parameter-efficient fine-tuning (e.g., LoRA and prefix-tuning) [60], for lightweight adaptation, ensure models retain efficiency while improving language-specific controllability. Despite these solutions, ensuring linguistic equity across languages remains constrained by an imbalance in data availability.

Future directions in multilingual and cross-cultural CTG include integrating multimodal controls to ground textual outputs in multi-sensory contexts [61]. These systems promise richer cultural representations, such as generating captions aligned with culturally specific visual cues. Additionally, emerging alignment methods like Knowledge-Aware Fine-Tuning (KAFT) [35] aim to dynamically reconcile factual inconsistencies or biases with cultural expectations, further broadening the scope of CTG in this domain.

In summary, CTG for multilingual and cross-cultural communication embodies significant advancements but faces persistent challenges in equitable resource usage, balance between global fluency and local norms, and cultural nuance embedding. Solutions integrating dynamic retrieval, efficient fine-tuning, and multimodal controls hold transformative potential for global inclusivity and accessible communication. However, the complexity of linguistic diversity necessitates continued innovation in aligning multilingual CTG with human-centric sociocultural expectations.

6.6 Safe and Ethical Text Applications

Controllable text generation (CTG) systems play an indispensable role in upholding ethical norms and societal standards, ensuring that generated outputs are aligned with principles of safety, inclusiveness, and fairness. By enabling text to be tailored along defined attributes such as tone, sentiment, and contextual appropriateness, CTG holds significant promise in mitigating toxicity, fostering inclusivity, and supporting adherence to regulatory constraints across diverse applications.

A prominent application of CTG lies in the mitigation of harmful or undesirable text, including toxic language, hate speech, and biased narratives. Plug-and-play methods, as demonstrated by "Plug and Play Language Models: A Simple Approach to Controlled Text Generation" [4], allow the on-the-fly steering of pretrained language models to suppress toxic outputs without necessitating retraining. These methods leverage lightweight attribute classifiers to dynamically adjust generation probabilities, enabling flexible output adaptation while maintaining fluency and diversity. Similarly, gradient-based decoding approaches like Future Discriminators (FUDGE) [7] refine model outputs probabilistically, employing Bayesian decomposition to make context-sensitive adjustments. Despite their efficacy in toxicity reduction, these techniques often face challenges in maintaining fluency while adhering to multiple intersecting attribute constraints, such as sentiment and contextual appropriateness, in complex generative scenarios.

Promoting inclusivity through language adaptation forms another cornerstone of ethical CTG. Techniques such as counterfactual augmentation and energy-based frameworks are used to mitigate biases in controlled settings. For instance, Distributional Approaches [37] utilize KL-divergence minimization to generate outputs that align with fairness metrics, addressing both specific and systemic biases. These strategies enable broader coverage of nuanced and intersectional biases in fine-tuned or instruction-tuned models. However, addressing fairness often introduces trade-offs, such as limiting legitimate linguistic variation or homogenizing outputs excessively. Disentanglement-based approaches, exemplified by MAGIC [68], explicitly separate style and content attributes, offering a more balanced mechanism to manage these trade-offs effectively.

Ensuring regulatory compliance is a critical consideration for CTG systems, particularly in domains such as legal documentation, healthcare communication, and finance, where accuracy and adherence to regulations are paramount. Methods leveraging dynamic attribute graphs [12] have demonstrated success in maintaining control over domain-specific lexicons, such as legal terminology or financial jargon, thereby minimizing risks associated with inaccuracies. These solutions dynamically adjust attribute-specific lexical constraints without requiring retraining, ensuring enhanced fluency and reduced perplexity while simultaneously meeting regulatory expectations.

CTG technologies also significantly contribute to content moderation by enabling automated systems that filter, guide, or flag inappropriate conversational behaviors. Reinforcement learning-based methods, such as DisCup [99], demonstrate improved moderation efficacy by leveraging discriminator models to optimize unlikelihood objectives, achieving better constraint satisfaction. Furthermore, decoding techniques like COLD [6] integrate constraining mechanisms within an energy-based probabilistic framework, facilitating precise and dynamic control of the output space. However, moderation systems must balance automated processes with human oversight to avoid amplifying undesired content due to inaccurate classifier predictions. Hybrid models that incorporate human feedback loops alongside machine optimization present a promising avenue for achieving greater reliability and practical applicability.

Emerging trends in ethical CTG underscore the growing need for highly granular, multi-attribute control to address the complexities of real-world applications. Innovations such as dynamic reward weighting [67] and counterfactual training paradigms show potential in resolving conflicts between overlapping attribute constraints, such as balancing safety with sentiment preservation. Moreover, the development of comprehensive evaluation frameworks capable of assessing fairness, inclusivity, and contextual integrity across diverse demographics is crucial for building trust and accountability in large-scale CTG applications.

While significant strides have been made in enhancing the ethical dimensions of CTG, challenges remain. Key areas for further research include improving interpretability within controlling mechanisms, addressing trade-offs between constraint fidelity and generalizability, and establishing standardized yet adaptable evaluation benchmarks. By confronting these challenges, CTG can become a cornerstone of responsible AI, enabling the safe, inclusive, and equitable deployment of language models across various domains.

7 Challenges and Ethical Considerations in Controllable Text Generation

7.1 Balancing Flexibility and Adherence

Balancing flexibility and adherence in controllable text generation presents a complex optimization problem where striking the right equilibrium between two often competing principles becomes crucial: maintaining the adaptability of large language models (LLMs) and ensuring strict fidelity to predefined control objectives. Excessive emphasis on either end of this spectrum risks impairing the model's overall effectiveness. On one hand, rigid adherence to constraints may lead to overfitting, reducing linguistic diversity and naturalness. On the other hand, excessive flexibility may yield outputs that deviate significantly from desired attributes or objectives. This subsection addresses the technical roots of this challenge, examines prominent methodologies, and proposes promising strategies for achieving a harmonious trade-off.

Overfitting to control parameters often results from models that assign undue priority to the predefined conditions, thereby curbing the model's inherent generative capabilities. For example, heavily parameterized control frameworks using fine-tuned models often suffer from a "degeneration effect," where lexically static outputs arise since the model sacrifices fluency and diversity to satisfy constraints [4]. This rigidity is amplified when incorporating hard lexical constraints or formal control codes [26], often yielding syntactically and semantically awkward outputs. Prior work consistently demonstrates that enforcing strict adherence to multiple simultaneous constraints, such as style, sentiment, and factuality, compounds this problem, as the control objectives can inherently conflict, thereby creating a highly restricted output space [25].

Conversely, inadequate enforcement of control parameters, often referred to as underfitting, leads to outputs that fail to respect critical constraints, diminishing trust in these models for high-stakes applications such as legal, medical, or regulatory document generation tasks [8]. For instance, simplistic approaches like top-k sampling or nucleus sampling (top-p) achieve flexibility at generation time but exhibit limited fidelity to specific constraints such as factual correctness or stylistic alignment [10].

Several mitigation strategies have emerged to address these trade-offs. Dynamic weighting mechanisms for control conditions are a promising direction, wherein constraint importance is modulated adaptively during generation to balance adherence and flexibility [6]. For instance, energy-based inference frameworks like COLD dynamically adjust to semantically plausible trajectories while preserving fluency. Moreover, retrieval-augmented generation introduces external evidence to ground outputs in factual context, thereby enhancing adherence without hampering linguistic adaptability [15]. These retrieval-augmented approaches are particularly effective for knowledge-intensive tasks and demonstrate improved control compliance compared to purely generative mechanisms [100].

Meta-learning techniques further extend this balance by training models to generalize across a distribution of control tasks without becoming overly specialized. For example, the use of meta-objective frameworks such as Gradient Episodic Memory (GEM) allows the model to manage trade-offs dynamically during multi-task fine-tuning, safeguarding against overfitting to narrow control objectives [2]. Lightweight control methods like contrastive prefix tuning also achieve remarkable success in steering outputs while maintaining fluency. These approaches leverage flexible, attribute-specific embeddings, bypassing fine-tuning of the main model and thereby limiting computational overhead [17].

Emerging directions in this domain focus on hybrid methods that combine hard constraints with soft, probabilistic mechanisms. For example, recent advances in sequential constraint satisfaction using dynamic graphs enable models to refine outputs progressively rather than enforcing constraints deterministically at each decoding step [12]. Furthermore, interpretable control methods based on structural causal models (SCMs) aim to disentangle the causal relationships between input attributes and generated text, enabling more nuanced control without compromising generality [25].

Looking ahead, the integration of reinforcement learning frameworks, such as Proximal Policy Optimization (PPO), with reward models tailored to multi-faceted control objectives holds promise for further enhancing adaptability [11]. Additionally, dynamic user feedback loops that iteratively refine model outputs in real-time open new avenues for balancing control adherence with user-defined preferences, thus improving applicability in interactive systems [56].

While considerable progress has been achieved, challenges remain. Control-induced biases and conflicts between constraints in multi-objective settings underscore the need for robust conflict resolution methods. Future breakthroughs will likely hinge on developing scalable, interpretable, and context-aware control mechanisms that minimize resource overhead while ensuring outputs remain both diverse and faithfully aligned with user-specified goals.

7.2 Biases and Fairness Concerns in Controlled Outputs

Biases and fairness concerns are critical ethical challenges in the realm of controllable text generation (CTG). Large language models (LLMs) inherit biases from the vast datasets they are trained on and are further shaped by control specifications during fine-tuning or prompt design. These properties risk enabling systems to perpetuate or even amplify harmful stereotypes, inequality, and exclusion. This subsection investigates the sources of such biases, analyzes the compounding effects of control mechanisms, and highlights strategies to detect and mitigate unfair outcomes in CTG systems.

A major source of bias in CTG originates from the training data used to pre-train LLMs, which are often derived from massive web-based datasets that mirror existing societal inequities. These datasets intrinsically exhibit imbalances in the representation of populations with respect to race, gender, and socio-economic context, leading to both explicit and implicit biases in generated outputs [5], [101]. When CTG systems layer additional control mechanisms—such as fine-tuning or reinforcement learning to regulate attributes like tone or sentiment—the biases embedded in pre-training may interact with the control objectives. This interplay can inadvertently reinforce harmful stereotypes or emphasize biased patterns inherent in the data. Attribute-specific fine-tuning, for example, may amplify latent biases tied to the specific control target, as shown in sentiment-controlled generation tasks where positive or negative sentiment amplification can significantly exacerbate gender stereotypes [47].

Moreover, biases are not only confined to dataset imbalances but can also emerge during the process of optimizing control parameters, which may exacerbate fairness issues. Fine-grained attribute control or multi-attribute generation often creates trade-offs between satisfying user-specific constraints and producing fair, inclusive outputs. For example, safeguarding against toxic content generation may lead models to systematically reduce representation or alter the portrayal of underrepresented demographics, a phenomenon observed in decoding strategies that strictly enforce toxicity mitigation protocols [11], [19]. Additionally, the use of auxiliary reward models in reinforcement learning frameworks introduces further risks, as the latent spaces optimized for alignment with such reward functions may encode and reinforce biased correlations, thereby diminishing transparency and fairness in output generation [37], [43].

To counter these challenges, several strategies for bias detection and mitigation have been developed, targeting both the pre-training and control dimensions of CTG systems. Causal modeling has emerged as a promising tool to disentangle spurious correlations and identify confounding factors within both input datasets and control specifications, allowing researchers to better understand and mitigate structural sources of bias [25]. Counterfactual data augmentation, which generates synthetic examples by altering sensitive attributes such as gender or race while maintaining other content, has also shown potential in debiasing LLMs and their controlled outputs [25], [37]. Similarly, adversarial training approaches—which minimize unintended correlations between controlled attributes and sensitive factors—represent another pathway for enhancing fairness without compromising generation quality [101].

Techniques grounded in retrieval-augmented generation offer additional opportunities for debiasing CTG systems. By incorporating external verified knowledge into the generation pipeline, retrieval-augmented methods can mitigate bias propagation while enhancing factual robustness. For instance, carefully curated retrieval corpora aligned with unbiased knowledge sources have proven effective in reducing bias risks without compromising content quality or coherence [48], [50]. At inference time, dynamic decoding strategies such as GeDi [19] and contrastive decoding [102] provide flexible mechanisms to steer outputs away from harmful biases, even in the absence of extensive model retraining.

Despite these innovations, substantial challenges remain. Measuring bias in generated outputs is an unresolved issue, complicated by the inherently subjective and contextual nature of fairness judgments, which can differ across cultures and domains [103]. Moreover, efforts to achieve fairness may introduce trade-offs with other generation qualities like fluency or diversity, requiring careful balancing. Adversarial debiasing and counterfactual approaches, while promising, face additional hurdles in ensuring adequate representation of diverse demographic groups in synthetic or real datasets [5]. Addressing these gaps necessitates interdisciplinary collaboration and the development of tools that integrate technical, sociological, and cultural perspectives.

Looking forward, embedding fairness principles directly into the formulation of control objectives offers a promising research direction. Explainable models capable of tracing how biases propagate during controlled generation can enhance transparency and accountability. Advances in causal inference, coupled with scalable, cross-lingual methods for debiasing, may improve fairness across multilingual and low-resource language contexts [25], [101]. Additionally, the creation of standardized evaluation metrics to measure fairness alongside fluency, diversity, and factual alignment will be crucial in establishing benchmarks for more equitable CTG systems.

Ultimately, addressing biases and fairness in CTG demands concerted effort from the AI community, emphasizing ethical accountability alongside technical innovation. Involving diverse stakeholders in the design, development, and evaluation of CTG systems will ensure that these technologies foster inclusivity, reduce harm, and contribute responsibly to socially impactful applications.

7.3 Hallucination and Factual Inconsistency

The phenomenon of hallucination and factual inconsistency in controllable text generation represents one of the most persistent and significant challenges in the application of large language models (LLMs). This issue arises when generated outputs include erroneous, fabricated, or unverifiable content—a critical concern in domains requiring factual rigor, such as medical, legal, or scientific documentation. In the context of controllable text generation, hallucinations can be exacerbated by the interaction between control attributes and the underlying language model's probabilistic tendencies. This subsection examines the roots of hallucinations, critically evaluates methodologies designed to mitigate them, and explores emerging paradigms aimed at enhancing factual robustness.

Hallucinations often stem from weak grounding in external or verifiable knowledge, which is compounded by the stochastic nature of autoregressive generation models. When constraints or control objectives impose demands outside the model's inherent knowledge scope, it may fill gaps with plausible but false information. For instance, the misalignment between pretrained knowledge representations and control attributes (e.g., domain-specific or stylistic constraints) can lead to hallucination, as observed in controllable dialogue systems where the push for informativeness compromises factual integrity [63]. Additionally, hallucinations can proliferate due to over-reliance on spurious correlations learned during training, where biases in data distributions can distort content fidelity [25].

One promising avenue for reducing hallucination involves retrieval-augmented generation (RAG), which dynamically incorporates external factual knowledge during text generation. By retrieving contextually relevant information from external knowledge bases, these models strengthen the grounding of generated outputs. For example, the integration of retrieval-based mechanisms has proven effective in improving the accuracy of controlled outputs without retraining the LLM [29]. However, RAG approaches face challenges in maintaining semantic coherence between retrieved content and generated text, particularly when multiple constraints, such as stylistic tone or sentiment, are applied concurrently [36].

Decoding-time interventions present another strategy for minimizing hallucinations. Methods such as constrained beam search and energy-based decoding frameworks dynamically adjust token probabilities to enforce factual consistency and suppress erroneously plausible alternatives [104]. While these techniques excel at incorporating constraints during inference, they often entail a trade-off: the more rigidly constraints are enforced, the greater the risk of reduced fluency or logical flow in the output. Techniques that combine soft constraints with adaptive weightings, such as MuCoCO [36], have shown promise in balancing factual fidelity with textual quality.

To address specific high-stakes use cases, approaches leveraging structured methodologies like chain-of-verification have gained traction. These methods compel the model to self-rationalize or follow logical steps during generation, thereby enhancing the factual grounding of complex outputs. For instance, structured guidance frameworks, such as dynamic attribute graphs, utilize interdependent attribute relationships to reduce factual inconsistencies in multi-constraint generation settings [12]. Despite their effectiveness, these methods remain computationally intensive and dependent on the quality of external attribute signals.

Beyond technical interventions, the development of robust evaluation metrics is essential for systematically addressing hallucinations. Current metrics often fall short in capturing nuanced factual inconsistencies, particularly in controlled outputs combining multiple generation objectives. Advanced alignment techniques, such as reinforcement learning with human feedback (RLHF), have demonstrated improvements but are limited by subjective variability in human preferences [44]. Multi-dimensional benchmarks incorporating task-specific factuality assessments—like cross-referencing generated facts with external databases—are increasingly necessary for substantive progress in this area [79].

Emerging trends suggest a paradigm shift toward hybrid frameworks that combine probabilistic generation, retrieval-based augmentation, and symbolic reasoning. The unification of these approaches could reduce hallucination by embedding stricter verification mechanisms during generation. In addition, advancements in causal modeling for controllability hold promise in disentangling spurious associations from true factual relationships, offering enhanced control over generated outputs [25]. Future directions should also consider methods to dynamically adapt models to shifting factual contexts, leveraging continual learning paradigms for real-time factual consistency.

Ultimately, while substantial progress has been made, achieving hallucination-free controllable text generation remains an open challenge. Innovations in knowledge integration, smarter decoding mechanisms, and rigorous evaluation frameworks are critical for mitigating the risks of hallucinated content and enabling the safe deployment of LLMs in critical application domains.

7.4 Computational Efficiency and Scalability

The rapid advancement of large language models (LLMs) has unlocked unprecedented capabilities in controllable text generation. However, the computational demands associated with achieving fine-grained control introduce significant challenges, particularly when optimizing these systems for efficiency and scalability. This subsection explores these challenges, evaluates strategies designed to address resource constraints in training and inference, and outlines potential avenues for future research to advance efficient and sustainable controllable text generation.

Large-scale LLMs inherently require substantial computational resources for both pretraining and fine-tuning, a demand that escalates further when integrating multiple control dimensions, such as sentiment, style, or length. Traditional approaches to task-specific fine-tuning, which involve updating the full set of model parameters for each control objective, are often computationally untenable—especially for models with billions of parameters. Studies have demonstrated that task-specific fine-tuning not only incurs considerable memory and energy overhead but also struggles to maintain scalability for evolving and diverse control objectives [74]. Moreover, inference-time controllability frequently relies on techniques such as plug-and-play methods or constrained sampling strategies, which add computational complexity through iterative adjustments to token probabilities [16], [80].

To alleviate these resource challenges, parameter-efficient fine-tuning methods, such as low-rank adaptations (LoRA) and adapter-based approaches, have come to the forefront. These techniques limit the need for updating the entire model by training only a small subset of parameters, thereby preserving the core functionality of the pretrained model while significantly reducing memory and computational costs [55]. For instance, the StructAdapt framework achieves strong task performance by training just 5.1

Another promising direction lies in hybrid modular frameworks that decouple control mechanisms from the core text generation process. By separating symbolic planning stages from neural realization, such frameworks can eliminate computational redundancy, focusing computational effort more selectively [32], [33]. For instance, dynamic content planning methods, such as those demonstrated in PLANET, leverage autoregressive mechanisms and coherence-enhancing objectives to efficiently generate long-form controlled text while maintaining consistency and fidelity across attributes [29].

In addition to training optimizations, decoding-time strategies play an equally crucial role in enhancing computational efficiency. Techniques such as constrained beam search and energy-based optimization frameworks enable models to simultaneously enforce multiple constraints during decoding without necessitating additional parameter updates [36], [58]. Moreover, alternative approaches like diffusion-based models, exemplified by Diffusion-LM, tackle the computational bottlenecks of autoregressive decoding while achieving sophisticated control capabilities [97].

Despite these advancements, significant challenges persist. Scaling controllable text generation to accommodate multiple attributes without interference between control mechanisms often involves complex computational tradeoffs [62]. Achieving a balance between constraint satisfaction, fluency, and diversity within a unified framework remains elusive, requiring further innovations in optimization algorithms and training paradigms [45]. Additionally, ensuring model performance and reliability in low-resource or real-time settings remains a pressing concern, particularly as models grow larger and increasingly complex [17].

Future research should prioritize the development of lightweight and modular architectures capable of incorporating new control dimensions on demand while minimizing resource consumption. Promising techniques include reinforcement-based alignment using external control signals and distributional optimization in latent spaces, which could both reduce computational costs and enhance control fidelity [69], [93]. Additionally, adaptive fine-tuning mechanisms that dynamically respond to changing task requirements or user preferences could facilitate greater scalability and adaptability in real-world applications. Crucially, progress in computational efficiency must be aligned with ethical imperatives, such as reducing the environmental impact of training and inference through carbon-aware strategies and energy-efficient architectures. Integrating such considerations will be pivotal to achieving sustainable and scalable controllable text generation, facilitating the responsible deployment of LLMs across diverse applications.

7.5 Security, Misuse, and Malicious Applications

Controllable text generation systems, while offering transformative potential across a variety of domains, also introduce significant security and misuse risks. The ability to shape and steer text generation towards specific outputs amplifies malicious opportunities, ranging from disseminating disinformation and creating persuasive propaganda to drafting contextually accurate but harmful content for phishing campaigns or personalized scams. This subsection delves into the manifold threats posed by such misuse, evaluates current mitigation strategies, and identifies gaps and challenges alongside emerging trends in tackling these threats.

The misuse of controllable text generation is particularly problematic in its potential to mass-produce disinformation and biased narratives. By conditioning outputs on attributes such as political sentiment, stylistic tone, or thematic framing, malicious actors can craft propaganda finely tuned to manipulate public opinion or reinforce harmful stereotypes [3]. For instance, with user-defined prompts and stylistic constraints, models can generate fake news articles exhibiting high contextual relevance and coherence, further undermining trust in legitimate information sources. Additionally, the automation afforded by such systems allows adversaries to scale campaigns cost-effectively, thereby exacerbating the societal impact.

The issue of bias amplification is another critical concern, as these generation systems inherently reflect biases present in their training data. When paired with controllable mechanisms, this tendency can enable even stricter reinforcement of harmful prejudices. For example, biases tied to gender, race, or ideology can become more pronounced when fine-grained control mechanisms are applied, as seen in techniques explicitly optimizing for specific attributes [25], [36]. Beyond amplifying pre-existing biases, adversarial actors can exploit these mechanisms to deliberately generate biased narratives aligned with their objectives, leading to ethical dilemmas.

A more specific application of concern is the use of controllable generation in socially harmful contexts, such as phishing and fraud. Actors can steer models towards generating phishing emails personalized to the recipient's preferences, regional dialect, or even conversational tone, making detection by users or even automated classifiers increasingly difficult. Advances in dynamic attribute graphs [12] have bridged the ability to modulate subtle, context-specific textual attributes, which could support deceptive campaigns that exploit vulnerabilities at unprecedented levels of efficacy and customization.

Existing defensive techniques, such as toxicity detection and filtering mechanisms, have shown some promise in mitigating certain risks but remain inadequate for addressing more nuanced threats. For instance, adversarial misuse relies on exploiting subtleties in controllable outputs, such as minor stylistic shifts that evade standard moderation systems. Anti-toxicity strategies often rely on fine-tuning language models with adversarial unlearning mechanisms [105] or adjusting decoding constraints [58]. While these techniques have successfully reduced some toxic outputs, they fall short in controlling harmful biases or ensuring ethical safety when multiple overlapping control dimensions are involved.

Emerging research seeks to develop more robust safeguard mechanisms. One promising approach involves retrieval-augmented systems, which ground generated outputs in verified knowledge sources, ensuring fidelity to factual constraints [48]. Additionally, adversarial testing frameworks are gaining traction, where models are stress-tested with synthetic malicious prompts to evaluate resilience against misuse scenarios [83]. These strategies indicate a shift towards proactive risk mitigation but require further refinement to approach reliability in real-world, high-stakes environments.

Looking forward, the development of robust governance frameworks is critical to mitigating the risks of misuse and ensuring alignment with ethical standards. Techniques integrating explainability into controllability systems—such as causal modeling frameworks for tracing the relationship between input constraints and generated outputs [25]—can provide valuable transparency to audit model behavior and prevent malicious exploitation. Similarly, real-time monitoring systems coupled with adversarial retraining pipelines could dynamically adapt models to counter emergent threats. However, the scalability and computational demands of implementing such mechanisms across diverse controllable applications pose non-trivial technical and infrastructural challenges.

In conclusion, the security and misuse risks posed by controllable text generation systems represent a pressing area of concern. While advancements in control mechanisms offer increased flexibility, they simultaneously broaden the scope for malicious applications. Mitigation strategies must evolve to comprehensively address disinformation, bias amplification, and adversarial exploitation, leveraging interdisciplinary solutions that balance technological safeguards with ethical responsibility. Further research into explainable and adaptive safeguards, as well as collaborative frameworks among academia, industry, and regulators, will be fundamental to ensuring safer deployments of these powerful systems.

7.6 Transparency, Explainability, and Ethical Responsibility

Transparency, explainability, and ethical responsibility are foundational to the safe and trustworthy development of controllable text generation (CTG) systems, particularly as these technologies increasingly influence both high-stakes domains (e.g., medical summarization, legal document drafting) and wider societal applications (e.g., conversational AI, creative tools). Although advances in CTG methodologies have yielded remarkable progress, these critical dimensions remain insufficiently addressed, raising significant concerns about their impact on trust, safety, and equity. This subsection delves into the technical and ethical challenges associated with achieving transparency, explores the evolving landscape of explainability techniques, and highlights the ethical imperatives inherent in deploying CTG technologies responsibly.

Achieving transparency in CTG is inherently challenging due to the black-box nature of large language models (LLMs) and their control mechanisms. Leading techniques such as Plug-and-Play Language Models [4] or reinforcement learning-based optimization methods [67] often rely on manipulating complex internal representations, such as activation patterns, logits, or latent states. These opaque processes obscure how specific controls—such as style, sentiment, or factual consistency—shape the generation process. However, recent efforts to enhance interpretability, such as attribute-level transparency frameworks [43] and energy-based methods like COLD decoding [6], provide promising pathways by framing control as a sequence of measurable decisions. For instance, energy-based approaches facilitate token-level modulation during decoding, creating opportunities for more granular insights into constraint enforcement. Nonetheless, such methods are computationally intensive and largely confined to specific types of controls, limiting their scalability and general practicality.

Explainability, while closely related to transparency, also includes fostering comprehension among end-users. Models such as Tailor [89] incorporate explainable control mechanisms that enable users to disentangle overlapping attributes in multi-aspect generation settings. Similarly, causal modeling frameworks [25] offer analytical insights into the causal relationships between input constraints and generated text. Meanwhile, disentangled counterfactual methods [68] aim to resolve complexities arising from interdependent attributes, thereby improving interpretability of multi-aspect controls. While these advances contribute to making CTG processes more understandable, they also underscore significant limitations—such as the need to tailor explainability approaches to diverse user requirements and contextual interpretations.

Ethical responsibility in deploying CTG systems represents a multidimensional challenge, encompassing issues such as bias mitigation, accountability, and societal alignment. Many pre-trained models underlying CTG capabilities inherit biases from their training datasets, risking the amplification of harmful stereotypes, toxic outputs, or content misaligned with user intentions. Counterfactual generation techniques [106] have illustrated some potential for mitigating such biases by enabling adjustments to generated outputs, yet transparency into constraint enforcement remains critical to avoid reinforcing bias unintentionally. Similarly, dynamic approaches like attribute graphs [12] have shown efficacy in balancing fairness with control fidelity by dynamically adjusting generation patterns to avoid disproportionately privileging specific attributes. Novel strategies like CriticControl [24], which integrates learned critic models, further exemplify how ethical considerations can guide the generation of content that adheres to predefined norms while safeguarding structural coherence and ethical integrity. However, the ethical deployment of CTG systems ultimately requires bridging technical solutions with proactive stakeholder engagement, accountability mechanisms, and alignment with evolving societal standards.

Emerging research trends suggest promising avenues for weaving together transparency, explainability, and ethical responsibility. For example, integrating causal inference methods with post-hoc interpretability tools may enhance both the transparency of CTG systems and user trust, enabling models to justify outputs through explicitly controllable mechanisms rather than opaque latent dynamics. Similarly, frameworks such as composable latent-space controls [69] open new possibilities for managing multiple constraints simultaneously, preserving high fidelity to input specifications while increasing interpretability.

As the field advances, efforts must prioritize the design of universally interpretable CTG systems that embed transparency as a core feature rather than an optional add-on. Enhancing system design with explainable visualizations (e.g., attention maps) and actionable user feedback loops during inference could significantly narrow the gap between model logic and user understanding. In addition, establishing ethical benchmarks similar to VisionPrefer [107] could serve to standardize metrics for transparency, fairness, and accountability within CTG systems. By aligning technical innovations with comprehensive ethical frameworks and societal expectations, the field can ensure that controllable text generation not only maximizes utility but also fosters equity, safety, and trustworthiness, laying the groundwork for its responsible integration into critical and everyday applications alike.

8 FUTURE DIRECTIONS AND EMERGING TRENDS

8.1 Integration of Multimodal Controllability

The integration of multimodal controllability into text generation systems represents a compelling frontier for advancing the coherence, adaptability, and contextual alignment of outputs across different modalities, such as images, audio, and video. By leveraging the interplay between modalities, these systems can generate text that is both grounded in non-textual cues and dynamically tunable based on desired constraints. This area of research holds transformative potential for applications ranging from video-script alignment to creative storytelling and cross-domain knowledge synthesis.

One prominent approach to multimodal controllability extends conditional text generation techniques to incorporate non-textual signals as context. Conditional frameworks like CTRL [3] and energy-based methods such as COLD Decoding [6] offer foundational insights into how multimodal data can influence text outputs. These frameworks are increasingly being adapted to align textual attributes (e.g., sentiment, tone, factuality) with cross-modal representations, such as the spatial information in images or sequential flow in audiovisual media. For instance, ControlGAN's ability to apply word-level attention to image regions during text-to-image generation [71] illustrates the benefits of fine-grained multimodal alignment. However, extending these mechanisms to bidirectional tasks—such as generating captions from videos and adapting those captions to stylistic or structural constraints—poses additional challenges.

The development of unified multimodal learning frameworks is critical for the seamless integration and control of text generation across modalities. These systems combine dynamic weighting techniques and modular architectures to fuse information from text, vision, and audio, optimizing both content generation and modality-specific constraints. For example, methods like Transformer-based latent variable models [2] or attribute conditioning using dynamic graphs [12] provide mechanisms for coordinating intermodal consistency. A central challenge lies in balancing the contributions of diverse modalities during inference, especially where priority conflicts arise. For instance, a system generating video narrations may need to resolve conflicts between high-precision audio cues and stylistic textual descriptions.

A key aspect of multimodal controllability hinges on attributing and aligning the semantic intensity of different modalities. Recent innovations in constrained decoding approaches, such as NeuroLogic Decoding [10], significantly enhance lexical and attribute-level enforceability. Nevertheless, applying these techniques across multimodal datasets reveals the complexity of achieving semantic groundedness while managing trade-offs in fluency and diversity. Similarly, consistency across modalities remains a bottleneck, with challenges emerging in use cases requiring coordinated manipulation of text and visual components, such as personalized storytelling or video caption optimization.

Promising applications of multimodal controllability extend beyond entertainment into domains like accessibility, where cross-modal systems can generate audio-guided textual descriptions for visually impaired users. Techniques like energy-based rewriting for constraint-driven editing [58] enable precise realignment of outputs without compromising nontextual integrity. However, these advancements remain limited by the availability of robust evaluation frameworks that holistically assess multimodal generation fidelity, as noted in CTRLEval [13]. Future benchmarking initiatives must incorporate metrics that capture inter-modal coherence, adherence to constraints, and user satisfaction.

Looking forward, the increasing convergence of multimodal pretrained models and domain-specific fine-tuning offers scope for creative breakthroughs. Approaches synthesizing transferable multimodal embeddings with causal frameworks, as proposed in A Causal Lens for Controllable Text Generation [25], can enable dynamic reasoning across modalities. Innovations in progressive generation strategies [91], when fused with multimodal contexts, show potential for scaling coherence in long-form generation tasks involving cross-modal interactions. Nonetheless, addressing computational overhead and scaling to low-resource or data-scarce domains will require open-source datasets and efficient retrieval-augmented generation pipelines.

Achieving seamless multimodal controllability represents a confluence of challenges in machine learning, data representation, and user interaction, necessitating interdisciplinary collaboration. This domain offers fertile ground for future research, with opportunities to refine multimodal adaptation frameworks, extend cross-modal applications, and enhance personalization. As research progresses, these endeavors are likely to redefine how machines interpret and generate information across diverse media, steering the development of next-generation interactive systems.

8.2 Real-Time Adaptability and Dynamic Feedback Systems

The increasing demand for nuanced, user-responsive systems has catalyzed advancements in real-time adaptability and dynamic feedback mechanisms in controllable text generation (CTG) for large language models (LLMs). These innovations facilitate the iterative refinement of outputs based on user feedback or progressively evolving contextual demands during interaction. By bridging human-machine collaboration with enhanced adaptability, these systems aim to elevate personalization, responsiveness, and overall quality in natural language generation.

At the heart of real-time adaptability are low-latency control strategies that enable models to promptly incorporate user feedback while sustaining fluency and semantic coherence. Techniques such as lightweight prompt tuning and reinforcement learning-driven approaches have proven instrumental in this regard. For example, reinforcement learning with human feedback (RLHF) optimizes generation policies through iterative alignment with contextual or user-specified constraints, leveraging reward models to guide output dynamically [108]. However, the scalability of RLHF often encounters practical challenges, as its reliance on intensive computational resources for continual training and policy updates limits real-time efficiency.

To address these challenges, parameter-efficient methods—such as prefix-tuning and LoRA (Low-Rank Adaptation)—have emerged as promising alternatives. These techniques achieve fine-grained adaptability with minimal updates to model parameters, thereby ensuring computational feasibility for repeated adjustments [49], [109]. While these methods significantly enhance responsiveness, they exhibit limitations in handling multi-attribute controls in dynamic, high-dimensional settings, underscoring the need for more versatile solutions.

Interactive refinement pipelines further advance real-time controllability by leveraging user-in-the-loop frameworks that iteratively tailor outputs based on feedback. Decoding-time interventions—using methods such as constrained beam search or energy-based optimization—play a pivotal role in this iterative process [10], [58]. Critic-Control frameworks, for instance, incorporate post-decoding critics to evaluate and refine generated text against defined metrics, such as stylistic consistency or disallowed content [24]. However, these approaches tend to falter in multi-turn interactions, where the recursive influence of prior outputs on ongoing refinement becomes necessary but inadequately supported.

A complementary advancement lies in retrieval-augmented generation (RAG) mechanisms, which integrate external memory systems to dynamically align generated outputs with contextual or user-provided requirements. By fetching relevant knowledge during interaction, RAG pipelines enhance factual grounding and real-time reliability in knowledge-intensive scenarios [48], [50]. Despite these advantages, the integration of retrieval systems can lead to increased latency and greater architectural complexity, especially when interfacing with expansive or evolving knowledge bases.

Emerging approaches aim to integrate explainable and transparent mechanisms into the feedback loop, providing dynamic insights into how user inputs influence outputs. For instance, models leveraging dynamic attribute graphs (DAGs) or probabilistic causal frameworks offer a structured way to modulate attributes like tone, sentiment, or inclusivity in real time [12], [25]. This transparency proves valuable in multi-attribute and multi-dimensional control, paving the way for more accountable and adaptable generation systems.

Despite significant progress, several challenges persist. One of the key issues involves optimizing trade-offs between competing constraints during interaction, where balancing parameters like semantic coherence, stylistic fidelity, and user-intended adjustments can be highly complex [62]. Additionally, designing intuitive and robust mechanisms for capturing ambiguous or conflicting user feedback remains an open problem. Equally pressing is the challenge of achieving low-latency, high-quality generation in resource-constrained settings, where computational and data limitations exacerbate systemic inefficiencies.

Looking forward, further development in hybrid frameworks is crucial to achieve scalable and flexible CTG systems. Integrating efficient parameter tuning with mechanisms for real-time retrieval, probabilistic editing, and energy-based evaluation offers an intriguing direction. Extending these advances to multimodal contexts, such as incorporating visual or auditory feedback, can further enrich adaptability and contextual alignment, enhancing cross-modal interaction and personalization.

Real-time adaptability occupies a pivotal role in advancing controllable text generation toward practical, user-centered applications. While existing methods showcase substantial promise, addressing the interplay of adaptability, efficiency, and interpretability remains a pressing frontier. Filling these gaps will pave the way for more versatile and collaborative text generation systems, reshaping how users interact with AI technologies in dynamic, real-world environments.

8.3 Enhancing Cross-Lingual and Low-Resource Controllability

Achieving robust and effective controllability in cross-lingual and low-resource language settings represents one of the most pivotal challenges in extending the global applicability of controllable text generation. Large language models (LLMs), while demonstrating exceptional performance in high-resource languages, often falter in low-resource or multilingual settings due to limited training data, cultural nuances, and linguistic variability. Addressing these deficits is essential for ensuring equitable access and enhancing the inclusivity of such technologies.

Cross-lingual transfer learning has emerged as a promising direction for enabling controllability across a wide array of languages, especially under low-resource settings. Pre-trained multilingual models such as mBERT and XLM-R have demonstrated the ability to transfer knowledge from high-resource to low-resource languages by utilizing shared subword tokenization and semantic embeddings. This foundational idea is further extended to controllable text generation through models like CTRL [3] and FUDGE [7], which allow conditioning on attributes. However, effective transfer, particularly for nuanced controls such as sentiment or stylistic adjustments, remains limited due to linguistic disparities. The recent work employing unified decoder architectures optimized for shared latent representations across languages has shown some improvement, enabling attributes such as emotional tone or formality to generalize across linguistic boundaries [110]. These methods provide an avenue for control-aligned transfer learning but require fine-tuning large multilingual datasets or attribute-specific auxiliary training, which remains resource-intensive.

Data-efficient techniques, such as synthetic data generation and active learning, are critical for improving performance on low-resource language controllability. Approaches that augment low-resource language datasets dynamically using parallel high-resource corpora have been shown to significantly improve controllability without extensive reliance on annotated data [51]. Similarly, back-translation with attribute-specific reconstruction losses has emerged as an effective tool for augmenting low-resource datasets [111]. These techniques concurrently preserve content and enable transformations in attributes such as sentiment or formal register. Nevertheless, issues of overfitting synthetic distributions arise, which can negatively affect model generalization to real-world inputs.

Cultural and ethical considerations are particularly significant in multilingual and cross-lingual controllability. For instance, models trained predominantly on high-resource data may inadvertently enforce stylistic or semantic biases, failing to capture cultural nuances embedded within low-resource languages. Methods such as culturally adaptive pre-training, which fine-tune models using culturally aligned corpora, have shown promise, yet challenges persist in balancing cultural representation with standardized controllability [72]. Moreover, new diagnostic tools designed to evaluate fairness and equity across linguistic dimensions, such as syntactic consistency and semantic alignment in cross-lingual generation tasks, remain necessary [112].

Dynamic attribute graphs (DAGs) and latent-space optimization techniques represent emerging paradigms for improving controllability across resource-scarce scenarios. Utilizing attribute graphs to modulate textual properties (e.g., tone or inclusivity) without necessitating explicit retraining allows for efficient adaptation to low-resource settings [12]. Additionally, incorporating parameter-efficient techniques such as prefix tuning and adapters permits simultaneous control across multiple languages with minimal compute requirements [89]. For instance, controlling stylistic tone across Hindi and Bengali while maintaining grammatical integrity has benefited from such modular approaches [81].

Future advancements in enhancing cross-lingual controllability will likely involve more sophisticated methods for disentangling linguistic attributes and language-agnostic representations. Techniques utilizing causal modeling [25] may provide a theoretical basis for isolating control-specific factors across languages, enabling fine-grained edits without impacting core linguistic structures. Moreover, incorporating unsupervised alignment techniques for emerging low-resource languages potentially aligns stylistic generation with cultural fidelity through latent-space contrastive optimization [97].

As multilingual applications continue to evolve, the integration of these methods through unified architectural frameworks, equitable evaluation benchmarks, and scalable optimization strategies will play a pivotal role in realizing globally relevant and culturally inclusive controllable text generation systems. Such advancements will not only deepen the applicability of LLMs but will also democratize access to advanced text generation tools across previously underserved linguistic communities.

8.4 Explainability and Transparency in Controlled Generation

The emergence of controllable text generation as a field of study is deeply intertwined with the inherent opacity of large language models (LLMs), which poses significant challenges for understanding and verifying how controllability constraints influence their outputs. This lack of explainability not only undermines user trust but also hampers debugging during model development and complicates the deployment of these systems in sensitive domains such as medical documentation or legal text generation. Addressing these challenges requires the introduction of interpretability mechanisms that enable stakeholders to gain actionable insights into how models balance adherence to constraints with output fluency, coherence, and contextual relevance.

One of the most promising paths toward improving transparency in controllable text generation involves the use of structural causal models (SCMs), which provide a principled framework for disentangling the causal effects of input constraints—such as stylistic tone, sentiment, or length—on generated outputs. By enabling interventions, counterfactual reasoning, and "what-if" analyses, SCMs make it possible to decompose and assess the contributions of specific attributes to controllability. For instance, causal approaches have been employed to unify attribute-conditional generation with text attribute transfer tasks, disentangling confounding factors to achieve more accurate control and reduce bias [25]. Nevertheless, these methods are computationally intensive and depend on domain expertise to encode causal structures effectively, raising concerns about their scalability in practical applications.

Complementary to causal approaches are post-hoc explainability techniques, such as attention-based interpretability methods and counterfactual explanations, which offer more accessible ways to evaluate and understand controllable text generation systems. Attention mechanisms, for example, have been explored to visualize how tokens, input parameters, or prompts influence model outputs, with dynamic attention calibration mechanisms proposed to optimize attention flows in real-time for improved control alignment without sacrificing fluency [113]. However, the interpretability of attention maps remains contentious, given that they often fail to establish true causal relationships in the decision-making process. Counterfactual explanations, where small perturbations to input prompts or constraints are analyzed to assess their impact on outputs, provide another avenue for interpretability. These methods have shown particular promise in applications like toxicity mitigation and bias reduction, though they remain underexplored in more complex settings involving multiple, potentially competing constraints [43].

Transparency solutions also extend to latent space-based methods, which focus on control within interpretable, compact latent representations rather than directly in the high-dimensional text sequence space. Researchers have developed latent variable models to disentangle specific attributes during generation, enabling precise manipulations while maintaining the interpretability of intermediate representations [54]. Latent-guided gradient-based optimization further refines this process, enhancing controllability and transparency simultaneously [69]. However, these methodologies often introduce trade-offs, such as requiring increased computational resources during inference, limiting their feasibility for real-time or large-scale applications.

Hybrid frameworks, which combine explainability with real-time user interaction, are also gaining momentum. Such frameworks enable users to iteratively refine generated outputs by providing feedback and dynamically visualizing attribute configurations [56]. These systems democratize the use of controllable text generation by making it accessible to non-expert users in fields such as creative writing and personalized storytelling. However, they also highlight the need for robust evaluation mechanisms to reconcile the sometimes subjective expectations of human users with system-defined control criteria.

Looking to the future, the creation of universal benchmarks and evaluation metrics will play a pivotal role in advancing transparency for controllable text generation. Among recent proposals is CoDI-Eval, a benchmark designed to evaluate controllability, fluency, and semantic coherence under diversified instructions while incorporating interpretability metrics to assess the alignment of control parameters with generated attributes [84]. Building on such efforts, the inclusion of causal modeling and post-hoc methods in evaluation frameworks could provide richer transparency insights, establishing standardized protocols to assess and enhance explainability in controlled generation systems.

In conclusion, achieving robust explainability and transparency in controllable text generation demands a multi-faceted strategy rooted in causal modeling, attention-based and counterfactual techniques, latent space exploration, and hybrid user interaction frameworks. These advancements are instrumental not only for fostering trust and usability in high-stakes contexts but also for overcoming technical barriers in scaling these systems to real-world, multi-constraint scenarios. As the field evolves, integrating interpretability mechanisms into evaluation pipelines and refining hybrid approaches will be crucial for realizing controllable systems that are both precise and reliably comprehensible.

8.5 Development of Universal Evaluation Benchmarks

The evaluation of controllable text generation (CTG) remains a critical bottleneck in advancing the field due to the lack of standardized and universal benchmarks that address the multidimensional nature of CTG. Existing evaluation frameworks often focus narrowly on specific dimensions, such as fluency or semantic coherence, while neglecting the interplay of diverse controllability constraints, including stylistic adherence, factuality, and alignment with desired attributes. A universal evaluation benchmark for CTG must establish a holistic and systematic approach that can comprehensively assess generation quality, constraint adherence, and usability across a wide range of tasks, domains, and user contexts.

A major challenge in developing such benchmarks is incorporating multidimensional evaluation metrics that reflect not only the intrinsic quality of generated text—such as fluency, grammaticality, and language novelty—but also the degree of control fidelity, which measures how accurately outputs adhere to predefined constraints. Current advancements in retrieval-augmented systems highlight the utility of task-specific metrics for factual accuracy [50], while efforts in reinforcement learning frameworks have introduced reward models tailored for specific dimensions of controllability, such as sentiment alignment or safety [105], [114]. However, these task- or attribute-specific metrics fall short of offering generalizability across diverse CTG applications, limiting their utility as universal evaluation protocols.

Crucially, the emerging paradigm of hybrid evaluation frameworks—combining automated metrics with human-centric criteria—has demonstrated considerable promise in addressing concerns about the subjectivity of control dimensions like creativity and emotional tone. Recent studies suggest that large language models (LLMs) themselves can serve as semi-automated evaluators, leveraging their interpretive capacities to assess alignment with stylistic, structural, and factual constraints [82], [115]. While these approaches have shown moderate success in correlating with human judgments, they often encounter biases resulting from prompt sensitivity and lack of robustness. Future frameworks must optimize such LLM-based evaluative capabilities while addressing their limitations through prompt optimization and calibration techniques [116].

Another key consideration is evaluating the trade-offs inherent in balancing conflicting constraints, such as fine-tuning outputs for sentiment while maintaining factual coherence or abstract stylistic consistency. Methods like energy-based models, which allow for constraint optimization during decoding, show potential for benchmarking such trade-offs with greater granularity [36], [58]. Furthermore, contrastive decoding has emerged as an effective technique to test the robustness of models in balancing parametric knowledge and contextual prompt adherence [65].

Another unresolved aspect lies in curating benchmark datasets that comprehensively capture the diversity of CTG tasks. Although datasets tailored for specific attributes, like toxicity mitigation or stylistic modulation, enhance focused evaluations [12], [20], general-purpose benchmarks remain scarce. Universal datasets must encompass multi-attribute control challenges, bridging low-resource and high-resource task settings while mitigating dataset-centered biases [85]. Cross-lingual evaluation is particularly underexplored; addressing linguistic and cultural variations in multilingual CTG tasks could yield inclusive benchmarks fostering global applicability [83].

The integration of human-machine hybrid methods offers a realistic path forward for universal evaluation. Methods like self-refinement, where models iteratively improve themselves based on automated or human feedback, could streamline evaluation protocols [38]. Additionally, the development of evaluative pipelines such as ProxyQA, which connects high-level metrics with domain-specific proxy questions, provides scalable and interpretable frameworks for assessing long-form and domain-specific generation outputs [117].

Key future directions involve the creation of modular and adaptive benchmarks that not only scale with evolving tasks and CTG paradigms but also integrate explainability mechanisms, offering transparency into metric scores and alignment to intended constraints. Establishing universal protocols to evaluate interdependencies between attributes, robustness across domains, and models' adaptability to user feedback is vital for driving the future of CTG evaluation frameworks. These advancements will be instrumental in closing the gap between task-specific evaluations and holistic, universally accepted benchmarks that support the broader adoption and ethical deployment of controllable text generation technologies.

8.6 Advances in Personalized and Idiosyncratic Control Mechanisms

Personalization in controllable text generation (CTG) offers transformative opportunities to tailor large language models (LLMs) to specific user preferences, enabling outputs that align with unique stylistic, linguistic, and contextual requirements. This dynamic adaptability necessitates methodologies that balance computational efficiency with the generalizability required to accommodate diverse user populations and task paradigms.

Historically, personalization in LLMs has relied heavily on fine-tuning, where models are adapted to individual users through retraining on small, user-specific datasets. While effective, this approach is limited by high computational costs, resource inefficiency, and scalability issues, particularly in large-scale deployment scenarios. To address these constraints, parameter-efficient training methods such as prefix-tuning and low-rank adaptations (LoRA) have gained significant traction [4]. By modifying only a subset of model parameters while retaining access to pre-trained representations, these methods enable personalized text generation with reduced computational overhead. For instance, LoRA focuses on fine-tuning rank-decomposed layers within the attention mechanism to encode users' stylistic or tonal preferences, striking a crucial balance between personalization performance and resource efficiency.

Effective personalization further demands the ability to capture the nuanced linguistic attributes of individual users, such as preferred syntax, lexicon, and stylistic tones. Disentangled latent variable models have been proposed as a promising solution, allowing representation spaces that facilitate control over finely grained linguistic dimensions [118]. These models disentangle elements such as tone, formality, and emotional inclination, enabling multi-attribute personalization. For example, hierarchical variational autoencoders (VAEs) effectively represent both macro-level stylistic types and micro-level attribute variations, providing a robust foundation for fine-tuned customization [119].

The emergence of dynamic, real-time adaptation marks another critical advancement in personalization. User-in-the-loop models enable iterative refinement of outputs based on real-time feedback, creating a continuous alignment between generated text and user preferences. Techniques such as reinforcement learning (RL) with continuous reward optimization and successor feature architectures have shown promise in this area [67]. However, achieving seamless responsiveness while ensuring text fluency remains a significant technical challenge, often complicated by latency constraints.

Another important direction involves hybrid frameworks that integrate external knowledge to enhance control tailored to user individuality. Dynamic attribute graphs (DAGs), for example, allow for the modulation of specific attribute keywords in real time, aligning outputs with personalized stylistic goals without compromising fluency [12]. Similarly, retrieval-augmented models leverage user-provided exemplars or external content to ground generated text in personalized contexts. These approaches effectively bypass the need for extensive retraining cycles, thereby enabling more scalable and efficient adaptation.

Personalization inevitably intersects with multi-attribute control, which introduces the challenge of harmonizing logical coherence and stylistic consistency across multiple dimensions simultaneously. Advanced composition-based techniques offer potential solutions, employing energy-based formulations or ordinary differential equations (ODEs) in latent spaces to balance various control demands. Methods such as composable latent operations redefine text generation as a search for intersections between attribute-specific distributions, allowing models to manage complex stylistic, topical, and syntactic requirements [69]. These frameworks are particularly crucial for applications where user profiles evolve or span multiple dimensions, requiring adaptable yet consistent outputs.

Despite the technological advancements in personalized CTG, ethical and technical challenges persist. User-driven fine-tuning risks amplifying biases or reinforcing harmful stereotypes, raising concerns about fairness and inclusion. Techniques such as counterfactual data augmentation and causal modeling have been employed to address these issues, disentangling spurious correlations to ensure equitable personalization [25]. Additionally, the opacity of personalized adaptations poses questions about explainability. Post-hoc interpretability tools, such as attention visualizations or auxiliary scoring mechanisms, can improve accountability and user trust without undermining model performance.

Future directions in personalized CTG demand the development of lifelong learning architectures capable of dynamically evolving alongside user interactions. Memory-aware models that encode long-term user preferences while mitigating issues like data drift and overfitting represent a promising path forward [35]. Beyond text, integrating multimodal personalization—such as incorporating voice, images, or behavioral data—can further expand the reach of adaptive systems. Parallel efforts are needed in constructing dedicated evaluation frameworks to rigorously benchmark personalization efficacy, measuring stylistic and contextual adaptability across diverse use cases.

In summary, personalization in CTG is redefining the capabilities of LLMs by enabling contextually and stylistically tailored outputs. Advances in adaptive training methods, multi-attribute control, and ethical personalization frameworks underscore the growing potential of user-centric generation systems. However, realizing this vision hinges on addressing scalability, fairness, and transparency challenges, requiring continued innovation across technical and interdisciplinary research domains.

9 Conclusion

This survey has comprehensively reviewed the landscape of controllable text generation (CTG) in the context of large language models (LLMs), delving into methods, challenges, evaluation paradigms, applications, and ethical considerations. The field has witnessed significant advancements, transitioning from rudimentary control approaches reliant on fine-tuned models to sophisticated, dynamic techniques capable of orchestrating multiple attributes simultaneously while maintaining fluency and coherence in generated text.

Core methods such as prompt engineering [4], prefix-based approaches [17], and fine-tuning frameworks [3] have demonstrated remarkable flexibility and modularity. Advanced decoding paradigms, including energy-based methods [6] and attribute-guided optimization [11], have shown efficacy in balancing tight constraints with high-quality natural language outputs. Architectural innovations such as control codes and memory mechanisms have further enabled intrinsic control, as exemplified by approaches like ControlGAN [71] and Knowledge-Aware Fine-Tuning (KAFT) [35]. Notably, hybrid methods integrating retrieval-based augmentation and symbolic reasoning [45] have advanced control in domain-specific, low-resource, or multimodal contexts.

Despite these advancements, challenges remain pervasive across technical, interpretability, and ethical dimensions. Technically, the scalability of multi-attribute control remains a barrier, with inherent trade-offs between control fidelity and fluency [10]. For instance, the strict enforcement of hard constraints often leads to degenerated diversity or overfitting [10]. Additionally, hallucination—the propensity to generate unfactual or misleading text—persists as a daunting issue, particularly in knowledge-grounded tasks [9]. Methods such as retrieval-augmented generation have emerged as promising mitigations [18], but they are not yet robustly generalizable.

Ethical and societal challenges are equally pressing. Bias in training data continues to influence CTG outputs, with frameworks like BOLD [15] underscoring the systemic nature of biases across gender, race, and sociopolitical dimensions. Similarly, controlling outputs to avoid toxicity or promote inclusivity remains an unresolved frontier, with approaches like anti-experts in DExperts [11] illustrating early but incomplete solutions. Furthermore, the potential misuse of CTG technologies for generating disinformation or deepfakes raises significant regulatory concerns, demanding reinforced safeguard mechanisms that go beyond traditional content filtering [120].

Evaluation methodologies, despite progress in incorporating LLM-assisted metrics [121], require standardization and greater interpretability. Metrics like CTRLEval [13] attempt to provide unsupervised, reference-free assessments of controllability and fluency, but subjective aspects like creativity and emotional tone remain challenging to quantify systematically. Additionally, the over-reliance on automated versus human-centric evaluation can introduce biases into benchmarks, as evidenced by discrepancies in LLM-based evaluators' consistency during abstractive summarization tasks [86].

Future research must focus on integrating more nuanced, interpretable control mechanisms while addressing bottlenecks in scalability and ethical accountability. Leveraging causal inference frameworks [25], developing universal multitask evaluation benchmarks [14], and incorporating multimodal modalities like text-to-image or text-to-audio generation [71] are promising directions. Additionally, personalization at scale, as explored in Persona-driven frameworks like MEGATRON-CNTRL [8], will redefine how CTG systems adapt to user-specific constraints in real-time.

In conclusion, controllable text generation holds transformative potential in advancing safer, more coherent, and purpose-driven natural language systems. However, an interdisciplinary, collaborative effort is indispensable for tackling open challenges, ensuring that the field evolves towards robust, ethical, and human-centric solutions. With continuous exploration of emerging trends such as dynamic adaptive mechanisms [56] and energy-efficient optimizations [58], the future of CTG is poised to reshape both industrial applications and academic paradigms.

REFERENCES

[1] R. Lebret, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain," in *Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1203–1213. 1

[2] J. Li, T. Tang, W. X. Zhao, and J. rong Wen, "Pretrained language models for text generation: A survey," *ArXiv*, vol. abs/2105.10311, 2021. 1, 25, 29

[3] N. Keskar, B. McCann, L. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *ArXiv*, vol. abs/1909.05858, 2019. 1, 4, 8, 11, 15, 21, 23, 28, 29, 31, 34

[4] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," *ArXiv*, vol. abs/1912.02164, 2019. 1, 2, 6, 10, 14, 15, 19, 23, 24, 28, 33, 34

[5] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, vol. 56, pp. 1 – 37, 2022. 1, 2, 6, 16, 23, 25, 26

[6] L. Qin, S. Welleck, D. Khashabi, and Y. Choi, "Cold decoding: Energy-based constrained text generation with langevin dynamics," *ArXiv*, vol. abs/2202.11705, 2022. 1, 2, 10, 15, 24, 25, 28, 29, 34

[7] K. Yang and D. Klein, "Fudge: Controlled text generation with future discriminators," *ArXiv*, vol. abs/2104.05218, 2021. 1, 6, 7, 12, 14, 15, 24, 31

[8] P. Xu, M. Patwary, M. Shoeybi, R. Puri, P. Fung, A. Anandkumar, and B. Catanzaro, "Controllable story generation with external knowledge using large-scale language models," in *Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 2831–2845. 1, 2, 4, 20, 24, 34

[9] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia, "Cognitive mirage: A review of hallucinations in large language models," *ArXiv*, vol. abs/2309.06794, 2023. 1, 34

[10] X. Lu, P. West, R. Zellers, R. L. Bras, C. Bhagavatula, and Y. Choi, "Neurologic decoding: (un)supervised neural text generation with predicate logic constraints," *ArXiv*, vol. abs/2010.12884, 2020. 1, 2, 3, 10, 11, 15, 16, 20, 24, 29, 30, 34

[11] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, and Y. Choi, "Dexperts: Decoding-time controlled text generation with experts and anti-experts," in *Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 6691–6706. 1, 3, 10, 20, 21, 25, 34

[12] X. Liang, H. Wang, S. Song, M. Hu, X. Wang, Z. Li, F. Xiong, and B. Tang, "Controlled text generation for large language model with dynamic attribute graphs," *ArXiv*, vol. abs/2402.11218, 2024. 1, 2, 3, 4, 5, 6, 7, 9, 12, 13, 17, 18, 19, 20, 21, 23, 24, 25, 26, 28, 29, 30, 31, 33

[13] P. Ke, H. Zhou, Y. Lin, P. Li, J. Zhou, X. Zhu, and M. Huang, "Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation," in *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 2306–2319. 2, 15, 16, 30, 34

[14] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018. 2, 10, 34

[15] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta, "Bold: Dataset and metrics for measuring biases in open-ended language generation," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021. 2, 6, 25, 34

[16] X. Lu, S. Welleck, P. West, L. Jiang, J. Kasai, D. Khashabi, R. L. Bras, L. Qin, Y. Yu, R. Zellers, N. A. Smith, and Y. Choi, "Neurologic a*esque decoding: Constrained text generation with lookahead heuristics," in *North American Chapter of the Association for Computational Linguistics*, 2021, pp. 780–799. 2, 4, 10, 13, 15, 22, 27

[17] J. Qian, L. Dong, Y. Shen, F. Wei, and W. Chen, "Controllable natural language generation with contrastive prefixes," in *Findings*, 2022, pp. 2912–2924. 2, 3, 6, 7, 11, 15, 20, 21, 22, 25, 27, 34

[18] H. Rashkin, V. Nikolaev, M. Lamm, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, "Measuring attribution in natural language generation models," *Computational Linguistics*, vol. 49, pp. 777–840, 2021. 2, 15, 34

[19] B. Krause, A. D. Gotmare, B. McCann, N. Keskar, S. R. Joty, R. Socher, and N. Rajani, "Gedi: Generative discriminator guided sequence generation," in *Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 4929–4952. 3, 7, 10, 21, 25, 26

[20] B. Alhafni, V. Kulkarni, D. Kumar, and V. Raheja, "Personalized text generation with fine-grained linguistic control," *ArXiv*, vol. abs/2402.04914, 2024. 3, 21, 33

[21] C. Zheng, P. Ke, Z. Zhang, and M. Huang, "Click: Controllable text generation with sequence likelihood contrastive learning," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 1022–1040. 3, 8, 10, 17

[22] S. Narayan, G. Simoes, Y. Zhao, J. Maynez, D. Das, M. Collins, and M. Lapata, "A well-composed text is half done! composition sampling for diverse conditional generation," in *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 1319–1339. 3, 16

[23] A. Chan, Y. Ong, B. Pung, A. Zhang, and J. Fu, "Cocon: A self-supervised approach for controlled text generation," *ArXiv*, vol. abs/2006.03535, 2020. 3, 21

[24] M. Kim, H. Lee, K. M. Yoo, J. Park, H. Lee, and K. Jung, "Critic-guided decoding for controlled text generation," in *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 4598–4612. 3, 5, 7, 8, 13, 21, 29, 30

[25] Z. Hu and E. L. Li, "A causal lens for controllable text generation," in *Neural Information Processing Systems*, 2022, pp. 24 941–24 955. 3, 4, 7, 15, 17, 18, 19, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 34

[26] A. Kumar, K. Ahuja, R. Vadapalli, and P. Talukdar, "Syntax-guided controlled generation of paraphrases," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 330–345, 2020. 4, 10, 15, 21, 24

[27] J. Guan, X. Mao, C. Fan, Z. Liu, W. Ding, and M. Huang, "Long text generation by modeling sentence-level and discourse-level coherence," *ArXiv*, vol. abs/2105.08963, 2021. 4

[28] K. Xie and M. Riedl, "Creating suspenseful stories: Iterative planning with large language models," in *Conference of the European Chapter of the Association for Computational Linguistics*, 2024, pp. 2391–2407. 4

[29] Z. Hu, H. P. Chan, J. Liu, X. Xiao, H. Wu, and L. Huang, "Planet: Dynamic content planning in autoregressive transformers for long-form text generation," *ArXiv*, vol. abs/2203.09100, 2022. 4, 12, 21, 26, 27

[30] Y. Gu, X. Feng, S. Ma, L. Zhang, H. Gong, W. Zhong, and B. Qin, "Controllable text generation via probability density estimation in the latent space," *ArXiv*, vol. abs/2212.08307, 2022. 4, 21

[31] R. Puduppully, L. Dong, and M. Lapata, "Data-to-text generation with content selection and planning," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 6908–6915. 4

[32] R. Puduppully and M. Lapata, "Data-to-text generation with macro planning," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 510–527, 2021. 4, 27

[33] A. Moryossef, Y. Goldberg, and I. Dagan, "Step-by-step: Separating planning from realization in neural data-to-text generation," *ArXiv*, vol. abs/1904.03396, 2019. 4, 27

[34] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8299–8308, 2018. 5, 22

[35] D. Li, A. Rawat, M. Zaheer, X. Wang, M. Lukasik, A. Veit, F. X. Yu, and S. Kumar, "Large language models with controllable working memory," *ArXiv*, vol. abs/2211.05110, 2022. 5, 23, 34

[36] S. Kumar, E. Malmi, A. Severyn, and Y. Tsvetkov, "Controlled text generation as continuous optimization with multiple con-

straints," in *Neural Information Processing Systems*, 2021, pp. 14 542–14 554. 5, 7, 8, 9, 12, 13, 26, 27, 28, 33

[37] M. Khalifa, H. ElSahar, and M. Dymetman, "A distributional approach to controlled text generation," *ArXiv*, vol. abs/2012.11635, 2020. 5, 8, 11, 12, 14, 24, 25

[38] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Welleck, B. P. Majumder, S. Gupta, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," *ArXiv*, vol. abs/2303.17651, 2023. 5, 9, 18, 33

[39] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *ArXiv*, vol. abs/2205.02655, 2022. 5

[40] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "Controlvideo: Training-free controllable text-to-video generation," *ArXiv*, vol. abs/2305.13077, 2023. 5

[41] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, S. Ermon, Y. Fu, and R. Xu, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *ArXiv*, vol. abs/2305.11147, 2023. 5, 19

[42] Y. Zhu, Z. Huang, Z. Dou, and J.-R. Wen, "One token can help! learning scalable and pluggable virtual tokens for retrieval-augmented large language models," *ArXiv*, vol. abs/2405.19670, 2024. 5

[43] W. Zhou, Y. Jiang, E. G. Wilcox, R. Cotterell, and M. Sachan, "Controlled text generation with natural language instructions," *ArXiv*, vol. abs/2304.14293, 2023. 6, 18, 19, 22, 23, 25, 28, 32

[44] D. Pu and V. Demberg, "Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 1–18. 6, 8, 15, 17, 18, 26

[45] F. Mireshghallah, K. Goyal, and T. Berg-Kirkpatrick, "Mix and match: Learning-free controllable text generationusing energy language models," *ArXiv*, vol. abs/2203.13299, 2022. 6, 12, 15, 17, 20, 27, 34

[46] T. Niu, C. Xiong, S. Yavuz, and Y. Zhou, "Parameter-efficient detoxification with contrastive decoding," *ArXiv*, vol. abs/2401.06947, 2024. 7

[47] L. Logeswaran, H. Lee, and S. Bengio, "Content preserving text generation with attribute controls," *ArXiv*, vol. abs/1811.01135, 2018. 7, 16, 17, 25

[48] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Lidén, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *ArXiv*, vol. abs/2302.12813, 2023. 7, 9, 11, 16, 26, 28, 30

[49] M. Li, H. Chen, C. Wang, D. Nguyen, D. Li, and T. Zhou, "Ruler: Improving llm controllability by rule-based data recycling," *ArXiv*, vol. abs/2406.15938, 2024. 7, 30

[50] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *ArXiv*, vol. abs/2005.11401, 2020. 7, 9, 14, 16, 19, 23, 26, 30, 32

[51] Z. Yang, Z. Hu, C. Dyer, E. Xing, and T. Berg-Kirkpatrick, "Unsupervised text style transfer using language models as discriminators," *ArXiv*, vol. abs/1805.11749, 2018. 8, 31

[52] J. Pei, K. Yang, and D. Klein, "Preadd: Prefix-adaptive decoding for controlled text generation," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 10 018–10 037. 8

[53] A. Bhargava, C. Witkowski, M. Shah, and M. W. Thomson, "What's the magic word? a control theory of llm prompting," *ArXiv*, vol. abs/2310.04444, 2023. 8, 17

[54] L. Fang, T. Zeng, C.-N. Liu, L. Bo, W. Dong, and C. Chen, "Transformer-based conditional variational autoencoder for controllable story generation," *ArXiv*, vol. abs/2101.00828, 2021. 8, 32

[55] L. F. R. Ribeiro, Y. Zhang, and I. Gurevych, "Structural adapters in pretrained language models for amr-to-text generation," *ArXiv*, vol. abs/2103.09120, 2021. 8, 27

[56] W. Zhou, Y. Jiang, P. Cui, T. Wang, Z. Xiao, Y. Hou, R. Cotterell, and M. Sachan, "Recurrentgpt: Interactive generation of (arbitrarily) long text," *ArXiv*, vol. abs/2305.13304, 2023. 8, 20, 22, 25, 32, 35

[57] L. Beurer-Kellner, M. Fischer, and M. T. Vechev, "Guiding llms the right way: Fast, non-invasive constrained generation," *ArXiv*, vol. abs/2403.06988, 2024. 8

[58] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 922–10 931, 2023. 9, 11, 13, 14, 15, 19, 23, 27, 28, 30, 33, 35

[59] Z. Li, B. Peng, P. He, M. Galley, J. Gao, and X. Yan, "Guiding large language models via directional stimulus prompting," *ArXiv*, vol. abs/2302.11520, 2023. 9

[60] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. 9, 13, 23

[61] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7001–7011, 2021. 9, 19, 23

[62] Y. Gu, X. Feng, S. Ma, L. Zhang, H. Gong, and B. Qin, "A distributional lens for multi-aspect controllable text generation," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1023–1043. 11, 16, 22, 27, 30

[63] Z. Wu, M. Galley, C. Brockett, Y. Zhang, X. Gao, C. Quirk, R. Koncel-Kedziorski, J. Gao, H. Hajishirzi, M. Ostendorf, and B. Dolan, "A controllable model of grounded response generation," *ArXiv*, vol. abs/2005.00613, 2020. 11, 26

[64] D. Wang, N. Jojic, C. Brockett, and E. Nyberg, "Steering output style and topic in neural response generation," *ArXiv*, vol. abs/1709.03010, 2017. 13

[65] Y. Su and N. Collier, "Contrastive search is what you need for neural text generation," *ArXiv*, vol. abs/2210.14140, 2022. 14, 33

[66] A. M. Turner, L. Thiergart, G. Leech, D. S. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid, "Steering language models with activation engineering," 2023. 14

[67] K. de Langis, R. Koo, and D. Kang, "Reinforcement learning with dynamic multi-reward weighting for multi-style controllable generation," *ArXiv*, vol. abs/2402.14146, 2024. 14, 20, 24, 28, 33

[68] Y. Liu, X. Liu, X. Zhu, and W. Hu, "Multi-aspect controllable text generation with disentangled counterfactual augmentation," in *Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 9231–9253. 14, 24, 29

[69] G. Liu, Z. Feng, Y. Gao, Z. Yang, X. Liang, J. Bao, X. He, S. Cui, Z. Li, and Z. Hu, "Composable text controls in latent space with odes," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 16 543–16 570. 15, 27, 29, 32, 34

[70] A. Madotto, E. Ishii, Z. Lin, S. Dathathri, and P. Fung, "Plug-and-play conversational models," *ArXiv*, vol. abs/2010.04344, 2020. 15

[71] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "Controllable text-to-image generation," *ArXiv*, vol. abs/1909.07083, 2019. 15, 20, 29, 34

[72] A. Liu, M. T. Diab, and D. Fried, "Evaluating large language model biases in persona-steered generation," *ArXiv*, vol. abs/2405.20253, 2024. 15, 17, 20, 31

[73] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, and D. Yang, "Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation," *ArXiv*, vol. abs/2310.15638, 2023. 16

[74] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro, "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model," *ArXiv*, vol. abs/2201.11990, 2022. 16, 27

[75] T. Buz, B. Frost, N. Genchev, M. Schneider, L.-A. Kaffee, and G. de Melo, "Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of reddit's showerthoughts," *ArXiv*, vol. abs/2405.01660, 2024. 16, 20

[76] J. Ficler and Y. Goldberg, "Controlling linguistic style aspects in neural language generation," *ArXiv*, vol. abs/1707.02633, 2017. 17

[77] S. Prabhumoye, A. Black, and R. Salakhutdinov, "Exploring controllable text generation techniques," in *International Conference on Computational Linguistics*, 2020, pp. 1–14. 17, 22

[78] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, and J. Wei, "A recipe for arbitrary text style transfer with large language models," *ArXiv*, vol. abs/2109.03910, 2021. 17, 22

[79] Y. Lyu, P. Liang, H. Pham, E. Hovy, B. P'oczos, R. Salakhutdinov, and L.-P. Morency, "Styleptb: A compositional benchmark for fine-grained controllable text style transfer," *ArXiv*, vol. abs/2104.05196, 2021. 17, 19, 26

[80] D. Pascual, B. Egressy, C. Meister, R. Cotterell, and R. Wattenhofer, "A plug-and-play method for controlled text generation," *ArXiv*, vol. abs/2109.09707, 2021. 17, 21, 27

[81] S. Mukherjee, A. K. Ojha, and O. Dusek, "Are large language models actually good at text style transfer?" *ArXiv*, vol. abs/2406.05885, 2024. 17, 31

[82] N. Wu, M. Gong, L. Shou, S. Liang, and D. Jiang, "Large language models are diverse role-players for summarization evaluation," in *Natural Language Processing and Chinese Computing*, 2023, pp. 695–707. 18, 33

[83] J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, J. Wieting, N. Peng, and X. Ma, "Evaluating large language models on controlled generation tasks," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3155–3168. 18, 19, 28, 33

[84] Y. Chen, B. Xu, Q. Wang, Y. Liu, and Z. Mao, "Benchmarking large language models on controllable generation under diversified instructions," *ArXiv*, vol. abs/2401.00690, 2024. 18, 32

[85] D. Ashok and B. Póczos, "Controllable text generation in the instruction-tuning era," *ArXiv*, vol. abs/2405.01490, 2024. 18, 19, 33

[86] C. Shen, L. Cheng, Y. You, and L. Bing, "Large language models are not yet human-level evaluators for abstractive summarization," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4215–4233. 18, 34

[87] Z. Zhao, E. Monti, J. Lehmann, and H. Assem, "Enhancing contextual understanding in large language models through contrastive decoding," *ArXiv*, vol. abs/2405.02750, 2024. 19, 23

[88] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. J. Ratner, R. Krishna, J. Shen, and C. Zhang, "Large language model as attributed training data generator: A tale of diversity and bias," *ArXiv*, vol. abs/2306.15895, 2023. 19

[89] K. Yang, D. Liu, W. Lei, B. Yang, M. Xue, B. Chen, and J. Xie, "Tailor: A prompt-based approach to attribute-based controlled text generation," *ArXiv*, vol. abs/2204.13362, 2022. 19, 21, 29, 31

[90] S. Yao, H. Chen, A. W. Hanjie, R. Yang, and K. Narasimhan, "Collie: Systematic construction of constrained text generation tasks," *ArXiv*, vol. abs/2307.08689, 2023. 19

[91] B. Tan, Z. Yang, M. Al-Shedivat, E. Xing, and Z. Hu, "Progressive generation of long text with pretrained language models," in *North American Chapter of the Association for Computational Linguistics*, 2021, pp. 4313–4324. 20, 30

[92] W. Zhao, Y. Huang, X. Han, C. Xiao, Z. Liu, and M. Sun, "Ouroboros: Speculative decoding with large model enhanced drafting," *ArXiv*, vol. abs/2402.13720, 2024. 20

[93] L. Kong, H. Wang, W. Mu, Y. Du, Y. Zhuang, Y. Zhou, Y. Song, R. Zhang, K. Wang, and C. Zhang, "Aligning large language models with representation editing: A control perspective," *ArXiv*, vol. abs/2406.05954, 2024. 21, 27

[94] R. Sinha, Z. Song, and T. Zhou, "A mathematical abstraction for balancing the trade-off between creativity and reality in large language models," *ArXiv*, vol. abs/2306.02295, 2023. 22

[95] D. Wingate, M. Shoeybi, and T. Sorensen, "Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5621–5634. 22

[96] X. Hua and L. Wang, "Pair: Planning and iterative refinement in pre-trained transformers for long text generation," *ArXiv*, vol. abs/2010.02301, 2020. 22

[97] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. Hashimoto, "Diffusion-lm improves controllable text generation," *ArXiv*, vol. abs/2205.14217, 2022. 22, 27, 31

[98] M. Zhou, J. Hou, C. Luo, Y. Wang, Z. Zhang, and J. Peng, "Scenex: Procedural controllable large-scale scene generation via large-language models," *ArXiv*, vol. abs/2403.15698, 2024. 22

[99] H. Zhang and D. Song, "Discup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3392–3406. 24

[100] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *ArXiv*, vol. abs/2006.14799, 2020. 25

[101] X. Liang, H. Wang, Y. Wang, S. Song, J. Yang, S. Niu, J. Hu, D. Liu, S. Yao, F. Xiong, and Z. Li, "Controllable text generation for large language models: A survey," *ArXiv*, vol. abs/2408.12599, 2024. 25, 26

[102] S. Kumar, B. Paria, and Y. Tsvetkov, "Gradient-based constrained sampling from language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2251–2277. 26

[103] J. Baan, N. Daheim, E. Ilia, D. Ulmer, H.-S. Li, R. Fernández, B. Plank, R. Sennrich, C. Zerva, and W. Aziz, "Uncertainty in natural language generation: From theory to applications," *ArXiv*, vol. abs/2307.15703, 2023. 26

[104] X. Liu, M. Khalifa, and L. Wang, "Bolt: Fast energy-based controlled text generation with tunable biases," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 186–200. 26

[105] X. Lu, S. Welleck, L. Jiang, J. Hessel, L. Qin, P. West, P. Ammanabrolu, and Y. Choi, "Quark: Controllable text generation with reinforced unlearning," *ArXiv*, vol. abs/2205.13636, 2022. 28, 32

[106] H. Yan, L. Kong, L. Gui, Y. Chi, E. P. Xing, Y. He, and K. Zhang, "Counterfactual generation with identifiability guarantees," *ArXiv*, vol. abs/2402.15309, 2024. 29

[107] X. Wu, S. Huang, and F. Wei, "Multimodal large language model is a human-aligned annotator for text-to-image generation," *ArXiv*, vol. abs/2404.15100, 2024. 29

[108] S. Kumar, V. Balachandran, L. Njoo, A. Anastasopoulos, and Y. Tsvetkov, "Language generation models can cause harm: So what can we do about it? an actionable survey," in *Conference of the European Chapter of the Association for Computational Linguistics*, 2022, pp. 3291–3313. 30

[109] Y. Liu and Y. Bu, "Adaptive text watermark for large language models," *ArXiv*, vol. abs/2401.13927, 2024. 30

[110] P. Xu, J. Cheung, and Y. Cao, "On variational learning of controllable representations for text without supervision," in *International Conference on Machine Learning*, 2019, pp. 10 534–10 543. 31

[111] J. He, X. Wang, G. Neubig, and T. Berg-Kirkpatrick, "A probabilistic formulation of unsupervised text style transfer," *ArXiv*, vol. abs/2002.03912, 2020. 31

[112] A. Rosenfeld and T. Lazebnik, "Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard," *ArXiv*, vol. abs/2402.14533, 2024. 31

[113] Z. Yu, Z. Wang, Y. Fu, H. Shi, K. Shaikh, and Y. Lin, "Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration," *ArXiv*, vol. abs/2406.15765, 2024. 32

[114] R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi, "Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization," *ArXiv*, vol. abs/2210.01241, 2022. 32

[115] K. Chu, Y.-P. Chen, and H. Nakayama, "A better llm evaluator for text generation: The impact of prompt output sequencing and optimization," *ArXiv*, vol. abs/2406.09972, 2024. 33

[116] T. S. Kim, Y. Lee, J. Shin, Y.-H. Kim, and J. Kim, "Evallm: Interactive evaluation of large language model prompts on user-defined criteria," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. 33

[117] H. Tan, Z. Guo, Z. Shi, L. Xu, Z. Liu, X. Li, Y. Wang, L. Shang, Q. Liu, and L. Song, "Proxyqa: An alternative framework for evaluating long-form text generation with large language models," *ArXiv*, vol. abs/2401.15042, 2024. 33

[118] W. Nie, A. Vahdat, and A. Anandkumar, "Controllable and compositional generation with latent-space energy-based models," in *Neural Information Processing Systems*, 2021, pp. 13 497–13 510. 33

[119] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," *ArXiv*, vol. abs/1810.07217, 2018. 33

[120] E. Crothers, N. Japkowicz, and H. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, vol. 11, pp. 70 977–71 002, 2022. 34

[121] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *ArXiv*, vol. abs/2402.01383, 2024. 34