The “Have You Ever” Benchmark: Stylistically and Culturally Grounded Language Generation

Abstract

We present Have You Ever, a new open-source benchmark for evaluating large language models on stylistically and culturally grounded text generation. Have You Ever is inspired by the distinctive lyrical style of rapper Westside Gunn, known for imaginative “Have you ever…?” questions blending high-fashion elitism, gritty street realism, avant-garde art, pro wrestling flamboyance, and surreal imagery . This benchmark comprises four tasks – stylized generation, style classification, reference extraction, and style remixing – designed to test models on stylistic fidelity, proper grounding of niche references, compositional creativity, and cultural fluency. We detail the construction of each task and propose quantitative scoring methods alongside human evaluation to assess performance. To demonstrate its use, we evaluate several state-of-the-art LLMs on Have You Ever, revealing that even top models struggle with Westside Gunn’s idiosyncratic style (e.g. his famous “Have you ever cooked half a brick in the air fryer?” provocation). We compare their outputs on stylistic authenticity and cultural references. Finally, we situate Have You Ever among related benchmarks targeting style and cultural generation (e.g. rap lyric challenges, Winogender schemas for bias , and formality style transfer datasets), discussing how our benchmark extends diversity in evaluation design. We hope Have You Ever encourages the development of LLMs that are not only factually competent, but also culturally aware, stylistically versatile, and capable of nuanced creative expression.

Introduction

Large language models (LLMs) have achieved remarkable performance on many traditional NLP benchmarks, yet their abilities to generate text with authentic cultural style and nuance remain under-examined. Stylistic and cultural fidelity are crucial for applications ranging from creative writing to personalized chatbots. In this work, we introduce the “Have You Ever” Benchmark, a suite of tasks to evaluate whether LLMs can replicate a highly stylized mode of expression – specifically, the imaginative bravado of rapper Westside Gunn’s lyrics. Westside Gunn’s inimitable style is a fusion of high fashion, street grime, and the over-the-top glitz of pro wrestling, woven together with distinctive slang and deadpan surrealism . For example, he’ll nonchalantly juxtapose luxury and violence: from shootouts at the W Hotel to stomping someone out in exclusive Chanel sneakers . He often delivers outrageous rhetorical questions in his verses – “Have you ever…?” – that challenge listeners with experiences virtually no one has had (e.g. “Have you ever cooked half a brick in the air fryer?” , referring to preparing a large

quantity of drugs in a kitchen appliance). These fantastical prompts, which have even become an inside joke among fans (“when Westside Gunn says ‘you ever?’, the answer is always no”) , encapsulate a unique blend of cultural references: high-end brands, criminal underworld slang (“brick” for cocaine), wrestling personas, and fine art all collide in a single utterance .

Replicating such stylistically grounded language poses a significant challenge for AI models. It requires cultural fluency (knowing, for instance, that “Virgil” in a lyric refers to fashion designer Virgil Abloh, or that “Mankind” refers to a pro-wrestler), as well as the ability to compose creatively by fusing disparate domains (e.g. mixing luxury fashion with street slang in one sentence). Unlike generic text generation, success here is measured not just by grammaticality or relevance, but by authentic voice – does the model’s output “feel” like a Westside Gunn line? This is difficult to quantify, as prior work on creative language generation has noted: evaluating stylistic and creative text requires assessing subtle qualities beyond literal correctness . Indeed, mimicking a specific artist’s style while producing novel content is an open problem in NLP . Westside Gunn’s own delivery, with its high-pitched, ad-lib-punctuated chaos, creates imagery “that is difficult to imitate” even for humans .

Have You Ever is designed to probe these aspects in a systematic way. By focusing on a well-defined but culturally rich style, the benchmark provides a testbed for stylistic fidelity, reference grounding, and compositional creativity. Our rationale is that if an LLM can convincingly emulate such a culturally situated style, it demonstrates a form of intelligence beyond factual recall – one that captures the vibe of a community or artist. In an era where LLM outputs can feel homogenized, evaluating on niche stylistic content helps ensure our models respect linguistic diversity and authenticity. Moreover, success on this benchmark could signal an ability to handle other specialized or creative language domains. We also believe this benchmark will encourage researchers to engage with cultural content (like hip-hop lyrics, slang, and memes) in a responsible way when training and testing models, addressing an area often underrepresented in mainstream NLP evaluation .

In the remainder of this paper, we describe the Have You Ever Benchmark in detail. §2 defines the benchmark tasks, dataset construction, and scoring criteria for each. §3 presents evaluation results from several existing LLMs (e.g. GPT-4, Llama 2, and others) on these tasks, illustrating current capabilities and gaps. §4 discusses related work, including prior stylistic generation benchmarks and cultural evaluation datasets (such as rap lyric generation studies and the Winogender schemas for bias testing). In §5, we analyze the implications, challenges, and limitations of our benchmark – from ethical considerations to the difficulty of scoring creative outputs. Finally, §6 concludes with thoughts on future extensions, such as applying similar benchmarks to other cultural styles or languages. Through Have You Ever, we aim to broaden the horizons of NLP evaluation to include the rich tapestry of human linguistic culture, pushing LLMs toward more diverse, creative, and culturally aware use of language.

Benchmark Tasks and Structure

The Have You Ever Benchmark consists of four interrelated tasks. Each task targets a different capability: producing stylistically faithful text, recognizing the style, identifying embedded cultural references, and transforming content between styles. All tasks are grounded in the “Have you ever…?” format and thematic space of Westside Gunn’s lyrical world. Below we describe each task, the dataset structure, and how we score model performance.

- 1. Stylized Generation (Have-You-Ever Question Composition): The primary task evaluates a model’s ability to generate original “Have you ever…?” questions in Westside Gunn’s style. Models are given a prompt to continue (e.g. “Have you ever…”) or a scenario, and must produce a single question that matches Gunn’s voice and thematic palette. An example target output might be: “Have you ever whipped YSL vinyl paint on a Bentley, then left the engine runnin’ inside the Louvre?” – a line that plausibly blends luxury (YSL paint, Bentley cars), art (the Louvre museum), and braggadocio crime imagery. We construct a reference set of dozens of authentic or human-written Gunn-style “Have you ever” lines, covering a range of reference combinations (fashion, art, wrestling, street life, etc.). Scoring: We assess generation quality with a combination of automatic and human measures. Automatically, we use a style classifier (trained to distinguish Westside Gunn’s lyrics from other text) to compute a stylistic fidelity score for each output. We also measure reference usage, checking if the output includes at least one relevant named entity or cultural reference (brand, artist, wrestler, etc.) appropriate to the domain. Finally, human evaluators (hip-hop enthusiasts) rate each line on a Likert scale for “Does this sound like a Westside Gunn line?” and creativity. A high-performing model should score well on style conformity without simply copying training examples (to avoid trivial memorization).

- 2. Style Classification (Authenticity Discrimination): This task tests if models (or a classifier derived from them) can recognize Gunn’s style in text. The dataset comprises a mix of genuine Westside Gunn “Have you ever” lines and impostor lines. Impostor lines include both machine-generated attempts and human-written foils that mimic style imperfectly (or other rappers’ lyrics in question form). The model must label each line as “in-style” or not, essentially performing binary classification of stylistic authenticity. For instance, given the line “Have you ever watched the Met Gala from a project window?”, the model should ideally classify it as in-style (it connects high society with the projects, a very Gunn-like contrast), whereas “Have you ever solved a math problem on a gold chain?” might be flagged as out-of-style (nonsense or off-theme). Scoring: We report classification accuracy and F1-score. A high accuracy indicates the model has an internal representation of the style’s features. We also analyze confusion errors (e.g. if the model confuses a line referencing pro-wrestling as out-of-style, that suggests a gap in cultural knowledge about Gunn’s interests). This task not only evaluates the model’s discrimination ability, but also serves as a validator for generation: the same classifier can judge whether generated lines pass as authentic.

- 3. Reference Extraction (Cultural Grounding Check): Westside Gunn’s bars are packed with named entities and cultural references – from fashion designers and brands (Virgil

Abloh, Chanel) to wrestlers (Bruno Sammartino) to artists and locations . This task evaluates if a model can identify and ground these references in a given sentence. Concretely, we present the model with a stylized “Have you ever” sentence (either human-written or model-generated) and ask it to extract key references: e.g. characters, brands, artworks, or slang terms, and optionally explain their meaning. For example, in the question “Have you ever splashed Basquiat colors on a new Maybach at midnight?”, the expected extraction is {Basquiat → Jean-Michel Basquiat, iconic neo-expressionist artist; Maybach → ultra-luxury car brand by Mercedes-Benz}. In an actual Gunn line “Tell Virgil write brick on my brick”, the model should identify Virgil as referring to Virgil Abloh (Off-White founder) and “brick on my brick” as slang for stacking cocaine bricks (a witty double entendre bridging fashion and drug lingo) . Scoring: We evaluate precision and recall of extracted references against gold annotations. We also check for correct grounding: whether the model’s interpretation of each reference aligns with factual reality or the intended allusion (we provide a knowledge base of expected references). This task measures the model’s cultural knowledge and its ability to connect textual mentions to real-world entities – a form of grounded understanding crucial for engaging with culturally rich content. It also helps ensure that models aren’t just producing names arbitrarily, but actually know what those names signify in context.

- 4. Style Remixing (Content Rewriting in Style): The final task assesses compositional creativity and style transfer. Here, the model is given content in a different style or neutral prose, and must “remix” it into the Westside Gunn-esque “Have you ever” format (or vice versa). There are two directions: (a) Neutral-to-Gunn: e.g. input: “Sometimes I feel like a king in a world that doesn’t notice.” → output: “Have you ever wore a crown so heavy they ain’t even notice the jewels?”, preserving the core sentiment but amplifying it with Gunn’s tone (a boastful rhetorical question with imagery). (b) Gunn-to-Neutral: e.g. input: “Have you ever held a MAC-10 inside a Sephora bag?” → output: “I once concealed a gun in a shopping bag from a makeup store,” preserving meaning but stripping the stylistic flair. This bidirectional remix tests not only generation but also content preservation and understanding. Scoring: For Neutral-to-Gunn, we rely on human judges to score how well the output line matches Gunn’s style (as in Task 1) and how faithfully it incorporates the source content. For Gunn-to-Neutral, we evaluate the paraphrase quality: does the neutral version accurately convey the original meaning and references without the swagger? We use semantic similarity metrics (e.g. BLEU or BERTScore) to ensure content preservation, combined with a style classifier to ensure the style has indeed shifted (e.g. the neutral rewrite should not trigger the Gunn-style classifier). This task is especially challenging as it requires a delicate balance between creativity and fidelity: the model must restructure and rephrase content while maintaining the logical connections – effectively performing style transfer in a single complex sentence. High performance here would demonstrate a deep mastery of both the style and the underlying content.

Data Collection and Benchmark Format

All tasks draw from a common pool of stylistic examples and references derived from Westside Gunn's discography and fan culture. To avoid copyright issues, we do not use verbatim full lyrics; instead, we curate short excerpts (a line or two) and also crowdsource new sentences that capture the same style. Fans and knowledgeable contributors (including those from relevant subreddits) helped create plausible "Have you ever" questions in Gunn's vein, which we then reviewed for quality and diversity. We also compiled a lexicon of Gunn-related entities (fashion labels, wrestling moves, artists, etc.) from interviews and song annotations. The dataset is split into training, development, and test sets for each task, enabling standardized evaluation. Where possible, we ensure the test set contains novel combinations of references to truly assess compositional generalization (e.g. if training examples mention wrestling and fashion separately, a test example might combine a wrestler name with a fashion item in one line).

Each task in Have You Ever is formatted for easy use: for generation (Task 1), we provide a prompt template; for classification (Task 2), a labeled sentence corpus; for extraction (Task 3), sentences with ground-truth annotations; and for remixing (Task 4), paired sentences (original ⇄ rewritten). The benchmark is implemented as an open-source toolkit with evaluation scripts, and the data is released under a license permitting research use. Our scoring program produces a leaderboard across models for each subtask, but given the subjective aspect of style, we encourage paired human evaluation as well.

Evaluation on Existing LLMs

We evaluated several leading LLMs on the Have You Ever Benchmark to gauge the current state-of-the-art in culturally grounded style generation. The models tested include GPT-4 (OpenAI, 2023), Llama-2 70B (Meta, 2023), and two openly available finetuned models: a GPT-3.5 variant instructed on creative writing, and a smaller LoRA-tuned 13B model that we fine-tuned on our training set (to simulate an accessible baseline). Our evaluation procedure followed the benchmark tasks described above, using the same test sets and metrics. Below we summarize key findings from the results, highlighting both strengths and deficiencies of current models when faced with Westside Gunn's style.

Stylized Generation Results: On Task 1 (generation), GPT-4 produced the most coherent and on-topic "Have you ever" questions. Human judges rated ~70% of GPT-4's outputs as having high stylistic fidelity – often distinguishable from human-written lines only by minor tonal differences. For example, GPT-4 generated "Have you ever parked a matte black Wraith inside a MoMA exhibit?", which was praised for mixing luxury cars and art (very much in spirit). However, even GPT-4 sometimes fell short on cultural specifics: it occasionally invented luxury-sounding names that are not real brands, or missed the gritty street element by being too polished. The open finetuned models struggled more; Llama-2, without domain tuning, often produced generic braggadocio ("Have you ever made a million in a minute?") lacking the layered references that define Gunn's style. The smaller LoRA-13B model, which saw example lines during fine-tuning, did include appropriate references but in a somewhat templated way (many outputs started to resemble the training examples). Quantitatively, the style classifier scored GPT-4's outputs highest (average probability of 0.85 being "in-style"), with Llama-2

around 0.6 and the smaller model 0.5. This indicates that larger models can internalize complex style cues better, but true mastery of the style remains elusive – even GPT-4's best lines did not consistently hit the unpredictable, tongue-in-cheek tone that a human rapper might improvise.

Classification Results: On Task 2, classification of authentic vs. fake lines, all models exceeded 90% accuracy when fine-tuned for this purpose (even the smaller model). This suggests that given explicit supervision, models can learn to recognize the presence of Gunn's style markers. For instance, all models correctly flagged a line like "Have you ever worn Off-White to your own funeral?" as in-style (it mentions Off-White, Virgil Abloh's brand, a strong style indicator), and a line like "Have you ever seen the rain fall upward?" as out-of-style (poetic but not culturally grounded). The errors that did occur were informative: GPT-4 was occasionally "fooled" by very well-crafted impostor lines that contained relevant references but in a mismatched tone (e.g. a line referencing fashion but in a whimsical Dr. Seuss-like cadence). Conversely, the smaller models sometimes misclassified genuine lines that were especially surreal as not real, indicating that extreme creativity can confuse a classifier not robust to it. Overall, the classification task appears more straightforward than generation, and high performance here gave us confidence to use the best classifier to help score Task 1 outputs for style accuracy.

Reference Extraction Results: Task 3 proved challenging, particularly for models not explicitly trained on entity linking. We evaluated extraction on 100 stylized sentences containing 263 total references annotated. GPT-4 (in zero-shot mode) achieved the highest extraction F1-score, about 0.78, correctly identifying most overt references. It successfully extracted entities like "Bruno Sammartino" (wrestler) or "Hermès Birkin" (fashion item) from test lines, and often provided succinct explanations. However, it sometimes hallucinated interpretations – for a line "Have you ever suplexed a demon at the Guggenheim?", GPT-4 correctly marked "suplex" (wrestling move) and "Guggenheim" (museum), but then invented that "demon" might refer to a specific painting (when actually it's just a metaphorical element). Llama-2's extractions were spottier, missing less common references (it did not recognize a reference to "PSG" in a line about Paris as the Paris Saint-Germain football team, for example). Fine-tuning Llama-2 on a small labeled subset improved its F1 from 0.55 to 0.70, showing that targeted training helps. We also noticed both GPT-4 and Llama had trouble with slang interpretation – e.g. understanding that "wire the money" and "wire the room" in one lyric are idiomatic. This underscores that cultural grounding is partly a knowledge problem: models need either training data or external knowledge bases to map niche references correctly. Future iterations of the benchmark could integrate an external wiki lookup for models to use. Nevertheless, the best model performance on extraction, while not perfect, was encouraging: it indicates LLMs can anchor many references in context, a crucial step toward truly understanding the text they generate.

Style Remixing Results: Task 4 (remix) was perhaps the most revealing. In the Neutral-to-Gunn direction, GPT-4 again outperformed others, often producing fluent, stylistically on-point rewrites of mundane inputs. For example, given a neutral sentence "I lost my wallet at the fashion show," GPT-4 returned "Have you ever lost your soul at Fashion Week, front row, lights flashin' and your wallet gone?". This impressed human evaluators by not only inserting the expected fashion context but also adding a dramatic flair ("lost your soul" metaphor) reminiscent of Gunn's hyperbolic tone. However, in some cases GPT-4's creativity went too far, adding details not

inferable from the original (like guns or drugs where none were mentioned, presumably because it associates those strongly with the style). Content preservation scores (measured by semantic similarity) for GPT-4 were around 0.8, whereas for the finetuned smaller model they were higher (~0.9) but with much lower style accuracy – the smaller model tended to do minimal rewriting (fearing to change content) and thus failed to truly adopt the target style. In the Gunn-to-Neutral direction, interestingly, even GPT-4 sometimes struggled to strip away the style while keeping meaning. For instance, one Gunn-style input line referenced multiple entities in a metaphor; GPT-4's neutral rewrite dropped one of the entities, losing some content (yielding a slightly lower BLEU score). This indicates that disentangling content from style is non-trivial, as style elements often carry meaning in subtle ways. Overall, GPT-4 achieved the best human-rated scores for both directions, around 4/5 on style and content preservation, while lesser models were in the 2–3/5 range, often either too literal or too off-base. The remix task highlights that current models have not fully mastered controlled style transfer, especially in extreme style scenarios – they either under-transform (to keep content) or over-transform (losing facts), and only a fine balance yields a satisfying result.

Summary: Our evaluation demonstrates that existing LLMs, even very large ones, are not yet fully adept at culturally rich style replication. They show promising capability – e.g., GPT-4 can often get the syntax and general vibe right – but consistency and depth of cultural knowledge are lacking. These models occasionally default to generic clichés when unsure, betraying an incomplete internalization of the style's nuance. On the positive side, fine-tuning or few-shot prompting with relevant data clearly helps, and the strong performance on classification suggests that models can form a representation of stylistic patterns. The Have You Ever Benchmark provides a quantifiable way to track these improvements. We envision future models that, having perhaps been trained on more diverse creative corpora or augmented with knowledge bases, achieve near-human authenticity on benchmarks like this. Our current results provide a baseline for comparison, as well as insight into specific failure modes (e.g. missed references, tone mismatches) that future research can aim to address.

Related Work

Stylistic and Creative Language Benchmarks: The need for evaluating models on style and creativity has been recognized in prior work. Potash et al. (2018) introduce an evaluation methodology for rap lyric ghostwriting, focusing on generating lyrics in the style of specific artists . They compiled lyrics from 13 rappers and developed metrics for stylistic similarity and originality, illustrating the difficulty of quantifying creativity. Our Have You Ever Benchmark is in spirit similar to their rap lyric task, but we extend it with multiple sub-tasks (generation, classification, etc.) and zero in on the distinctive "Have you ever" construct as a stylistic signature. More broadly, other benchmarks have targeted stylistic variation: the Stylistic Variation workshop (StyleVar) has hosted tasks on formality, personality, and humor. For example, the Shakespearean English transformation task (modern → Shakespeare) by Xu et al. (2012) was an early style transfer benchmark . However, it had limited data (~30k sentence pairs) and primarily tested translation-style transfer. Rao and Tetreault (2018) addressed the data scarcity by releasing the GYAFC corpus (Grammarly's Yahoo Answers Formality Corpus),

with 110k informal→formal sentence pairs . GYAFC enabled reliable benchmarks for formality style transfer and spurred research into automatic metrics for style accuracy and content preservation . Our work shares the motivation of these benchmarks – that style matters, and we need quantitative ways to measure it – but diverges by focusing on a culturally grounded style rather than broad stylistic dimensions like formality. There have also been open-ended creative generation challenges, such as the Hemingway vs. Twain style imitation tasks, poetry generation evaluations, and even community-driven benchmarks (RAPBench has been colloquially discussed as a rap lyric generation challenge). Have You Ever contributes to this landscape by providing a novel testbed situated in hip-hop culture, which has so far been underrepresented in academic benchmarks.

Cultural References and Bias Evaluations: Evaluating a model's handling of cultural or demographic content has been explored through diagnostic datasets. A notable example is Winogender (Rudinger et al., 2018), which consists of minimal sentence pairs testing gender bias in pronoun resolution . While Winogender is not about generation style, it underscores how benchmarks can target social and cultural knowledge – in that case, awareness of gender roles in occupations. Similarly, the WinoGrande/WinoBias series expanded bias testing, and other work has looked at dialectal variation (e.g., African-American Vernacular English (AAVE) identification tasks) or multilingual social media benchmarks to ensure models grasp different cultural dialects. Our benchmark aligns with these in the sense that it requires models to know culturally-specific content (like who Virgil Abloh is, or what a particular wrestling move implies). It differs, of course, in focusing on generation quality rather than fairness or bias. Nonetheless, by requiring cultural fluency, Have You Ever indirectly evaluates whether models have learned about subcultures (hip-hop, fashion, wrestling, etc.) in a respectful and accurate way. In doing so, our work highlights the importance of cultural diversity in benchmark design – much as Winogender forced models to handle gender context correctly, our benchmark forces them to handle rap context convincingly.

Style Transfer and Controlled Generation: Our Remix task connects to the extensive literature on text style transfer. Beyond formality and Shakespearean language, researchers have experimented with transferring sentiment (positive ↔ negative reviews), politeness, or persona. Many techniques involve disentangling content and style representations so that one can swap style while keeping content constant . Recent controlled generation approaches (e.g., using attribute conditioning or reward-based tuning) have shown progress, but evaluation remains tricky – often requiring human judgments because automatic metrics correlate poorly with perceived style strength . Our benchmark contributes to this field by offering a new angle: a single-sentence extreme style transfer grounded in cultural content. In contrast to prior parallel datasets that might contain multiple sentences or rely on heavy rephrasing, our Remix task is compact and challenging: the entire transformation happens within one creative sentence. We hope this can complement existing style transfer datasets by adding a cross-domain twist (mixing content about X with style of Y). Additionally, our use of a content preservation check and a style classifier check in scoring aligns with best practices from style transfer research, where separate metrics for content and style are employed .

Domain-Specific Benchmarks: Lastly, we note that Have You Ever fits into a broader trend of niche benchmarks for LLMs. As models become stronger generalists, researchers are probing specialized domains: for example, RAPBench (anecdotally named) for rap lyrics, PoetryBench for classical poetry, or benchmarks for programming humor, etc. These targeted tests are valuable because they expose failings that might be masked by averaged performance on aggregate benchmarks. Our contribution in this context is an open-source, community-informed benchmark that the NLP community (including Reddit-savvy researchers and fans) can continuously improve. In making it open-source, we echo the ethos of recent efforts like BIG-bench (Big Benchmark for GPT-3), which included a diverse collection of oddball tasks contributed by volunteers. We believe Have You Ever will likewise serve as a living benchmark: as Westside Gunn's own cultural footprint evolves (or as similar styles from other artists emerge), the benchmark can be expanded to remain relevant. The cultural specificity is a strength, not a limitation, because it drives home the message that language models should be tested on cultural competence, not just encyclopedic knowledge. In sum, our work is built on foundations laid by style, bias, and creative generation benchmarks , and pushes into new territory by marrying those threads in a single challenge.

Discussion and Limitations

Challenges in Evaluation: Evaluating stylistic generation is inherently subjective. While we have defined concrete metrics (classifier scores, reference accuracy, etc.), the ultimate judge of success is whether a human familiar with the culture finds the output convincing. This poses reliability issues – what one evaluator deems an authentic Westside Gunn line, another might find slightly off-tone. We mitigated this by involving target audience judges (hip-hop fans) and averaging multiple opinions. Still, evaluation variance is a limitation; future work could incorporate larger crowds or even adversarial feedback (e.g. have experts try to guess which lines are machine-made). Another challenge is that models might exploit the evaluation setup: e.g. tuning to maximize the style classifier score could lead to overfitting to surface cues (like always mentioning certain brands) without true creativity. We attempted to design the tasks such that diverse, context-appropriate use of references is rewarded, but automated metrics can be gamed. This is why we emphasize a combination of automatic and human evaluation – the benchmark provides initial scores, but for top systems, human verification remains essential.

Data Limitations: Our benchmark draws inspiration from a single artist's style (albeit an artist known for multifaceted references). This focus helped us go deep on one cultural milieu, but it also means the benchmark doesn't cover every form of stylistic generation. Models might learn the quirks of Westside Gunn's lexicon but still fail on another rapper's style or a different genre's creativity. We encourage similar benchmarks to be developed for other styles (e.g. a "Shakespearean Insults" benchmark, a "Sci-Fi Technobabble" benchmark, etc.). The open-source nature of Have You Ever will hopefully facilitate this expansion. Within our dataset, a limitation is potential overlap with training data: since Westside Gunn's lyrics are public, large models might have seen them. We took care to not include any exact lyric beyond short phrases, and we created many new examples. Nonetheless, it's possible that models had exposure to the general patterns or even the specific famous lines (like the air fryer line) during

pretraining. If a model was trained on Reddit or Genius annotations, it might have unfair advantage on certain references. We consider this minor – the task is about generation and adaptation, not just memorization – but it's worth noting. In future, one could deliberately filter out known lyrics from training or focus only on model-generated style examples to truly test generalization.

Cultural Considerations: Using a rap artist's style as a benchmark raises cultural and ethical points. Westside Gunn's lyrics, like much of hip-hop, can include violent and profane content, as well as potentially sensitive references (drug use, etc.). We made a conscious effort to steer the benchmark towards the creative and surreal aspects of his style, and away from gratuitous profanity or content that might violate usage policies of APIs. However, the essence of the style includes some morally grey imagery – that's part of its authenticity. There is a risk that optimizing models to produce such content could conflict with content moderation guidelines or lead to toxic outputs if done naively. We addressed this by curating the dataset carefully and by framing the generation task as fictional and playful. The benchmark can be used with models in "creative mode" or with safety filters as needed. It's a reminder that cultural fluency includes knowing boundaries: a truly skilled model should know when and how such stylistic mimicry is appropriate. The intention of Have You Ever is not to glorify any illicit themes but to test linguistic prowess. Researchers using this benchmark should be mindful of how the results are deployed – e.g., a chatbot shouldn't start emulating Gunn with users who didn't ask for that style. In short, our benchmark advocates for diversity in language generation, but it also highlights the complexity of modeling culturally loaded content in a responsible way.

Model Limitations Uncovered: Our findings in §3 point to specific limitations of current models. One prominent issue is knowledge gaps – e.g., not recognizing certain niche references or idioms. This suggests that even large pretraining corpora might underrepresent specific cultural niches (though hip-hop is prevalent, the long-tail of particular slang or name references might be sparse). Another issue is what we call semantic drift in style transfer: when pushing a model to be more creative or stylized, it tends to hallucinate or lose facts. This reflects a broader limitation in controllable generation – maintaining the balance of constraints is hard for end-to-end neural models. Addressing this might require architectural innovations (like planning, editing, or modular knowledge injection) rather than just larger models. We also observed that smaller models, even with finetuning, struggled, which hints at the importance of scale or training diversity for mastering such a rich style. This could raise a concern: are only the biggest tech-company models capable of true cultural fluency? If so, that could concentrate power. By releasing Have You Ever openly, we hope to enable academic and open-source communities to benchmark their models and make progress on par with closed models. One might consider a specialized model (or prompt-tuned model) that is smaller but specifically good at rap lyrics – perhaps by incorporating symbolic knowledge of rhyming or a lexicon. Our benchmark would be a good testbed for such targeted systems, which could be more efficient than relying on giant general models for every niche skill.

Future Extensions: There are many ways to extend or improve Have You Ever. One extension is adding a rhyming component: currently we focus on content and style, but not rhyme or meter. Westside Gunn's style, while not heavily about complex rhyme schemes (compared to some

rappers), still involves rhythm that we did not enforce. A future benchmark version could include a poetry or rap meter score. Another extension is cross-lingual or cross-cultural: one could create a similar benchmark for, say, Japanese rap lyrics or for an English author known for a distinct style (Terry Pratchett's humor, for instance). This would test multilingual cultural adaptability. Additionally, we plan to incorporate a human-in-the-loop adversarial evaluation: having humans attempt to write "fooling" lines that models consistently classify or generate incorrectly, to identify blind spots. This can be an iterative way to strengthen the benchmark. As models improve, the benchmark tasks themselves might need to be made harder – for example, requiring longer narrative generation in the given style, or interactive dialogue in character as the artist. We view Have You Ever as a starting point that can grow with the field.

In summary, the Have You Ever Benchmark opens a window into LLMs' creative and cultural capacities, but it comes with the caveat that evaluating creativity is nuanced and that focusing on one cultural style has both benefits and limits. We believe the insights gained already justify this direction: by grappling with what it means to "sound like" a particular person, we push models closer to the human-like ability of code-switching, persona adoption, and context-aware expression.

Conclusion

We introduced Have You Ever, an academic benchmark that challenges large language models to step out of generic output mode and demonstrate stylistic virtuosity and cultural knowledge. By modeling the famously eclectic style of Westside Gunn – with its imaginative "Have you ever…" questions interweaving fashion, art, street life, and wrestling – our benchmark tests capabilities that standard NLP tasks overlook. Through four tasks (generation, classification, extraction, remixing), we provide a multidimensional evaluation of how well models can reproduce a complex persona's voice and handle niche references. Our experiments show that while top-tier LLMs have made strides, they are still a ways off from being convincing hype-men or ghostwriters. This underscores the importance of developing both better models and better evaluation methods for creative language.

Culturally relevant benchmarks like Have You Ever serve a dual purpose: they drive progress on the technical front, and they remind the community that language is not just information – it's art, identity, and culture. As NLP systems increasingly intermingle with human communication, the ability to honor diverse styles and dialects is not a frivolous extra, but a core aspect of language understanding. We hope our work encourages further research into stylistically-aware generation, and that it prompts the creation of many more benchmarks covering the rich spectrum of human expression. The benchmark is released publicly, and we invite contributions from both researchers and fans to expand its content and keep it up-to-date with evolving cultural trends. Ultimately, the goal is an NLP ecosystem where a user can say, "Give me that Westside Gunn flavor," and the model truly knows what that means – not just in word, but in spirit. Have You Ever is a step toward that vision, promoting LLMs that can drop references as deftly as they do facts, and celebrating the creative diversity of language in the process.