

# CineGPT: Generative A.I. In Screenwriting

Ryan Harding  
ry287616@ucf.edu  
University of Central Florida  
Orlando, Florida, USA

## Abstract

The use of A.I. in Screenplay has been a controversial topic in the film industry, with arguments for it as an escape from barriers of entry and arguments against dealing with ethical and authorship issues. To determine both the potential uses and limitations, as well as to validate and test concerns with the technology, I created an original story treatment and screenplay and ran through several generative A.I. experiments to test the potential of ChatGPT in these fields. I generated A.I. screenplays from a treatment and compared against my original work, and used A.I. to critique and analyze the resulting scripts.

## Artifact and Resource Repository:

<https://github.com/rharding8/CineGPT>

## 1 Introduction

As someone invested both in technology and filmmaking, the use of A.I. in filmmaking is one that gives me pause. Some screenwriters noted fears over A.I. replacing them, or losing valuable credit and residuals due to the use of A.I., during the 2023 Writer's Guild of America Strikes [5]. Other screenwriters instead have expressed belief in A.I. being unable to eclipse human work, supporting its use in brainstorming, storyboard, and revision. Even some of these writers, however, have noted these uses can encroach on entry-level writer positions in television [3]. Because of this, I decided to research within Topic 14 (Natural Language to Video Storyboarding), to determine where A.I. falls short in filmmaking, where it succeeds, and how to limit its role in replacing human creativity while enhancing its role in aiding it.

## 2 Related Work

Many studies have gone into the technical and creative sides of this issue, exploring how screenplay analysis and generation and storyboard generation work computationally, the concerns associated with such, and the creative implications. A study by Sabyasachee Baruah and Shrikanth Narayanan went into detail about the limitations and concerns of character coreference resolution, a subset of named-entity-recognition (NER) related to characters within the screenplay. Baruah and Narayanan discussed how screenplay length often exceeds the limits of traditional transformers, requiring a new set of approaches using BERT models as well as Bidirectional RNNs to parse and resolve references in a screenplay. They also noted concerns about non-linear stories and characters being referred to as a plural (E.g. "Kids"), which I've taken into account to make sure my work can be analyzed properly [1].

NER concerns were also noted by Kyle Jorgenson and Haohong and Mea Wang from the University of Calgary, who conducted a study and experiment to create a simple script-to-film pipeline using A.I. to generate an animated film from a human-written screenplay.

They believed such a tool could remove barriers of entry to people looking to make films, allowing writers to take control into their own hands and produce their vision without any other technical skill required [4]. Their program, using script coverage and NER through transformers, takes advantage of pre-made assets and a Unity environment to allow a writer to automatically generate environments and place characters within it. They can even line-up recorded voices with dialogue in the script to automatically voice the film [4]. The program essentially takes the role of the crew, while the writer becomes director and producer of the film. However, even this experiment still required considerable user-input and decision-making, proving A.I. analysis and placement was still imperfect [4].

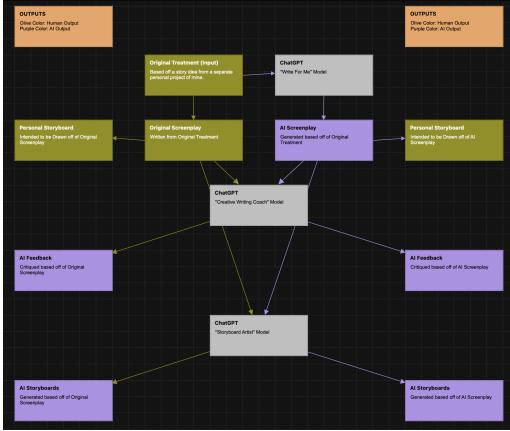
A more creatively-minded study into A.I. for script generation, rather than just analysis and visual generation, was conducted by Susan Cake, who explored using ChatGPT and other models as both a collaborator and primary writer. Cake noted that such tools had already become commonplace in analytics and post-production tasks, risking infringing on jobs in those fields, and could raise concern for writers as well [2]. Cake explored risks such as the homogeneity of A.I.-produced scripts, copyright infringement concerns about scripts used to train Large Language Models (LLMs), and legal concerns around guidelines in different nations around media involving sensitive subjects [2]. The latter being especially relevant given recent controversy around DeepSeek and censorship around subjects such as Tiananmen Square [6].

Cake was also concerned about cultural prejudices, especially from a non-American perspective she found ChatGPT to typically lean into American culture and screenwriting conventions even when given Australian stories [2]. She ultimately found that results were not up to par with human work, with ChatGPT 3.5 creating cliched dialogue and generic plotlines, sometimes lacking basic cause-and-effect. True to her concerns, she found the results could contain copyrighted material and harmful stereotypes and often leaned towards generic "happy endings" [2]. Cake also noted that ChatGPT was better at critiquing work than generating it, but that critiques of A.I.-generated material were almost universally positive and failed to point out obvious flaws, being inconsistent with the quality of the material [2]. Because of this I decided to keep an eye out for these issues in my experiment.

## 3 Methodology

### 3.1 The First Act

The experiment began by coming up with an initial story idea. I took a piece of an original screenplay I'd already been working on, the opening montage, and used it as a seed to create a short film story. I focused on writing a heavily detailed treatment. Then I got to work on the actual screenplay based on the treatment. It was during this time when I also did most of my wider research



**Figure 1: The original intended pipeline for this project, before the screenplay generation experiment was expanded and the storyboard generation experiment was truncated**

for the project, as writing the script itself was the longest portion of the project. I initially sought to write at least one draft of the screenplay *before* any NLP usage was involved, however I decided to initially focus on a first act for the purposes of the milestone earlier this semester. I took that complete first act through the pipeline in Figure 1, minus the storyboard generation.

I started by giving ChatGPT, using the "Write For Me" model, the first act of my treatment and instructed it to write a 5-8 page first act screenplay based on it. I then made hand annotations as notes on the screenplay, as well as similar notes on my original screenplay. Finally, I used OpenAI's "Creative Writing Coach" model in ChatGPT to give AI-Generated Notes on both screenplays.

### 3.2 Milestone Results and Adjustments

I started with a wide range of expectations for the A.I. screenplay, going from an incoherent mess to human quality. I was surprised to see the generated first act was coherent, but only by virtue of following the treatment nearly verbatim. Despite entering a prompt asking for a 5-8 page act, the resulting script was only 3 pages, one more than my treatment. For the most part, the A.I. script just put the treatment into screenplay format, expanded a few lines of dialogue, and added brief moments of imagery. One such moment actually impressed me enough to make note of it for making an image I hadn't thought of, which can be found in Figure 2 [7].

However, the barebones nature of it made it more of an extrac-

INT. DETENTION ROOM - LATER  
Carlos slouches, doodling in the margins of his worksheet. A sharp SHADOW crosses his desk. The Teacher picks up his notebook, flipping through it. — *Actual decent imagery*  
The Teacher's stern face SOFTENS as they take in the astonishingly detailed artwork.  
TEACHER  
Carlos... these are incredible. Why aren't you in an art class?

**Figure 2: A decent use of imagery generated by the first act of the first A.I. script**

tive translation than an abstractive expansion. It also made key mistakes in formatting such as using capitalization across the script at random, not just for character names [7]. And the changes it did make to the characters or narrative were either confusing, clichéd, or actively undermining the intent behind the story.

For example, Carlos is explicitly defined as arrogant and womanizing in my treatment, leading to a moment where he defends a woman but also hits on her. In the AI screenplay, his tone is not only changed to be harmlessly playful, but his comment is entirely one of support [7]. The two examples can be seen below, along with my note about Carlos being portrayed as "too nice".

A young woman who is unmistakably the woman from the opening curtains walks into the room. She is carrying a painting and is painting out of her bag awkwardly. A rude MALE STUDENT yells at her to get out. Sera's already won Carlos hearts this and, in a moment of pure joy, she turns around while implicitly hitting on the girl, DANIELA WILKINS.  
A Rude Male Student answers.  
MALE STUDENT  
Give it up, sera already won.  
Carlos snarks, but his play tone now has a playful edge.  
SERA  
C'mon, man. Let's see what she's [redacted] for my  
got.

**Figure 3: Original Treatment Description**

**Figure 4: Sanitized A.I. Script Interpretation**

The A.I. first act made no reference to Carlos being a womanizer at all. This overt sanitization felt like the LLM wanting to avoid any sort of prejudice or bias, without recognizing that characters being flawed can be intentional.

I had also tested the ability of A.I. to critique and give notes on my own first act, which I had made several critiques and notes on myself. When the "Creative Writing Coach" was tested on my own first act, it gave me valuable feedback including pacing issues, inconsistencies, and how I could improve the commentary of the script [7]. However, not only did it miss some obvious problems with my first act, including logistical issues of realism, but it too fell prey to the LLM's attempt to avert bias and prejudice by sanitizing some characters. I refer to this criticism of Dahlia's portrayal:

Maybe have her arrive in a different way—she doesn't just rush in breathless but makes an entrance that asserts her presence in a powerful way [7].

This ignores that Dahlia is set up as struggling a bit, yet proving herself fierce nonetheless. The A.I. suggestion wasn't concerned with staying true to the character or her arc, but rather trying as hard as possible to prove itself unbiased.

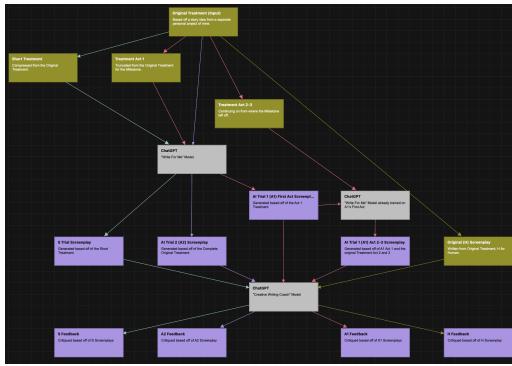
Finally, I had set ChatGPT to critique the A.I.-generated first act. Consistent with Cake's observations, the resulting critique was almost universally positive, giving the LLM-generation more credit than it deserved and identifying character motivation and detail completely absent from the script itself.

At that point I decided to make some adjustments to my process moving forwards:

- (1) In addition to continuing the existing generated screenplays, I would also do a second A.I. screenplay from scratch once the script was finalized. My theory was that having a complete treatment given for a complete script could give more context to character arcs and potentially ease the bias issues I noted.
- (2) I would add yet a fourth screenplay to the pipeline by simplifying my treatment down to a two-page treatment, and set ChatGPT to create a screenplay from that as well. My theory being that less railroading and information might give the model more flexibility.

- (3) Shortly thereafter, due to the expansion of the screenplay generation experiments, time constraints, and early personal storyboard experiments not being useful due to the lack of a proper talented storyboard artist, I decided to drop the storyboard generation portion of the research. That deserves its own experiments and study, as well as a proper baseline with genuine storyboards that I couldn't guarantee in the duration of this project.

### 3.3 The Complete Pipeline



**Figure 5:** The final pipeline for this project. Each line color traces a different variation [H, S, A1, A2] of the script.

The above is the more complex, comprehensive process I ended up using once the milestone passed. I split my workflow into four separate screenplay projects:

**H** The 'H' screenplay, or "Human" screenplay, was my own script that I worked on almost all semester long, building off of my original treatment.

**A1** The 'A1' screenplay, or "A.I. Trial 1" screenplay, was the continuation of the screenplay that ChatGPT started back in February. I fed the model both its own Act 1 script and my treatment for Act 2 and 3 to get a consistent result.

**A2** The 'A2' screenplay, or "A.I. Trial 2" screenplay, was a from-scratch screenplay that ChatGPT built off of my complete treatment.

**S** The 'S' screenplay, or "Short Treatment" screenplay, was a screenplay that ChatGPT built off of a much-truncated and compressed version of my treatment.

Once these screenplays were complete, I followed the same process for each, making my own notes as well as feeding them back into the "Creative Writing Coach" model for feedback, and this time receiving quantitative measurements defined by the model based on the metrics I set out for it. All screenplays, treatments, as well as complete conversation histories are found in the Github Repository.

## 4 Evaluation and Results

### 4.1 Metrics

Early on in the project I knew I wanted to judge the screenplays on coherence, pacing, and imagery. For the A.I.-generated critiques, the model defines how to measure these values. For my own human evaluation I decided to lay out a simple points system as follows:

**Coherency** For coherency, I start reading one of the screenplays with 10 points of coherency, and mark off one point for every break in logic or character. Below 5 points I would deem in need of significant editing, below 0 being a mess.

**Pacing** Similarly, for pacing I start out at 10 points and mark off one for every instance where the story feels noticeably rushed or compressed.

**Imagery** For imagery, I start off at 0 and add a point for every standout image or piece of dialogue I find. Above 5 is good, but 8-10 is where I'd like to be at for a 20+ page screenplay.

In terms of measuring the critiques generated by the A.I., I decided to judge on accuracy and constructiveness:

**Accuracy** For Accuracy, I follow the same pattern as Coherency and Pacing and subtract points for every problem, only from 5 instead of 10. In this case a problem being the critique extrapolating something clearly not there, embellishing one of the A.I. screenplays, or claiming a flaw as a positive.

**Constructiveness** Much like imagery, this is an additive metric. Every actionable suggestion for improvement on a screenplay is a point, with 5 being a positive result.

### 4.2 Human Evaluation

#### A1 A.I. Trial 1

*Logical Placement*  
 A hush falls over the room. The crowd shifts—many leaving Carlos's side to admire Dahlia's masterpiece. Carlos, frozen, studies the painting. Even his admirers can't hold his attention. He steps closer, finally speaking.

**Figure 6:** A moment where the logical placement of people in a scene feels completely off in the A.I. generated script

**Coherency** was a weak spot with A1. The first act that I had already studied by the milestone contained 5 breaks in coherency alone, including deviations-of-character such as Carlos being sanitized and less of a bully, wrong location placements, and even the staging of an early scene being off, which can be seen in Figure 6. Results actually worsened in the second and third act, with multiple out-of-character decisions and dialogue, unclear time jumps, incorrect formatting and terminology leading to a confusing screenplay, and one particularly odd moment of narration being added to the screenplay despite otherwise completely lacking it, as can be seen in Figure 7. Acts 2 and 3 featured 9 breaks in coherency, rendering my score for this section at a -4/10 overall.

CARLOS (V.O.)  
 I only ever painted longing. And now I have everything.  
*No narration*

**Figure 7:** A sudden change in tone occurs when the A.I. decides the script suddenly has voiceover

**Pacing** and length in general wound up being a recurring problem across all the A.I. scripts. Despite constantly prompting for a 20-30 page screenplay, none of the A.I.

screenplays go over 10 pages long, and most of them hang around the exact same length as the treatment I fed them. As such, pacing is practically non-existent across these scripts.

A1 consistently rushed past major, important scenes, while also giving unnecessary time to unimportant ones. I made note of 9 times I felt the pacing was rushed in unnecessary montages or scenes lacking detail or breathing room, rendering my score a 1/10.

**Imagery** actually wasn't terrible here. Dialogue was often cliche and never really stood out, but a few visual images did. I will once again highlight Figure 2 as decent image, though it was the only one that stuck with me enough to give it a single point.

## A2 A.I. Trial 2

Stunned. Dozens of jaw-dropping sketches.  
 Why the ~~hell~~ TEACHER (CONT'D) = Unprofessional  
 My parents. They say art won't pay ~~rent~~ = Decent  
 You have a gift. Let me talk to someone. CARLOS  
 Carlos hesitates. Nods.  
 They walk to the door...

**Figure 8: A teacher suddenly uses language they never used in the treatment, though Carlos' response is a memorable line**

**Coherency** was improved a slight bit compared to A1, only having 12 logical breaks rather than 14, giving a -2/10 score. However, some of the breaks from logic here are particularly odd. A high school teacher who didn't swear in the treatment suddenly does (Figure 8), Carlos still acts too "nice", the script implies Dahlia and Carlos already knew each other before meeting, there's conversations happening in different places to in the treatment, the hospital doesn't seem to have doctors, and not only is there once again a strange added voiceover, there's even some sort of omniscient narrator added simply dubbed "Night" (Figure 9).

WEDDING - DAY  
 Joy. Artist friends jealous. Promises of love and family.  
 Carlos sneaks to the canvas. He stares. Nothing comes.  
 -- Dahlia thrives. Carlos tries to paint. Nothing.  
 MONTAGE:  
 ? NIGHT ?  
 What?

**Figure 9: A.I. can still struggle with NER, and believe a time of day may be a character or narrator**

**Pacing** is once again a big problem with this script overusing montages and rushing past scenes far too fast. In fact, many scenes are rushed enough to become classified as coherency errors rather than just pacing errors. However, it was also still an improvement over Trial 1, having 6 major pacing infractions as opposed to 9.

INT. HOUSE - NIGHT  
 He shows her his paintings. She asks why he stopped.  
 CARLOS  
 I can only paint Mommy. And that hurts.  
 BAILEY  
 Then paint me too. Okay

**Figure 10: A rare moment of a memorable piece of dialogue in these scripts**

**Imagery** A2 did not have any memorable images like A1 did, but it did contain two moments of memorable dialogue, including one for the little girl that gives her a moment of direct agency over her father that actually serves the scene quite well (Figure 10). Thus I'll give A2 a 2 for imagery.

## S Short Treatment Script

**Coherency** was a bit interesting with this script. In terms of actual hard breaks of logic or character or context, I only really noticed 5, but that may be because the script is so marathon in pacing that it doesn't really even feel like there's a flow to "break" from necessarily. Some context is missing, but the characters somehow fit my original treatment better than they did in the other A.I. trials. I'd say a 5/10 coherence score is odd, but technically fair.

**Pacing** This script has the pacing problems the worst. As can be spotted in Figure 11, the model chose to space out lines way more than is standard, meaning this script is not only the shortest of the bunch, but is shorter even compared to my original treatment. Everything is rushed through to the point where I stopped counting the breaks from pace at some point. Looking over it again, I'd say every single one of the 16 scenes feels extremely rushed, so I'm going to judge this at a -6/10.

INT. HOSPITAL - DAY  
 Dahlia cradles the baby, kisses her, begins to wheeze. Machines SCREAM.  
 Flatline.  
 Carlos cradles his daughter, forcing a smile.  
 Decent imagery

**Figure 11: A memorable audiovisual image in an otherwise extremely barebones script**

**Imagery** This script lacked memorable dialogue entirely, but did have so far the best use of audiovisual imagery in the A.I. scripts, which I show in Figure 11. So another 1 point for imagery here.

**H** In terms of judging my own screenplay, I felt it wouldn't be fair to try to present an objective critique of it. I will say that I caught some coherence and pacing issues and feel I could improve the imagery. I annotated my script the same way I did the A.I. ones in the Git Repo. But for an actual measurement I decided to leave that to the ChatGPT model.

### 4.3 Critiques

#### A1 A.I. Trial 1

**What's Working**

**Coherence: 9/10**  
The story flows clearly and consistently from scene to scene. Carlos's arc—from underappreciated teen artist to burned-out golden boy, to grieving father, to emotional healing through his daughter—is well-constructed and emotionally resonant. The transitions especially between art school → gallery → domestic decline are tight and make the character journey easy to follow without losing nuance.

The circular structure—the opening painting reappearing at the emotional climax—is elegant. The emotional beats hit hard and feel earned.

**Pacing: 8/10**  
The screenplay moves efficiently, never lingering too long, but also not rushing through moments. The quick-cut montages (e.g., “Dahlia rises,” “Carlos falls”) are effective in conveying time passage and emotional descent. One or two sections, particularly the jump from marriage to Dahlia’s illness, could benefit from a breather to deepen their impact, but the pacing overall works well.

**Imagery: 9/10**  
This is where the script really shines. Visual metaphors are rich throughout:

- The **opening painting** is a symbolic split between aspiration and adoration
- Carlos’s **sketchbook** vs. the **whiteboard**—his creativity in contrast to rigid expectation
- Dahlia’s **curved garden** vs. the **straight-lined city**—her artistic freedom vs. society’s control
- **Black canvas** as emotional harbinger—he paints when he’s longing, stops when he’s full
- The **final shared canvas** with Dahlia and Bailey—a perfect visual culmination

The imagery is never heavy-handed, and it deeply enhances the emotional storytelling.

Figure 12: ChatGPT’s review of Script A1

**Accuracy** in these metrics is very flawed. The LM understands the fundamental idea of a character arc for coherence and not wanting to drag pace, but not the idea of logical flow to such things. The existence of Carlos’ character arc alone does not indicate coherence, nor does the existence of quick montages indicate an engaging pace. The LM is extrapolating a deeper meaning to the arc that wasn’t conveyed in the script, as well as misreading a rushed summary as an effective montage. As well, all five points of imagery noted are points of symbolism, not descriptive imagery, which brings it up to 7 inaccuracies with the script being judged, for an accuracy score of 3/10.

#### S Suggestions for Improvement

1. **Deepen the Emotional Pivot Post-Marriage**  
The shift from joy (marriage) to Dahlia’s illness is structurally sound but emotionally abrupt. Consider a short scene or beat showing their family life before illness strikes—maybe Carlos finally painting again, or Bailey bonding with Dahlia. This would make Dahlia’s collapse more gut-wrenching.
2. **More Dialogue in Later Scenes**  
As Carlos grows quiet in grief, his silence works—but a touch more dialogue from Bailey (especially in “why no pictures of me?” scene) could add warmth and help transition us into the hopeful finale.
3. **Characterize Dahlia Slightly More in Early Scenes**  
While Dahlia is portrayed through her art and impact, she could use a bit more dialogue or internal contrast to distinguish her beyond being a muse or rival. A line or two revealing her perspective on art or struggle would make her later fate hit harder.

Figure 13: ChatGPT’s improvements for Script A1

**Constructiveness** unlike all the other metrics I’ve used so far, was actually considerably successful in these trials. It was the one thing the model consistently did well, give actionable improvements to the script. All three suggestions for this script are good ideas that would not only make the script at hand better, but are ideas that would make my own better as well. That’s a solid 3 points from every suggestion.

#### A2 A.I. Trial 1

**Accuracy** is once again an issue. Some breaks in accuracy here include praising the rushed style as “minimalist”, claiming the characters have conveyed motivations in this script, and claiming Carlos’ response to Dahlia being pregnant was “emotional”, when Carlos didn’t actually

#### Coherence: 8.5/10

The narrative stays faithful to the original script established in Version A1. Following Carlos’s rise, fall, and healing, this version trims some transitions and dialogue, keeping most in a more minimalist aesthetic style. That streamlining helps the plot stay clear, though it comes at the cost of some emotional richness. The story remains easy to follow, and character motivations are clear.

One big plus: Dahlia’s pregnancy is better integrated. The hospital moment (“I’m pregnant”) clarity is in a more natural point in the timeline, and Carlos’s emotional response gives that beat clarity and power.

#### Pacing: 7/10

This version moves faster—almost too fast. Some beats, particularly around Carlos’s decline and emotional lows, are abbreviated. I’m in agreement.

- The **bar scenes** and **A1 Recovery** are compressed.
- Dahlia’s **rise to fame** feels more told than shown.
- The **romantic rekindling** and **proposal** are sweet but feel rushed, packed into a quick montage.

As a result, the emotional highs and lows don’t hit quite as deeply as in Version A1. The trade-off is a more efficient, visually-driven narrative—but it sacrifices some depth.

#### Imagery: 8.5/10

Still strong, but slightly less poetic than A1. Here’s what stands out:

- The **opening canvas metaphor** is maintained and remains powerful.
- Dahlia’s **curved garden** and the final **family canvas** are still excellent visual symbols.
- The **napskin sketch proposal** is a sweet, visually charming moment.

Figure 14: ChatGPT’s review of Script A2

respond to Dahlia being pregnant at all, it was skipped over. Imagery is once again confused for symbolism, outside of the napskin sketch proposal moment, which is an insightful praise to give considering I cut it out of my own script. The pacing faults are actually quite accurate, albeit undercut by not recognizing the same faults in A1. Still, that leaves the accuracy at a 5/10.

#### S Suggestions for Improvement

1. **Restore Some Emotional Beats**  
A couple of beats—especially post-breakup, during Carlos’s depression and recovery—could use more space. The current pacing clips through what should be devastating moments, reducing their emotional impact.
2. **Let Carlos’s Arc Breath Post-Marriage**  
The montage after the proposal is a bit fast and surface-level. A short scene where Carlos genuinely tries to paint and fails—something more than a quick “stares, nothing comes”—could add resonance to his creative block.
3. **Add More Distinctiveness to Dahlia**  
Dahlia gets slightly more dialogue here (the “I eclipsed you the day we met” line is sharp), but she still feels more like a symbol than a fully realized character. A scene that gives her a vulnerable or reflective moment—not just powerful ones—would round her out.

Figure 15: ChatGPT’s improvements for Script A2

**Constructiveness** once again was a success. The second two suggestions are ones that my script could benefit from, while the first still finds a room for improvement in the A.I. script being analyzed. Once again, a full 3 points for every suggestion.

#### S Short Treatment Script

#### Coherence: 8/10

This version is narratively clear and emotionally focused. The story remains easy to follow and hits the key beats from the other versions—Carlo’s artistic awakening, his rise and fall, the tension with Dahlia, loss, and eventual healing with Bailey.

It does a good job maintaining the **emotional throughline** and **character arcs**, despite being a very stripped-down version. The narrative still has a satisfying arc and a poignant close.

#### Pacing: 8.5/10

The pacing here is **tight and effective**. This is the most economical version—it wastes no time. Every scene exists to push the story forward or land an emotional beat.

However, this comes at a cost. While it flows swiftly, **some transitions are abrupt**, especially:

- From college rivalry → love → tension → fallout (a lot of complexity is compressed).
- The **reunion and recovery** feels like a quick pit stop instead of a deep transformation.

So the pacing is sharp, but emotional pacing could benefit from more nuance.

#### Imagery: 6.5/10

This is where the draft is the thinnest. The imagery is functional rather than evocative. Yes, the **split-canvas** painting returns, and Dahlia’s legacy lives on in Bailey’s drawing—but compared to the lush metaphor work in other drafts, this version is much more reserved.

Figure 16: ChatGPT’s review of Script S

**Accuracy** is at an all-time low with this review. Emotion was hardly the focus on the S script, there weren’t really character arcs, scenes don’t land emotional beats or really push the story forwards. Once more, imagery is

confused for metaphor and symbolism, rather than descriptive details and setting. The pacing faults, once again, aren't necessarily wrong so much as they are issues the LM should've caught in the other scripts. This reads at around 4/10 accuracy.

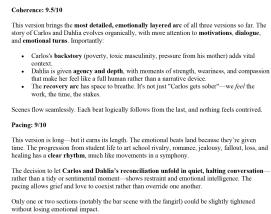
#### ⌘ Suggestions for Improvement

1. **Layer in a Few Visual Motifs**  
Add back some visual/emotional touchstones from the earlier versions—Carlos's “blank canvas” as metaphor, the recurrence of specific images (trees, longing, light vs shadow, etc.) to deepen the imagery.
2. **Let Major Emotional Beats Breathe**  
Just one or two more lines or moments of internal struggle (e.g., at the AA meeting, or after Dahlia's death) would let the emotional stakes settle and resonate.
3. **Give Dahlia Slightly More Voice**  
She's a powerful figure, but only gets a few sharp lines. Just one small scene that shows her reflecting on her art, or her fears about motherhood, could make her death more devastating.

**Figure 17: ChatGPT's improvements for Script S**

**Constructiveness** is slightly less well-handled here. Suggestions 2 and 3 are legitimate improvements, while suggestion 1 is spent on simply asking for more metaphors, which doesn't really seem like the solution to the larger problems of the script. 2/3.

#### H Human Script



**Figure 18: Coherence and Pacing scores on my screenplay**

**Accuracy** is certainly better when given a human script to judge. That the model understands the point of Carlos' flaws and toxicity is a significant step up from what we've seen so far. However, it could still be more critical. The recovery arc doesn't really have space to breathe, it's jumped over in a timeskip even in my script. And once again, every attempt to explain imagery is once again overly focused on symbolism rather than the visual detail I was looking to judge. 6/10 accuracy here.

#### ⌘ Suggestions for Improvement

1. **Trim Some Redundancy**  
The bar scene with the fan girl, while adding realism, could be shortened slightly—it risks repeating emotional beats we've already seen (Carlos's bitterness, public perception, temptation).
2. **Polish Some Dialogue for Pacing**  
The emotional conversations are strong, but occasionally too long. Shortening some exchanges slightly—without losing depth—could keep scenes crisp.
3. **Visual Transitions Could Be Sharper**  
A few scene transitions (e.g., from the slums-dream to AA) could use visual contrast or crossfades to make time jumps smoother. This is a minor note but would help a filmed version flow more cleanly.

**Figure 20: ChatGPT's improvements for my screenplay**

**Constructiveness** is actually considerably weaker here. Many of the suggestions offered the A.I. scripts are more tangible and realistic improvements. These suggestions include a slight redundancy as two of them deal mostly with trimming down overlong dialogue scenes, and the last one is a visual nitpick rather than a deeper issue. Only 1/3 this time.

## 5 Discussion and Challenges

### 5.1 Insights

In spite of how far LLMs have come, it's clear in these results that there's still a fundamental lack of understanding of creative or narrative structures and storytelling in them. Given the concerns raised during the WGA strikes, I had expected to find A.I. at least capable of creating something approaching human level. But what I found is that even with detailed instructions, these models won't really generate much beyond what you give them. Once more, I instructed every single model to write at least 20 pages, and the most I ever received back was 6.

More telling was how poorly ChatGPT seemed to understand analyzing its own scripts, often finding its own work to be near-perfect in spite of clear, visible flaws. It took viewing a complete human screenplay for the model to show even a rudimentary understanding of the intended arcs and ideas of the story it had been given to begin with. The fact that it was able to reach that point, and consistently hinted towards it when giving suggestions for improvement, is proof that this potential does lie in these models, there's just something holding them back.

### 5.2 Threats to Validity and Improvements

Some limitations and potential confounding variables in this study included the use of specialized ChatGPT models such as "Creative Writing Coach" and "Write For Me". Perhaps a vanilla ChatGPT session would've granted different results, or a non-ChatGPT-based model. The model never being exposed to the fully fleshed-out screenplay until critiques could also be a concern, as being trained on the final human output could potentially benefit generation. Testing entirely on one already defined story idea may also have been a limiting factor. Had I taken the idea from brainstorming with ChatGPT, it may have been able to develop the idea better. In the future it would also be wise to find ways to force the model to adhere to page counts and reach the same length of screenplay as the human script, in order to have space to generate more than a basic rewording of the treatment.

## 6 Conclusion

While this study is hardly conclusive, it does add to the increasing litany of research showing that A.I. has harsher limits in these creative endeavors than some might've believed. At no point in the process, other than suggestions for improvement during critique, did these A.I. models ever create anything approaching human quality. Ethically, that may even be a good thing, keeping these models relegated to editors and aids to professional writers, rather than competition. Even so, there is a clear potential buried underneath for these models to improve understanding of creative concepts. How that will be used is yet to be seen.

## References

- [1] Sabyasachee Baruah and Shrikanth Narayanan. 2023. Character Coreference Resolution in Movie Screenplays. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10300–10313. <https://doi.org/10.18653/v1/2023.findings-acl.654>
- [2] Susan Cake. 2025. Artificial intelligence as a collaborative tool for script development. *Media Practice and Education* 0, 0 (2025), 1–16. <https://doi.org/10.1080/25741136.2025.2454074> arXiv:<https://doi.org/10.1080/25741136.2025.2454074>
- [3] Frank Deese. 2024. Screenwriters vs. the machine: Embracing AI in the storytelling process. <https://www.rit.edu/news/screenwriters-vs-machine-embracing-ai-storytelling-process>
- [4] Kyle Jorgensen, Haohong Wang, and Mea Wang. 2023. From Screenplay to Screen: A Natural Language Processing Approach to Animated Film Making. In *2023 International Conference on Computing, Networking and Communications (ICNC)*. 484–490. <https://doi.org/10.1109/ICNC57223.2023.10074526>
- [5] Molly Kinder. 2024. Hollywood writers went on strike to protect their livelihoods from Generative AI. their remarkable victory matters for all workers. <https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/>
- [6] Donna Lu. 2025. We tried out DeepSeek. it worked well, until we asked it about Tiananmen Square and Taiwan. <https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan>
- [7] OpenAI. 2025. ChatGPT. <https://chatgpt.com/>