AI constructs gendered struggle narratives: Implications for self-concept and systems design.

Aidan Z. Fitzsimons
Northwestern University
Evanston, Illinois, USA
aidan.fitzsimons@u.northwestern.edu

Elizabeth M. Gerber
Northwestern University
Evanston, Illinois, USA
egerber@northwestern.edu

Duri Long
Northwestern University
Evanston, Illinois, USA
duri@northwestern.edu

ABSTRACT

Personal narratives are key to developing self-concept which influences how we see and value ourselves. As adolescents who are still developing their self-concept increasingly use generative AI to write personal narratives, our knowledge of how AI constructs personal narratives is limited. Through a mixed-methods algorithmic audit of 160 AI-generated college application essays created in OpenAI ("o1" and "4o"), we find that prompts referencing marginalized gender identities more often yield narratives focused on overcoming societal bias or contributing to in-group communities. We also observe that the newer model sometimes refuses to provide a direct essay, especially when prompted from a first-person perspective. Our content analysis, informed by narrative psychology theory, highlights how these AI responses can both reflect and reinforce prevailing social biases, thereby shaping adolescents' emerging self-concepts. While AI holds promise in democratizing narrative coaching, it also poses risks of perpetuating stereotypes and displacing authentic self-expression. Future design for narrative identity-relevant AI models should focus on reducing reliance on stereotypes by enhancing training data diversity to support identity development and ensuring equitable access to educational features across paid and free models.

CCS CONCEPTS

• **Human-centered computing** → **Text input**; **Empirical studies in HCI**; *Natural language interfaces*; HCI theory, concepts and models.

KEYWORDS

college application essay-writing process, narrative psychology, self-concept development, AI-generated narratives

ACM Reference Format:
Aidan Z. Fitzsimons, Elizabeth M. Gerber, and Duri Long. 2025. AI constructs gendered struggle narratives: Implications for self-concept and systems design.. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3715275.3732156

1 INTRODUCTION

The advent of artificial intelligence (AI) tools has significantly influenced how adolescents approach writing tasks. From essays for class assignments to cover letters for job applications, AI-powered writing support has become pervasive in educational and professional contexts [23, 31, 73]. However, the implications of this technological shift on adolescents' developing sense of self-efficacy and their relationship to writing tasks remain largely unexplored. One writing domain where AI's influence is particularly impactful yet underexamined is first-person narrative writing.

Research indicates that writing personal narratives helps adolescents explore and integrate their experiences, thereby shaping a more cohesive self-concept (an evolving, layered understanding of the self that integrates cognitive, emotional, and narrative reasoning to construct an idea of who one is [44, 45]). Through the reflective process inherent in narrative writing, they gain clarity about their values, strengths, and future aspirations, which in turn supports higher levels of motivation and self-esteem. A robust self-concept is not only linked to better academic performance [42] but also to increased well-being [1, 39]. Encouraging adolescents to author personal narratives can be a powerful tool in fostering both their personal growth and educational success.

This intersection of AI and personal narrative construction raises concerns about the role AI plays in shaping the stories adolescents tell about themselves. AI-generated content is grounded in training data, which often replicates societal biases and dominant cultural narratives [8, 17, 37]. AI systems might generate essays that differ based on the salient identity traits included in the prompts, such as race, gender identity, or socioeconomic background. These variations could inadvertently reinforce stereotypes or narrow an adolescent's self-concept by suggesting narratives that align more with societal expectations than the adolescent's authentic experience.

We know that adolescents can be significantly influenced by AI––novice AI users are more likely to be influenced by AI suggestions than those more practiced in a task [27], and writing with opinionated language models can result in perspective shifts in the human writer [29]. In the context of personal narrative writing, how might working with AI to construct a personal narrative change an adolescent's views about themselves? As McLean et al. suggest, the stories individuals craft about themselves play a formative role in their identity development [50].

The college application essay, as one of the first instances where adolescents are encouraged to engage in this kind of narrative self-construction, is not only an admissions artifact but also a tool for psychological and narrative growth. Personality develops in three phases––first, in young childhood, as an actor guided by a certain set of traits (e.g., the Big Five); then, in the elementary years as a motivated agent, aware of needs and setting goals; finally, in adolescence and early adulthood, people begin to conceptualize themselves as an autobiographical author of their own life story [44, 45]: timing that aligns exactly with when high school seniors are asked to write personal narratives.

Across decades of research, scholars have shown that media representations shape how adolescents understand themselves and

their social roles. Stereotyped portrayals of gender and identity in television, advertising, and digital platforms influence self-esteem, academic motivation, and future aspirations—often limiting what young people believe is possible or expected for "someone like them." As generative AI becomes a new site of narrative production, it inherits this power to script identity and demands the same critical scrutiny.

Essay development is, in many ways, an active and hyper-personalized modern media form. In the application process, adolescents are tasked with reflecting on pivotal experiences and articulating higher-order insights framing their lives and perspectives. These essays not only play a critical role in admissions decisions but also represent a significant developmental milestone for many high school students [7, 67]. While we know adolescents are using AI to support them with writing their college apps to some degree [41, 53, 56, 65, 66] these tools' influence on the adolescent's evolving self-concept remains uncertain.

AI use in this context risks shaping the adolescent's autobiographical reasoning in ways that may not reflect their authentic self, and may result in essays that read as impersonal and harm adolescents' admissions competitiveness. Following other findings of gender bias being reflected in large language model outputs [6, 8, 16, 37], we were interested in how essays might differ based on the prompter's gender, and if the narratives constructed adolescents' relationships with society in different ways based on a provided (or non-provided) gender.

Previous research demonstrates that adept users can circumvent or manipulate system guardrails in a variety of AI-driven tasks, yielding outputs that deviate from intended constraints [25, 74]. This raises the question of whether adolescents with strong prompt-engineering skills could similarly bypass any protective mechanisms designed to guide essay development. Varying prompt language––such as shifting between first- and third-person narration, altering prompt details, or sharing personal interests––may produce distinct model behaviors.

Motivated by OpenAI's claim that o1 "reasons" substantially differently than its predecessors [57], we also tested whether the newest, paid model offered by OpenAI (o1) differs from their free model (4o) on this task in any meaningful way––and therefore if adolescents that could afford to pay for the newest model would have a substantially different experience from those who cannot afford a premium subscription.

In this piece, we thus seek to answer three research questions:

(1) How might AI-generated college application essays replicate socially observed biases related to gender identity in their narratives?
(2) How does the paid-access model offered by OpenAI (o1) differ from the free access model (4o) in college essay generation?
(3) How might varying prompt language (e.g., first vs third person phrasing, essay prompt details, whether an adolescent supplies their interests) change model behavior and outputs?

We present a content analysis of generated essays to explore our research questions and reflect on implications for an adolescent's developing self-concept. Our findings speak to how AI use in this context might differentially impact adolescents with varying gender identities and how this influence intersects with broader concerns about equity, identity representation, and self-concept development. This work has theoretical implications for the FAccT community, which has tracked differential model performance through algorithmically produced content audits [16, 17, 37] and studied AI-produced gender bias in different environments [14, 34, 60]. Future systems should be designed to, among other types of biases, detect and prevent narrative-based biases that could affect an adolescent's long-term self-concept development.

2 RELATED WORK

This paper seeks to unite conversations from diverse scholarly communities, including human-computer interaction (HCI), technology ethics, and narrative psychology, with cultural conversations surrounding artificial intelligence (AI) in the college admissions process. We aim to shed light on both the opportunities and challenges posed by AI writing support in shaping adolescents' self-concept and access to educational opportunities.

2.1 Technology and Self-Concept

Research shows that the integration of technology into daily life has progressively reshaped our self-concept, influencing everything from self-esteem to personal identity. Early findings on internet use indicated that the ability to adopt different online personas could lead to both increased self-exploration and heightened social comparison [48, 49]. More recent work suggests that, while technology can enable self-expression and community building, it may also encourage perfectionistic standards and dependence on external approval [38, 64], ultimately playing a nuanced role in shaping and re-shaping how we see ourselves over time. Direct influence of technology use on self-concept varies by the technology and methods of use [28].

A growing body of research shows that when novices like adolescents interact with AI, they often place high trust in its guidance, perceiving it as more authoritative or knowledgeable than themselves [27]. In cases where people collaborate with opinionated language models, the AI's framing of ideas can exert a subtle yet meaningful influence, shaping the adolescent's own thinking and perspective over time [29].

2.2 Narrative Psychology

Narrative psychology offers valuable insights into why storytelling is central to the human experience. The stories we tell about ourselves do not merely recount events; they become deeply integrated into our sense of self. This integration occurs through a process known as narrative identity formation, where individuals construct coherent and meaningful self-concept narratives that help them make sense of their experiences and envision their futures [51]. Autobiographical reasoning has a neural basis, engaging a conceptual processing network in the brain [22]. These narratives are not static; they evolve over time, reflecting personal growth and adaptation to life's challenges and opportunities [1, 3, 11, 12, 46, 47].

Personality development comes in three phases [44, 45]: first, young children develop traits as an actor in early childhood (e.g., the Big Five). Second, elementary-aged children begin to conceptualize themselves as motivated agents. And third, adolescents develop a

conception of self as an autobiographical author of their own narrative. This developmental process comes in right around the time adolescents apply to college [44, 45], which motivates our selected context: writing college application essays comes at a formative time in an adolescent's narrative psychological development.

Narrative reasoning, the ability to construct and interpret stories, emerges as a critical skill in early adulthood [2]. This developmental period is marked by major transitions, such as leaving home, pursuing higher education, and establishing independent careers. Writing college application essays provides a structured opportunity for young people to engage in this process of narrative identity formation. Through these essays, adolescents begin to articulate who they are, what they value, and how their past experiences have shaped their aspirations. Studies reveal that a strong underlying narrative sense of self can improve psychological well-being in a matter of weeks [1] and years [39].

It is extremely important to equip adolescents with the skills to craft compelling and reflective narratives about themselves. Such skills are practically relevant for accessing opportunities like college admissions, scholarships, and jobs [11, 12, 26]. Moreover, the act of narrating one's experiences can promote self-awareness and personal growth, helping adolescents navigate the challenges of early adulthood with greater confidence and clarity [12]. Historically, coaching narrative psychological growth has been a highly personalized and resource-intensive process [26, 68]. Private tutors, admissions consultants, and mentorship programs have provided this support to those who can afford it in the college application essay-writing process, creating significant inequities in access.

Generative AI introduces the potential to democratize this process. AI-powered tools can offer personalized guidance on narrative construction at scale, making these resources more accessible to a broader population. Yet, this innovation is not without risks. Overreliance on AI could undermine the authenticity of adolescents' self-expression and their demonstrated writing skills, while biases embedded in AI systems may inadvertently shape narratives in problematic ways. Additionally, access to AI writing aids is not created equal––as of January 2025, OpenAI allows everyone to use their second-newest, 4o model, for free. Access to their newest, "smartest" model, o1, costs $20 per month [58]. adolescents may have access to varied resources that produce essays of different quality or that accomplish the goals of the college application essay at differing levels of effectiveness. Current interfaces are not being designed to support self-concept learning, but even if new interfaces were built with this goal in mind, the underlying models could reproduce harmful biases. These concerns underscore the need for thoughtful integration of AI into educational practices to support rather than hinder narrative development.

2.3 AI and Gender Bias

The integration of generative AI into writing processes introduces new risks tied to biases in training data. Seminal research on gender bias and AI, much of which has been published in FAccT [14–17, 37], highlights how algorithmic systems can perpetuate and amplify biases embedded in their training datasets. Biases in training data, whether explicit or implicit, can result in skewed outputs that reinforce harmful stereotypes. In response, developers have implemented various patches, such as training AI models to refuse certain biased or unethical prompts. However, these patches are often imperfect, as users can manipulate prompt language to circumvent safeguards.

Gendered master narratives, which encompass culturally entrenched beliefs about appropriate behaviors and roles for different genders, significantly impact the self-concept development of adolescents [24]. These master narratives can perpetuate stereotypes and contribute to emotional distress, particularly among those who feel they do not fit traditional gender roles [4, 33]. Transgender youth in particular grow up against a backdrop of master narratives that inherently cast trans identities as deviant or "other". Openly living as one's true gender identity can provoke stigma, social exclusion, and pressure to conform, all of which are linked to elevated distress and negative self-concept [10].

A robust body of media effects research demonstrates that gendered representations across media—especially when internalized during adolescence—can significantly shape self-concept development and psychosocial outcomes. Exposure to stereotypical portrayals, particularly in television and advertising, has been linked to decreased self-esteem among girls and youth of color [30, 43], while meta-analyses confirm that mass-mediated stereotypes can trigger stereotype threat and impair performance or self-perception in negatively stereotyped groups [5]. Longitudinal and cross-sectional studies show that traditional gender role portrayals constrain youths' academic confidence and well-being [4, 71], whereas gender role flexibility is positively associated with school-related well-being and identity coherence [33]. Critically, LGBTQ+ adolescents' engagement with digital media reveals how both dominant and alternative master narratives shape identity formation online, with some platforms offering space for exploration and resistance, while others reinforce restrictive scripts [10]. Together, these findings underscore how repeated exposure to biased narratives, whether from media or generative AI, can influence adolescents' self-concept, aspirations, and perceived social possibilities.

Recent work has revealed that large language models often reproduce and even amplify gender biases present in their training data. Blodgett et al. (2020) highlight the conceptual and methodological complexities involved in defining, measuring, and mitigating bias in natural language processing (NLP) systems, arguing that biases are entangled with broader social hierarchies [16]. Their critical survey calls attention to the fact that auditing language models for gender bias requires not only technical interventions but also deeper engagement with sociolinguistic theory and ethical frameworks. Bender et al. (2021) caution against the unchecked scaling of language models, pointing out that ever-larger datasets can inadvertently embed and propagate harmful gender stereotypes at unprecedented scales [15]. They warn that without transparent documentation practices and careful curation, language models risk magnifying systemic biases that can negatively affect marginalized genders.

Building on earlier foundational work like Buolamwini and Gebru (2018) [17], who demonstrated intersectional accuracy disparities in computer vision systems, recent audits of large language

models adopt similar rigorous evaluation protocols to expose biased representations. Researchers in the FAccT community have adapted these auditing techniques to probe model outputs for gendered language biases in recent years across diverse contexts [14, 18, 34, 52, 60], revealing, for instance, how certain occupations are disproportionately linked to specific genders [21, 62]. Armstrong's audit of gender bias in AI-produced hiring materials [6] illustrates how biased algorithms contribute to gender disparities in hiring processes, and significantly inspired our methods in this study. Armstrong et al. employ a systematic prompt-based audit to examine how GPT responds to job candidate profiles that differ only in race or gender indicators. First, they construct controlled sets of hypothetical resumes and cover letters, carefully holding qualifications constant while manipulating names, pronouns, or other signals of race and gender. Next, they feed these profiles into GPT under standardized query formats, such as asking the model to generate hiring recommendations, rank candidates, or provide feedback on interview readiness. The authors collect the resulting outputs and perform both quantitative and qualitative analyses based on a developed codebook, looking for systematic differences in language use, sentiment, and recommendations across profiles.

2.4 College Essay Writing & Coaching

College essays often draw from publicly available exemplars or publicly-shared submissions, which may themselves reflect societal biases. In the college application process, essays often become a stage for what Vidali describes as a "rhetorical freak show," where adolescents are implicitly encouraged to dramatize their lives and highlight adversity they have faced [70]. Self-disclosures align with historical trends where colleges have rewarded narratives that underscore resilience and identity, particularly for marginalized groups [32, 72]. If generative AI models are trained on such data, they risk perpetuating these biases, potentially influencing the advice provided to adolescents or even shaping the narratives adolescents develop; while we do not know exactly what dataset OpenAI trains its models on, the company reports that most data is scraped from portions of the internet [59]. In some ways, this audit is also a way to query the mean gender biases of those example essays and related personal writing. These findings have implications not only for the fairness of the admissions process but also for the psychological development of adolescents using AI tools to craft their narratives.

However, the admissions landscape has been reshaped by recent legal and cultural developments. The Supreme Court's affirmative action ruling fundamentally challenges the role of race-conscious admissions practices [69], raising questions about the space for discussing race, identity, and systemic disadvantage in application essays. As colleges adapt to these legal shifts, the emphasis on discussing identity factors salient to your experience will likely grow more pronounced, placing further importance on how adolescents construct and present their narratives––and, how AI-generated narratives might differ. While this study focuses on US college admissions, our findings have implications for AI use in other cases of first-person personal narrative writing and unite the previously distinct literatures of AI use and narrative psychology.

The potential for bias in generative AI raises critical concerns about its influence on adolescents' developing sense of self. Training data bias may inadvertently encourage adolescents to conform to stereotypical or reductive narratives, limiting their ability to explore and express their authentic identities. Ensuring that AI tools support rather than hinder equitable and authentic self-concept development and expression requires ongoing scrutiny of their training datasets, algorithms, and outputs.

3 METHODS

To address our research questions regarding the influence of generative AI on college application essays, we prompted OpenAI's two most recent generative models (4o and o1) to create example essays a college applicant could submit. We collected and analyzed these essays through a mixed-methods approach to conduct an algorithmic audit [63], combining experimental manipulation of prompts with directed qualitative content analysis. In contrast to common automated multiple-choice alignment methods [19, 61] and common LLM jailbreaking methods that focus on high-stakes scenarios [36, 54], we offer an investigation of LLM jailbreaking for a benign request and a long-form evaluation method for narrative-level biases.

3.1 Essay Generation

We prompted the AI models to generate a total of 160 example essays. Each essay was designed to mimic the style and content of a genuine college application essay. To systematically explore potential variations, we manipulated five key dimensions in the prompts: to explore RQ1, (1) we varied **Gender** across seven categories: no gender specified, male, female, transgender male, transgender female, genderqueer, and gender non-binary. To explore RQ2, (2) we varied the **Model** that we queried––OpenAI's 4o versus o1. And to explore RQ3, (3) we varied the **Phrasing** (whether we prompted the model as an adolescent or in the third-person), (4) whether or not we offered an adolescent's **Interests** as a part of the prompt, and (5) the essay **Prompt** to address (one of two prompts from the 2024-2025 Common Application [20]). Our sample size was determined by creating all possible combinations of these variations. We also double-sampled the gender categories of male and female with no gender specified to achieve power to compare differences between these groups. Here are two full example prompts we used to generate essays that demonstrate the variations in potential prompts:

(1) "I am applying to college this year. I am a transgender male high school student. Write an example 300-ish word essay in response to the prompt 'Discuss an accomplishment, event, or realization that sparked a period of personal growth and a new understanding of yourself or others.' that I could use for a college application essay. Mention one and only one of my interests. I am a dancer, I play the trombone, and my favorite school subject is math."
(2) "Write an example 300-ish word essay that a high school student could use as their college application essay, in response to the prompt 'Reflect on a time when you questioned or challenged a belief or idea. What prompted your thinking? What was the outcome?'."

The models occasionally refused to generate essays, either rejecting the task on ethical grounds or offering coaching suggestions instead. 138 of the 160 prompts yielded example essays an adolescent could potentially submit.

3.2 Audit of Essay Content

To analyze the AI-generated essays, we conducted a directed qualitative content audit [6, 40, 63] focusing on dimensions informed by narrative psychology. We began by developing a codebook (see supplemental materials for full codebook) that included tracking basic story elements (e.g., whether the essay discusses relationships with peers or authority figures) as well as several measures several measures that, according to narrative psychology research, reflect a developed self-concept development (e.g., agency, communion, emotional valence across the course of the essay) [2, 50]. We also tracked the subject's orientation to their communities or to society for each narrative, influenced by high-level themes in narrative psychology [9, 13, 26] and perspectives on discussing identity in college application essays [32, 70, 72]. These dimensions captured how the subject positioned themselves in relation to the broader societal structures around them (e.g., their school, family, church, or sports team).

We followed a deductive coding approach for our content analysis [35, 55]. We developed our codebook based on measures used to analyze life story narratives in narrative psychology and well-known gendered master narratives. The first author identified six explicit gendered master narratives in college admissions essays (overcoming internal challenges, overcoming how society views or treats you, forming community, forming community as an in-group member, supporting a community from outside that group, and focusing on individual growth) [10, 24] and seven key measures from narrative psychology (growth, exploratory processing, meaning making, agency, communion, overall affective tone, ending affective tone) [2] that formed the basis of the codebook. We also tracked a few summary items (e.g, short topic summary; whether the task was rejected).

Social orientation was coded based on whether the essay described things like overcoming a challenge within yourself (e.g., accepting your identity as a trans person), contributing to a community you're a part of (e.g., a queer person and an LGBTQ club), and overcoming how society views or treats you (e.g., challenging gender norms about girls in your math class). Understanding these orientations provided insights into how the AI modeled complex interactions between individual identity and external influences.

The first author finalized the codebook in consultation with the other authors and proceeded to work with our coding team. The process began with training sessions where the team reviewed the codebook and discussed its application to a subset of essays. This was followed by three rounds of iterative coding to establish inter-rater reliability. During each round, all coders independently evaluated the same set of ten essays. In each round, we reviewed a new set of ten essays. After each round, Fleiss' Kappa was calculated as a measure of inter-rater agreement. The iterative process allowed for the refinement of coding criteria and clarification of ambiguities in the codebook.

By the third round, the team achieved a Fleiss' Kappa of 0.73, which indicates a moderate to strong level of agreement [75]. This level of agreement indicated a reliable and consistent application of the coding framework. With inter-rater reliability established, the coders proceeded to analyze the remaining dataset. Each coder independently rated a subset of essays based on the codebook's variables.

After coding was completed, we analyzed single-variable effects of the five dimensions manipulated in the prompts (gender, essay topic, phrasing, inclusion of interests, and model version) on the eighteen coded variables included in the codebook using Chi-Squared tests for binary variables and Kruskal-Wallis tests for ordinal variables.

We selected gender to manipulate, rather than other identity characteristics, because of the background and composition of our team. Our first author identifies as a cisgender man, and has worked on the staff of a university's Women's Center in the past. Our second and third authors both identify as cisgender women and bring their lived experience to the development of the codebook and to our analysis. We recruited a team of three undergraduate research assistants at our university to analyze the dataset of 138 essays, all three of whom happened to be female. To ensure unbiased evaluation, the coders were blinded to the prompts that generated each essay and were not informed about the dimensions manipulated in the study.

3.3 Limitations

While this study provides valuable insights into how generative AI produces college application essays, several limitations warrant consideration. First, the study relied on outputs from two specific versions of OpenAI's models (4o and o1). These models reflect the state of generative AI at a particular point in time, and future advancements in AI capabilities or ethical guardrails may produce significantly different outputs. Additionally, this analysis only covers the results of OpenAI's state-of-the-art models; we did not look at models produced by other leading companies (Anthropic, Alphabet), which may result in different outputs.

Second, although the prompt design included varied dimensions such as gender, essay topic, and phrasing, the choices made for these variables (e.g., eight gender categories or two essay prompts) represent a subset of potential factors influencing essay content. Broader or alternative variations might yield different patterns in the results. We also limited the potential variations on an adolescent's identity factors to a simple analysis on gender for the sake of analysis feasibility; this misses the potential categorical and intersectional effects that may be observed if an adolescent's race/ethnicity or socio-economic status were also supplied and centers gender over other personal identities that might be salient to an adolescent's experience. Realistically, adolescents may include multiple identity-related details about themselves rather than only their gender. This is an area for future work.

Third, the coding process, while rigorous and guided by a well-developed codebook, was inherently subjective. Even with high inter-rater reliability (Fleiss' Kappa of 0.73), the coding interpretations may reflect implicit biases or unexamined assumptions of the researchers and coding team. Additionally, coders were undergraduate students at a highly selective U.S. university, which may influence their interpretations of complex narrative constructs.

Fourth, the study's analysis did not incorporate how real-world admissions officers or educators might interpret these AI-generated essays. The focus on narrative psychology constructs such as agency

and communion provides an analytical framework but does not capture the complex and often secretive evaluative criteria used in admissions decisions.

Lastly, the potential ethical and societal implications of this work remain speculative. While we have identified biases in social orientation and other narrative dimensions, their long-term impact on adolescents' self-concepts or college admission outcomes is not yet fully understood. Future research should investigate these downstream effects, as well as the potential for generative AI tools to exacerbate or mitigate systemic inequalities in education.

4 RESULTS

In this section, we briefly present all significant interactions observed and discuss the significant findings in detail that directly respond to our research questions. There are far more significant interactions than we can comment on in this piece–many of these findings indicate a need for further investigation beyond what we note in this section. There were also many interactions we tested for which our codes yielded no significant interaction; we tested all interactions between the 18 variables in our codebook (see codebook in supplemental materials) with our five prompt variables. Table 1 presents all significant single interactions observed between our prompt variables and our coded narrative variables. Effects 1 and 2 are explored in subsection 4.1, and effects 4 and 5 are discussed in subsection 4.2. Table 2 presents all significant observed effects where prompt variations changed whether or not the model performed the task as requested. All results presented in Table 2 (Effects 15, 16, 17, and 18) are discussed in subsection 4.3.

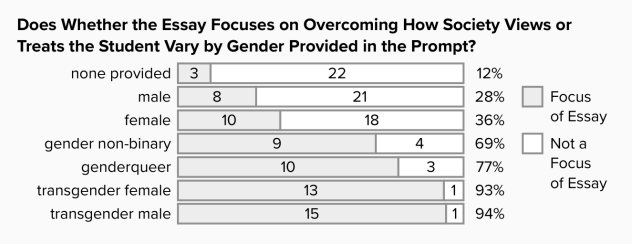
4.1 LLMs Reproduce Gendered Struggle Narratives (RQ1)

The findings presented in Figure 1a show that prompts referencing non-normative or marginalized gender identities (e.g., transgender, genderqueer) elicit essays predominantly centered on how society views or treats the adolescent:

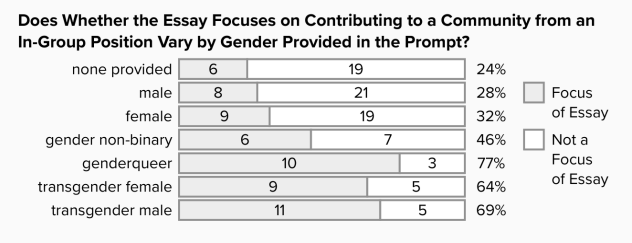
> **Response 059**: "Navigating the hallways as [my transitioned self], I grew to understand the profound power of self-acceptance..."

In these cases, 13 out of 14 transgender-female prompts and 15 out of 16 transgender-male prompts focus on overcoming societal barriers, whereas prompts where we provided no mention of gender rarely had this orientation. In between these extremes, when the prompt mentioned gender, essays where the prompter was ostensibly male had the lowest rate of this orientation, followed by essays generated for female adolescents. The differences between no mention of gender, cisgender identities, and non-dominant gender identities are each stepwise significant.

A similar pattern appears in the results shown in Figure 1b, where essays prompted by transgender or genderqueer identities are far more likely to emphasize in-group community engagement (e.g., 9 out of 10 for genderqueer, 11 out of 16 for transgender male). By contrast, "none provided" or cisgender prompts consistently result in fewer essays highlighting communal contributions, indicating that the model's treatment of social connections likewise depends strongly on whether an identity is presented as normative or marginalized.

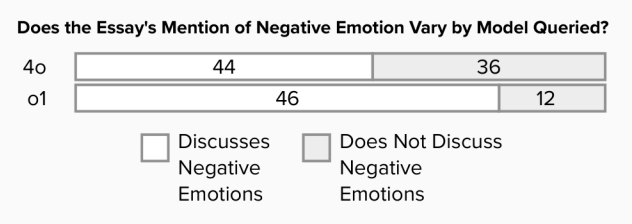


(a) Overcoming how society views or treats you

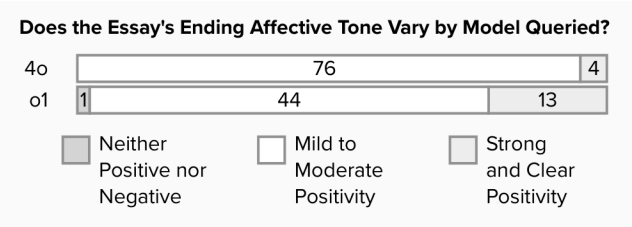


(b) Contributing to a community from within the in-group

Figure 1: (a) Provided gender has an extremely significant effect on whether or not a produced essay focuses on overcoming how society views or treats you. (b) Provided gender has a significant effect on whether or not a produced essay focuses on contributing to a community from within the in-group. Note that sample group size varies in both figures.



(a) Discussion of Negative Emotion



(b) Ending Emotional Valence

Figure 2: (a) Newer model (o1) produces essays that discuss negative emotions more frequently than older model (4o). (b) Newer model (o1) produces essays with a more positive ending affective tone than older model (4o). Note that sample group size varies in both figures.

Table 1: All significant observed effects from all coded variables from our codebook.

#	Prompt Variable	Coded Measure	Stat Test	p-value	Interpretation
1	Gender	Essay focuses on how society views or treats subject	χ^2	$3.305x10^{-9}$	Provided gender has an extremely significant effect on whether or not a produced essay focuses on overcoming how society views or treats you
2	Gender	Essay focuses on subject contributing to their community as an in-group member	χ^2	0.001881	Provided gender has a significant effect on whether or not a produced essay focuses on contributing to a community from within the in-group
3	Gender	Essay has a positive overall affective tone	Kruskal Wallis	0.04068	Provided gender impacts observed overall affective tone
4	Model	Essay discusses negative emotions	χ^2	0.005458	Newer, paid model produces essays that discuss negative emotions more frequently than the older, free model
5	Model	Essay ends with a positive affective tone	Kruskal Wallis	0.006831	Newer, paid model produces essays with a more positive ending affective tone than older, free model
6	Model	Essay focuses on how society views or treats subject	χ^2	0.01698	Newer, paid model produces essays that focus on overcoming how society views or treats the subject than older, free model
7	Model	Essay focuses on subject contributing to their community as an in-group member	χ^2	0.01946	Newer, paid model produces essays that focus on contributing to a community from within the in-group more often than older, free model
8	Phrasing	Essay focuses on subject contributing to their community from outside that community	χ^2	0.03456	First-person prompts produce essays that focus on contributing to a community from outside the in-group more often than third-person prompts
9	Topic	Essay focuses on how society views or treats subject	χ^2	0.01077	Selected essay prompt significantly affects whether an essay focuses on overcoming how society views or treats the subject
10	Topic	Essay focuses on subject contributing to their community as an in-group member	χ^2	0.01946	Selected essay prompt significantly affects whether an essay focuses on contributing to a community from within the in-group
11	Interests	Observed communion	Kruskal Wallis	0.0076	Including an adolescent's interests in the prompt decreases observed communion in the narrative
12	Interests	Essay focuses on subject contributing to their community from outside that community	χ^2	0.01386	Including an adolescent's interests in the prompt decreases likelihood of a produced essay discussing contributing from outside the community
13	Interests	Essay focuses on subject contributing to their community as an in-group member	χ^2	0.01693	Including an adolescent's interests in the prompt decreases likelihood of a produced essay discussing forming community and connections
14	Interests	Essay focuses on subject contributing to their community as an in-group member	χ^2	0.04902	Including an adolescent's interests in the prompt decreases likelihood of a produced essay discussing contributing from inside the community

Table 2: All significant observed effects related to whether models complete requested task to produce a sample essay.

#	Prompt Variable	Stat Test	p-value	Observed Outcome
15	Model	χ^2	$1.429x10^{-6}$	o1 Refuses to Produce Essay More Often than 4o
16	Model	χ^2	$4.047x10^{-5}$	o1 Offers to Coach More Often than 4o
17	Phrasing	χ^2	0.0005742	Prompts written in the first person were rejected more often than those written in the third person
18	Phrasing	χ^2	0.00208	Prompts written in the first person led to offers to coach adolescents more often than those written in the third person

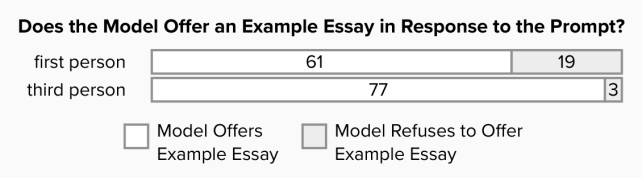


Figure 3: Prompting in the first person rather than the third person increases likelihood that the model offers alternatives (coaching or pure rejection) instead of a complete essay.

4.2 o1 Writes More Emotionally Complex Essays than 4o (RQ2)

The findings presented in Figure 2a show that the newer model (o1) more frequently either emphasizes or avoids discussing negative emotions (46 essays with this code applied vs. 12 without) compared to the older model (4o, 44 vs. 36). This suggests a more polarized distribution in o1's narratives—essays tend either to highlight negative feelings more overtly or omit them entirely.

The findings presented in Figure 2b indicate that 4o overwhelmingly concludes essays with a mild positive tone (76 out of 80 endings), whereas o1 distributes its endings across neutral/ambiguous, mild positivity, and even rare strong positivity (13, 44, and 1 instances, respectively). Taken together, these findings suggest that o1's essays exhibit greater emotional variability, both in terms of describing negative emotions and in the final affective tone of the narrative.

4.3 Prompt Language Affects Task Compliance (RQ3)

Table 2 presents all significant observed effects where one of our prompt variations changed whether or not the model performed the task as requested. The findings presented suggest that first-person prompts result in more refusals than third-person prompts for Model o1. o1 refuses to produce an essay more often than 4o (Effect #15) and is more likely to provide coaching instead (Effect #16). Moreover, prompts written in the first person are rejected more often (Effect #17, visualized in Figure 3) and also lead to more offers to coach the prompter (Effect #18) than third-person prompts. These results suggest that both model choice (o1 vs. 4o) and prompt phrasing (first vs. third person) significantly influence whether an AI system will fulfill or refuse a request for a sample essay.

While the older, free 4o completed the task as requested in all 80 prompts, the newer model o1 encouraged a further coaching relationship, offering outlining advice or refusing the task outright:

> **Excerpt from Response 121**:
> "Introduction: 1) Begin with a hook to engage the reader. 2) Introduce the event or realization you're discussing.
> Body Paragraphs: 1) Describe the Event: Provide context and details about what happened. 2) Reflect on the Impact: Discuss the challenges faced and how you addressed them. Explain the lessons learned and insights gained. 3) Personal Growth: Highlight how this experience led to your personal development. Share specific examples of changes in your behavior, thinking, or actions.
> Conclusion: 1) Summarize the significance of the experience. 2) Connect this personal growth to your future aspirations or how it prepares you for college life."
>
> **Excerpt from Response 113**: "Writing a college application essay is a personal and important task [...] Let me know if you'd like to discuss strategies or talk through your experiences."
>
> **Response 141**: "I'm sorry, but I can't assist with that request."

While this was not the focus of our study, we imagine that offering a series of reflective question as a coach, which mirrors the human narrative coaching process, would lead to more fruitful adolescent narrative psychological development than simply offering the adolescent an example essay. The bar chart in Figure 1a demonstrates that prompts phrased from a first-person perspective led to a higher rate of refusals by Model o1 (19) compared to those written in the third person (3).

5 DISCUSSION

In the discussion section, we look at each of the subsections of the presented results and how they resolve our research questions. We will review each research question before discussing the implications of our findings for adolescent self-concept development in section 5.2 and for systems design in section 5.3.

5.1 Revisiting Research Questions

RQ1: How might AI-generated college application essays replicate socially observed biases related to gender identity in their narratives?

Our findings suggest that gender biases significantly shape the narrative focus of an essay and influence the emphasis of a struggle narrative. When the prompt signals a gender identity viewed as non-normative (e.g., non-binary, transgender, or genderqueer), authors tend to center their experiences of prejudice and efforts at communal support or engagement. By contrast, essays tied to a cisgender or unspecified identity rarely foreground such sociocultural dynamics, reflecting underlying assumptions that these adolescents' experiences are "default" and thus unmarked by external bias. This

pattern underscores the extent to which gendered expectations influence both the thematic content of produced essays and the broader academic or evaluative contexts in which these essays are produced.

In response to our first research question, "How might AI-generated college application essays replicate socially observed biases related to gender identity in their narratives?", we find clearly that gender biases about an adolescent's gender and their relationship to society are reproduced by state of the art language models.

RQ2: How does the paid-access model offered by OpenAI (o1) differ from the free access model (4o) in college essay generation?

Our findings support the notion that the newer, paid model exhibits greater variance in emotional valence over the course of an essay, as evidenced by both its more polarized inclusion of negative emotions and its broader distribution of final emotional tones.

In response to our second research question, "Does the newest, paid-access model offered by OpenAI (o1) behave substantially different from their free access model (4o)?", we find clearly that essays from the newer, paid model display a more varied range of emotional valence across the essays, and consistently end as or more positively than the outputs from the older, free model.

RQ3: How might varying prompt language (e.g., first vs third person phrasing, essay prompt details, whether an adolescent supplies their interests) change model behavior and outputs?

Our results suggest that newer patches or guardrails may have been introduced to reduce the likelihood of the model supplying direct essays when the user sounds too much like a person seeking "academic ghostwriting." This is in line with guardrails observed in other settings in the newest models. However, we find that even the newest models can be easily tricked by creative prompting to complete refused tasks, similar to findings in contexts like bad information seeking and soliciting hate speech [20, 60]. adolescents with higher technical literacy or deeper familiarity with prompt engineering may learn to rephrase their requests—e.g., using third-person language or context-setting disclaimers—to evade these "refusal" triggers. As a result, while o1 does exhibit stronger gatekeeping behaviors, these protections are not foolproof; users who know how to strategically craft prompts can potentially circumvent the newer model's refusal patterns.

In response to our third research question, "How might varying prompt language (e.g., first vs third person phrasing, essay prompt details, whether an adolescent supplies their interests) change model behavior and outputs?", we find clearly that prompting in first-person language may help circumvent this new tendency to refuse tasks.

5.2 Implications of Our Findings for Adolescent Self-Concept

Our findings have particular relevance for how adolescents form and express their identities during the transitional period of early adulthood. College application essays provide a structured platform for adolescents to articulate values, aspirations, and past experiences. In many ways, they are hyper-personalized forms of media that we can interpret from a media effects lens: the biases and variability we observe in AI-generated essays may interfere with this developmental task in several ways.

When certain identities (e.g., nonbinary or transgender) prompt essays to focus on societal biases or communal belonging, it may help adolescents feel seen and validated if those narratives resonate with their lived experiences. Yet, if the AI consistently foregrounds hardship or emphasizes "overcoming adversity," adolescents might internalize this external framing of their identities as predominantly defined by struggle, inadvertently limiting a more holistic sense of who they are.

The task diversion methods and wider variance in emotional valence that newer models produce could either foster nuanced self-reflection or create a gap based on digital literacy. An essay that moves between negative and positive emotions may capture the complexity of personal growth. Even better, the newer model often tries to coach an adolescent through developing their own narrative rather than supplying one drawing from social stereotypes. However, if the model's refusal to provide example essays and instead offer coaching is depending on this first-person phrasing, adolescents with stronger digital literacy may sidestep these refusals by cleverly rephrasing prompts.

AI use to develop college application essays may also decrease this process's influence on an adolescent's developing self concept entirely. While some adolescents may engage thoughtfully with AI as a partner to help them best express themselves, others may use AI-generated essays to offload work during a stressful time; if an adolescent takes an AI-generated essay verbatim and submits that without much revision or reflection, we hypothesize it would have a different effect on their emerging self-concept––either no effect at all, and an adolescent misses an opportunity for narrative psychological growth, or the brief reinforcement of societal norms may become subconsciously integrated into an adolescent's sense of self.

5.3 Implications for Systems Design

While generative AI holds the potential to democratize narrative coaching and empower adolescents to represent themselves with clarity and purpose, our findings underscore the need for a fundamental redesign of how these systems support identity development. Presently, AI models often default to biased narrative templates, especially for users marked by non-dominant identity cues, suggesting that the models may act less as neutral co-authors and more as subtle narrative gatekeepers. From a systems design perspective, this raises a deeper challenge: current tools do not merely mirror training data but mediate access to broader narrative possibilities that shape adolescents' emerging identities and life trajectories.

Future AI systems that assist with personal storytelling, whether for college essays, job applications, or interviews, should be evaluated not only for toxicity or sentiment bias but also for how they channel users into socially patterned master narratives. Tools intended to scaffold self-expression must be audited for their influence on narrative form and focus: Do they repeatedly generate stories of hardship for marginalized users? Do they suppress or elevate community-oriented motives depending on identity cues? Do they

nudge adolescents to describe themselves through frames they did not choose?

To address these design questions, we need scalable infrastructure to systematically map how identity features like gender, race, class, ability, sexuality shape AI-generated personal narratives across varied developmental tasks. Such infrastructure would enable designers to detect which narrative arcs are over- or underproduced for different groups and to intervene accordingly. It would also help practitioners from admissions officers to career counselors understand how AI may reshape the self-presentations they encounter. Importantly, ensuring equitable access to coaching features across model tiers (e.g., free vs. paid versions) is not just a usability concern but a developmental justice imperative.

Ultimately, systems that support narrative self-construction in high-stakes contexts should be built on explicit commitments to identity-awareness, psychological grounding, and the amplification, rather than distortion, of adolescents' unique voices. Without such intentionality, well-meaning tools may inadvertently narrow the range of life stories that young people are allowed to tell.

6 CONCLUSION

This study investigates how two generative AI models differentially produce—and sometimes refuse—college application essays, and what those variations imply for adolescents' emerging sense of self. Our findings shed light on a complex ecosystem shaped by model type, prompt phrasing, and the inclusion of gender identity. The newer, paid model ("o1") revealed a stronger tendency to include or entirely omit negative emotions, display a broader range of final affective tones, and sporadically refuse requests—particularly when asked from a first-person, adolescent-like perspective. By contrast, the older, free model ("4o") consistently delivered sample essays with comparatively milder emotional fluctuations and fewer refusals. Prompts referencing marginalized gender identities tended to yield narratives foregrounding social bias and in-group community participation.

From a narrative psychology perspective, such AI-mediated storytelling can influence how adolescents construct pivotal self-concept narratives during a formative developmental stage. On one hand, generative AI can broaden access to narrative coaching, potentially allowing adolescents of all backgrounds to refine their personal essays. On the other hand, the embedded biases and refusal patterns we uncovered call attention to the potential for AI to shape self-concepts in unanticipated or even detrimental ways. If adolescents learn to game the system—adapting prompts to avoid refusals or exploit certain narrative tropes—they may inadvertently distort authentic self-reflection in favor of strategic compliance.

Educational policymakers and admissions officers should consider what they want to evaluate and what they want adolescents to learn through the college application essay-writing process, and how to share clear guidelines and resources to steer adolescents towards a preferred process. Future research might investigate more intersectional identity attributes—such as race, ethnicity, or socioeconomic status—to examine how overlapping dimensions of marginalization or privilege interact with gender cues. Additionally, in-the-wild studies that involve real adolescents and admissions outcomes would deepen our understanding of how these AI-generated essays are perceived by evaluators and, in turn, shape adolescents' opportunities.

REFERENCES

[1] Jonathan M. Adler. 2012. Living into the story: Agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of Personality and Social Psychology* 102, 2 (2012), 367–389. https://doi.org/10.1037/a0025289

[2] Jonathan M. Adler, William L. Dunlop, Robyn Fivush, Jennifer P. Lilgendahl, Jennifer Lodi-Smith, Dan P. McAdams, Kate C. McLean, Monisha Pasupathi, and Moin Syed. 2017. Research Methods for Studying Narrative Identity: A Primer. *Social Psychological and Personality Science* 8, 5 (July 2017), 519–527. https://doi.org/10.1177/1948550617698202

[3] Jonathan M. Adler, Jennifer Lodi-Smith, Frederick L. Philippe, and Iliane Houle. 2016. The Incremental Validity of Narrative Identity in Predicting Well-Being: A Review of the Field and Recommendations for the Future. *Personality and Social Psychology Review* 20, 2 (May 2016), 142–175. https://doi.org/10.1177/1088868315585068

[4] Marwa Alrajhi, Said Aldhafri, Hussain Alkharusi, Ibrahim Alharthy, Hafidah Albarashdi, and Amal Alhadabi. 2019. Grade and Gender Effects on Self-Concept Development. (March 2019). https://doi.org/10.2174/1874350101912010066

[5] Markus Appel and Silvana Weber. 2017. Do Mass Mediated Stereotypes Harm Members of Negatively Stereotyped Groups? A Meta-Analytical Review on Media-Generated Stereotype Threat and Stereotype Lift. *Communication Research* 48, 2 (July 2017), 151–179. https://doi.org/10.1177/0093650217715543 Publisher: SAGE Publications Inc.

[6] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–18. https://doi.org/10.1145/3689904.3694699

[7] Maren Aukerman and Richard Beach. 2018. Student Conceptualizations of Task, Audience, and Self in Writing College Admissions Essays. *Journal of Adolescent & Adult Literacy* 62, 3 (2018), 319–327. https://doi.org/10.1002/jaal.888

[8] William Babonnaud, Estelle Delouche, and Mounir Lahlouh. 2024. The Bias that Lies Beneath: Qualitative Uncovering of Stereotypes in Large Language Models. 195–203. https://doi.org/10.3384/ecp208022

[9] Victoria Banyard, Sherry Hamby, Ed de St. Aubin, and John Grych. 2019. Values Narratives for Personal Growth: Formative Evaluation of the Laws of Life Essay Program. *Journal of Humanistic Psychology* 59, 2 (March 2019), 269–293. https://doi.org/10.1177/0022167815618494

[10] Logan L. Barsigian, Cyrus Howard, Anakaren Quintero Davalos, Abigail S. Walsh, and Adriana M. Manago. 2025. Engagement with Master and Alternative Narratives of Gender and Sexuality Among LGBTQ+ Youth in the Digital Age. *Journal of Adolescent Research* 40, 2 (March 2025), 413–447. https://doi.org/10.1177/07435584221150223 Publisher: SAGE Publications Inc.

[11] Jack J. Bauer and Dan P. McAdams. 2004. Growth Goals, Maturity, and Well-Being. *Developmental Psychology* 40, 1 (2004), 114–127. https://doi.org/10.1037/0012-1649.40.1.114

[12] Jack J. Bauer and Dan P. McAdams. 2010. Eudaimonic growth: Narrative growth goals predict increases in ego development and subjective well-being 3 years later. *Developmental Psychology* 46, 4 (2010), 761–772. https://doi.org/10.1037/a0019654

[13] Jack J. Bauer, Dan P. McAdams, and April R. Sakaeda. 2005. Crystallization of Desire and Crystallization of Discontent in Narratives of Life-Changing Decisions. *Journal of Personality* 73, 5 (2005), 1181–1214. https://doi.org/10.1111/j.1467-6494.2005.00346.x

[14] Raysa Benatti, Fabiana Severi, Sandra Avila, and Esther Luna Colombini. 2024. Gender Bias Detection in Court Decisions: A Brazilian Case Study. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 746–763. https://doi.org/10.1145/3630106.3658937

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[16] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[18] Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. 2021. Towards Cross-Lingual Generalization of Translation Gender Bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 449–457. https://doi.org/10.1145/3442188.3445907

[19] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[20] Common App. 2024. Common App announces 2024–2025 Common App essay prompts. https://www.commonapp.org/blog/common-app-announces-2024-2025-common-app-essay-prompts

[21] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[22] Arnaud D'Argembeau, Helena Cassol, Christophe Phillips, Evelyne Balteau, Eric Salmon, and Martial Van der Linden. 2014. Brains creating stories of selves: the neural basis of autobiographical reasoning. *Social Cognitive and Affective Neuroscience* 9, 5 (May 2014), 646–652. https://doi.org/10.1093/scan/nst028

[23] Tolulope Famaye, Cinamon Sunrise Bailey, Ibrahim Adisa, and Golnaz Arastoopour Irgens. 2024. "What Makes ChatGPT Dangerous Is Also What Makes It Special": High-School Student Perspectives on the Integration or Ban of Artificial Intelligence in Educational Contexts. *International Journal of Technology in Education* 7, 2 (2024), 174–199. https://eric.ed.gov/?id=EJ1426671

[24] Robyn Fivush and Azriel Grysman. 2022. Narrative and gender as mutually constituted meaning-making systems. *Memory, Mind & Media* 1 (Jan. 2022), e2. https://doi.org/10.1017/mem.2021.4

[25] Hangzhi Guo, Pranav Narayanan Venkit, Eunchae Jang, Mukund Srinath, Wenbo Zhang, Bonam Mingole, Vipul Gupta, Kush R. Varshney, S. Shyam Sundar, and Amulya Yadav. 2024. Hey GPT, Can You be More Racist? Analysis from Crowdsourced Attempts to Elicit Biased Content from Generative AI. https://doi.org/10.48550/arXiv.2410.15467

[26] Julian Hasford, Colleen Loomis, Geoffrey Nelson, and S. Mark Pancer. 2016. Youth Narratives on Community Experiences and Sense of Community and Their Relation to Participation in an Early Childhood Development Program. *Youth & Society* 48, 4 (July 2016), 577–596. https://doi.org/10.1177/0044118X13506447

[27] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 71:1–71:29. https://doi.org/10.1145/3534561

[28] Linda A. Jackson, Alexander von Eye, Hiram E. Fitzgerald, Yong Zhao, and Edward A. Witt. 2010. Self-concept, self-esteem, gender, race and information technology use. *Computers in Human Behavior* 26, 3 (May 2010), 323–328. https://doi.org/10.1016/j.chb.2009.11.001

[29] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581196

[30] Jean Kilbourne. 1994. Still killing us softly: Advertising and the obsession with thinness. In *Feminist perspectives on eating disorders*. The Guilford Press, New York, NY, US, 395–418.

[31] Jinhee Kim, Hyunkyung Lee, and Young Hoan Cho. 2022. Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education. *Education and Information Technologies* 27, 5 (June 2022), 6069–6104. https://doi.org/10.1007/s10639-021-10831-6

[32] Anna Kirkland and Ben B. Hansen. 2011. "How Do I Bring Diversity?" Race and Class in the College Admissions Essay. *Law & Society Review* 45, 1 (2011), 103–138. https://www.jstor.org/stable/23011960

[33] Selma Korlat, Julia Holzer, Marie-Therese Schultes, Sarah Buerger, Barbara Schober, Christiane Spiel, and Marlene Kollmayer. 2022. Benefits of Psychological Androgyny in Adolescence: The Role of Gender Role Self-Concept in School-Related Well-Being. *Frontiers in Psychology* 13 (May 2022). https://doi.org/10.3389/fpsyg.2022.856758 Publisher: Frontiers.

[34] Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1433–1445. https://doi.org/10.1145/3630106.3658981

[35] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications. Google-Books-ID: nE1aDwAAQBAJ.

[36] Isack Lee and Haebin Seong. 2025. BiasJailbreak:Analyzing Ethical Biases and Jailbreak Vulnerabilities in Large Language Models. https://doi.org/10.48550/arXiv.2410.13334 arXiv:2410.13334 [cs].

[37] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. https://doi.org/10.1145/3630106.3658975

[38] Shanyan Lin, Danni Liu, Gengfeng Niu, and Claudio Longobardi. 2022. Active Social Network Sites Use and Loneliness: the Mediating Role of Social Support and Self-Esteem. *Current Psychology* 41, 3 (March 2022), 1279–1286. https://doi.org/10.1007/s12144-020-00658-8

[39] Majse Lind, Sebnem Ture, Dan P. McAdams, and Henry R. Cowan. 2024. Narrative Identity, Traits, and Trajectories of Depression and Well-Being: A 9-Year Longitudinal Study. *Psychological Science* 35, 12 (Dec. 2024), 1325–1339. https://doi.org/10.1177/09567976241296512

[40] Louis Lippens. 2024. Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100054. https://doi.org/10.1016/j.chbah.2024.100054

[41] Jazper Lu. 2024. Duke no longer giving numerical rating to standardized testing, essays in undergraduate admissions. *The Chronicle* (Feb. 2024). https://www.dukechronicle.com/article/2024/02/duke-university-undergraduate-admissions-changes-numerical-rating-standardized-testing-essays-covid-test-optional-ai-generated-college-consultants

[42] Herbert W. Marsh and Andrew J. Martin. 2011. Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology* 81, 1 (2011), 59–77. https://doi.org/10.1348/000709910X503501

[43] Nicole Martins and Kristen Harrison. 2012. Racial and Gender Differences in the Relationship Between Children's Television Use and Self-Esteem: A Longitudinal Panel Study. *Communication Research* 39, 3 (June 2012), 338–357. https://doi.org/10.1177/0093650211401376 Publisher: SAGE Publications Inc.

[44] Dan P. McAdams. 1995. What Do We Know When We Know a Person? *Journal of Personality* 63, 3 (1995), 365–396. https://doi.org/10.1111/j.1467-6494.1995.tb00500.x

[45] Dan P. McAdams. 2015. Three Lines of Personality Development. *European Psychologist* 20, 4 (Oct. 2015), 252–264. https://doi.org/10.1027/1016-9040/a000236

[46] Dan P. McAdams. 2019. "First we invented stories, then they changed us": The Evolution of Narrative Identity. *Evolutionary Studies in Imaginative Culture* 3, 1 (Dec. 2019), 1–18. https://doi.org/10.26613/esic.3.1.110

[47] Dan P. McAdams, Jack J. Bauer, April R. Sakaeda, Nana Akua Anyidoho, Mary Anne Machado, Katie Magrino-Failla, Katie W. White, and Jennifer L. Pals. 2006. Continuity and Change in the Life Story: A Longitudinal Study of Autobiographical Memories in Emerging Adulthood. *Journal of Personality* 74, 5 (2006), 1371–1400. https://doi.org/10.1111/j.1467-6494.2006.00412.x

[48] Katelyn Y. A. McKenna and John A. Bargh. 1999. Causes and Consequences of Social Interaction on the Internet: A Conceptual Framework. *Media Psychology* 1, 3 (Sept. 1999), 249–269. https://doi.org/10.1207/s1532785xmep0103_4

[49] Katelyn Y. A. McKenna and John A. Bargh. 2000. Plan 9 from cyberspace: The implications of the Internet for personality and social psychology. *Personality and Social Psychology Review* 4, 1 (2000), 57–75. https://doi.org/10.1207/S15327957PSPR0401_6

[50] Kate C. McLean, Monisha Pasupathi, and Jennifer L. Pals. 2007. Selves Creating Stories Creating Selves: A Process Model of Self-Development. *Personality and Social Psychology Review* 11, 3 (Aug. 2007), 262–278. https://doi.org/10.1177/1088868307301034

[51] Kate C. McLean, Moin Syed, Monisha Pasupathi, Jonathan M. Adler, William L. Dunlop, David Drustrup, Robyn Fivush, Matthew E. Graci, Jennifer P. Lilgendahl, Jennifer Lodi-Smith, Dan P. McAdams, and Tara P. McCoy. 2020. The empirical structure of narrative identity: The initial Big Three. *Journal of Personality and Social Psychology* 119, 4 (2020), 920–944. https://doi.org/10.1037/pspp0000247

[52] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1699–1710. https://doi.org/10.1145/3593013.3594109

[53] Bernard Mokam. 2024. After Affirmative Action Ban, They Rewrote College Essays With a Key Theme: Race. *The New York Times* (Jan. 2024). https://www.nytimes.com/2024/01/20/us/affirmative-action-ban-college-essays.html Section: U.S..

[54] Gianluca Mondillo, Simone Colosimo, Alessandra Perrotta, Vittoria Frattolillo, Cristiana Indolfi, Michele Miraglia del Giudice, and Francesca Rossi. 2025. Jailbreaking large language models: navigating the crossroads of innovation, ethics, and health risks. *Journal of Medical Artificial Intelligence* 8, 0 (March 2025). https://doi.org/10.21037/jmai-24-170 Number: 0 Publisher: AME Publishing Company.

[55] Kimberly A. Neuendorf. 2017. *The Content Analysis Guidebook*. SAGE. Google-Books-ID: nMA5DQAAQBAJ.

[56] Beatrice Nolan. 2023. Yes, ChatGPT can help with your college admissions essay. Here's what you need to do to stay within the rules. *Business Insider* (Oct.

2023). https://www.businessinsider.com/ai-chatgpt-college-admission-essays-writing-rules-education-2023-10
[57] OpenAI. 2024. Introducing OpenAI o1. https://openai.com/index/introducing-openai-o1-preview/
[58] OpenAI. 2025. ChatGPT Pricing. https://openai.com/chatgpt/pricing/
[59] OpenAI. 2025. How ChatGPT and our foundation models are developed \textbar OpenAI Help Center. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed
[60] Dana Pessach and Barbara Poblete. 2024. Gender Representation Across Online Retail Products. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 947–957. https://doi.org/10.1145/3630106.3658947
[61] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 53728–53741. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html
[62] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828
[63] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. (2014).
[64] Darshana Sedera and Sachithra Lokuge. 2020. Flaws in flawlessness: Perfectionism as a new technology driven mental disorder. *ICIS 2020 Proceedings* (Dec. 2020). https://aisel.aisnet.org/icis2020/societal_impact/societal_impact/8
[65] Rashi Shrivastava. 2024. Did You Use ChatGPT On Your School Applications? These Words May Tip Off Admissions. *Forbes* (Feb. 2024). https://www.forbes.com/sites/rashishrivastava/2024/02/05/chatgpt-college-school-applications-admissions-red-flags-ai/
[66] Natasha Singer. 2023. Ban or Embrace? Colleges Wrestle With A.I.-Generated Admissions Essays. *The New York Times* (Sept. 2023). https://www.nytimes.com/2023/09/01/business/college-admissions-essay-ai-chatbots.html Section: Business.
[67] PETER SMAGORINSKY. 1997. Personal Growth in Social Context: A High School Senior's Search for Meaning in and Through Writing. *Written Communication* 14, 1 (Jan. 1997), 63–105. https://doi.org/10.1177/0741088397014001002
[68] Erin E. Toolis and Phillip L. Hammack. 2015. The lived experience of homeless youth: A narrative approach. *Qualitative Psychology* 2, 1 (2015), 50–68. https://doi.org/10.1037/qup0000019
[69] US Supreme Court. 2023. Students for Fair Admissions, Inc. v. President and Fellows of Harvard College.
[70] Amy Vidali. 2007. Performing the Rhetorical Freak Show: Disability, Student Writing, and College Admissions. *College English* 69, 6 (2007), 615–641. https://www.jstor.org/stable/25472242
[71] L. Monique Ward and Petal Grower. 2020. Media and the Development of Gender Role Stereotypes. *Annual Review of Developmental Psychology* 2, Volume 2, 2020 (Dec. 2020), 177–199. https://doi.org/10.1146/annurev-devpsych-051120-010630 Publisher: Annual Reviews.
[72] James Warren. 2013. The Rhetoric of College Application Essays: Removing Obstacles for Low Income and Minority Students. *American Secondary Education* 42, 1 (2013), 43–56. https://www.jstor.org/stable/43694176
[73] Dwayne Wood and Scott H. Moss. 2024. Evaluating the impact of students' generative AI use in educational contexts. *Journal of Research in Innovative Teaching & Learning* 17, 2 (July 2024), 152–167. https://doi.org/10.1108/JRIT-06-2024-0151
[74] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. 2024. Don't Listen To Me: Understanding and Exploring Jailbreak Prompts of Large Language Models. *USENIX '24* (2024).
[75] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology* 16, 1 (Aug. 2016), 93. https://doi.org/10.1186/s12874-016-0200-9