

RESEARCH REPORT

From Evidence to Edits

Rule-Guided Language-Model Video Editing and a
Small On-Device Student Model

Senotel Research

October 2026

Building what comes next.



RESEARCH REPORT

From Evidence to Edits: Rule-Guided Language-Model Video Editing and a Small On-Device Student Model

Senotel Research

October 2026

ABSTRACT

A companion report condensed the evidence on what keeps viewers watching short-form video into hook principles, retention principles, genre templates and length bands [1]. Here we ask whether those findings make automated editing better, and whether a small model that runs on a phone can learn to edit as well as a large one. We give the findings to a language-model editor in two ways, as an ordered reasoning brief and as a deterministic post-processor that enforces the mechanical rules on any plan, and evaluate on fourteen openly licensed recordings (52.9 minutes). Compared with the same editor without the evidence, the evidence-guided editor made edits 20.3 seconds shorter on average ($p = .010$), passed more of six automatic rule checks ($p = .014$) and brought the first payoff forward from 24.6 to about 15 seconds; a blind, order-balanced language-model judge preferred its edits on 7 recordings and the baseline's on 3, with 4 ties, a direction too small a sample to confirm ($p = .344$). We then taught the same material, lesson by lesson, to Senotel Vision-1.0, a 35 KB rule-and-weights editor that plans an edit in about 20 milliseconds without a network connection. It passed the rule checks at least as often as the large editor but did not reach it: it chose the same material far less often than the large editor agrees with itself (F1 0.33 against 0.59), the judge preferred the large editor on 11 of 13 recordings ($p = .006$), and fitting its weights to the large editor's choices did not beat the evidence-derived priors on held-out recordings. We report both results in full and describe what the negative one teaches.

Keywords: automated video editing; short-form video; language models; knowledge distillation; on-device models; evaluation

1. Introduction

Automated editors can now cut a long recording into a short video in seconds. Whether the result holds a viewer depends on decisions that are easy to state and hard to make: which moment opens the video, what context the viewer needs, what to leave out, and where to end. A companion report reviewed the evidence on these decisions and summarised it as principles labelled by the strength of their support [1]. This report tests whether that evidence improves an automated editor in practice.

We study two kinds of editor. The first is a large language model that reads the transcript and selected frames of a recording and returns a plan for the edit. The second is Senotel Vision-1.0, a small, deterministic model designed to run on a phone, which we taught the same material as a fourteen-lesson curriculum and then tried to fit to the large editor's decisions. Our contributions are:

1. a method for giving retention evidence to a language-model editor, both as an ordered reasoning brief and as a deterministic post-processor, with a computational value-of-continuing proxy and six automatic rule checks (Section 3);
2. a paired evaluation on fourteen openly licensed recordings, with two samples of the stochastic editor and a blind, order-balanced judge (Sections 5 and 6);
3. a curriculum-taught on-device student model and a frank negative result: it follows the rules but does not choose the material a large model chooses, and we identify why (Sections 4, 6 and 7).

2. Related work

Computational editing of speech-driven video. Berthouzoz and colleagues linked transcripts to interview footage and visualised good cut points from the words and the picture together [2]. QuickCut aligned narration with annotated footage and chose frame-level cuts by dynamic programming under film-editing guidelines [3]. Leake and colleagues selected a take for each line of a scripted scene with a probabilistic model of editing idioms that the user can switch on and off [4]. Fried and colleagues edited talking-head video by editing its transcript and re-rendering the face [5], and LAVE placed a language-model agent in charge of planning and executing edits from natural-language requests [6]. These systems help an editor make the edit the editor intends. The editors studied here are asked to choose the story itself, which requires an explicit account of what keeps viewers watching; that account is the subject of the companion report [1]. Our editors cut real footage only and never synthesise speech or faces.

Measuring retention. Viewing logs show that duration explains much of how much of a video is watched [7], that interest peaks coincide with visual transitions [8], and that viewers abandon slow starts within seconds [9]; short-video recommenders correct watch time for duration [10–12]. None of this measures what an editor can change inside a short video, which is why we rely on a proxy and a judge, and why we treat both with caution.

Language models as judges. Using a model to compare two outputs is cheap but biased: judges favour a position, longer answers and their own outputs [13]. We follow the published remedy of judging each pair in both orders and counting only verdicts that agree, and we use a judge from a different developer than the editor.

3. From evidence to an editor

3.1 The value-of-continuing proxy

The companion report proposes that a viewer keeps watching while the expected value of the next few seconds exceeds that of the next video in the feed [1]. We implement this as a discrete-time hazard over the edit, in steps of 0.25 seconds. At each step t the hazard of leaving is

$$h(t) = b(t) \cdot \exp(-1.4 Q + 1.0 \min(1.2, E) + 0.6 P) + 0.6 C + 0.3 \max(0, B - 3)$$

where Q is the strength of the open question (set by the hook, refreshed by lines that open a new gap, decaying with a time constant of $\max(20 \text{ s}, 0.6 \times \text{length})$ and capped low in the final 15%); E is recent value (each sentence carries a value from 0.05 for a greeting to 1 for a laugh or a high-stakes line; cuts, cutaways and on-screen text add smaller amounts; decaying with a 2.5 s time constant); P is the pull of progress, stronger when the content is visibly numbered; C is processing cost (more than two new elements at once, or speech faster than four words per second); and B is the time since anything new, with silence counting one and a half times. The base rate $b(t)$ is 0.20 per second in the first second, 0.10 until three seconds and 0.035 thereafter, and half as high again if the first word comes after 0.6 seconds, which encodes the stay-or-swipe window [9, 14]. Survival is $S(t) = \exp(-\int h)$, and the score reported here is $100 \times (\frac{1}{2} S(T) + \frac{1}{2} \text{mean } S)$, an average of estimated completion and estimated share watched.

THE PROXY IS NOT A MEASUREMENT

Its constants express directions and rough magnitudes taken from the evidence. They were fixed once, before any comparison, and never tuned, but they have not been fitted to any audience. We use the proxy only to locate weak spots and to compare edits of the same footage, and Section 7 explains why a higher score is weak evidence of a better edit.

3.2 Six automatic checks

Check	Passes when	From [1]
H1	The first word is spoken within 0.6 s of the start	RQ1 [9]
H1b	The edit does not open on a greeting	RQ1 [15, 16]
H5	Hook text is on screen within 1.5 s, has ten words or fewer, and is not a vague opener	RQ1 [17]
RQ2	The first payoff (the first line after 2 s carrying more than plain information) lands within the length band's target	RQ2 [18]
R1	No stretch longer than the band's limit passes without a new value event	RQ4 [8, 19]
R10	The edit ends within 1.2 s of the last word and not on a sign-off	RQ10 [18, 20]

Table 1. The automatic checks. The band targets for RQ2 and R1 are those of the companion report, with one second of tolerance on R1.

3.3 The language-model editor

The editor receives the transcript of a recording with word timings, a small set of frames and audio measurements, and plans the edit in two passes: a story pass that decides what the video is about and how it is told, and an edit pass that turns the story into clips, on-screen text, cutaways, sound and music. In our experiments it was a Gemini Flash-family model served through the Gemini API [21]; the service substitutes a smaller model of the same family when it is busy, and the model used for each individual run was not recorded. The *baseline* brief described a good edit in general terms. The *evidence-guided* brief changes three things.

An ordered reasoning brief. The story pass must reason in a fixed order: the genre; the length band; several hook candidates, each with its type, the gap it opens and where the footage pays it off; one hook chosen by the hook principles; then the beats laid out on the genre template with the main question and the smaller ones. It returns the genre, band, hook type, gap and payoff time, which are validated. The edit pass receives the retention principles and must report where the payoff lands. The brief is generated from the same data as the checks, so the editor and the evaluation quote one source.

A deterministic post-processor. Whatever plan comes back, a post-processor applies the mechanical principles: a greeting at the very start and a sign-off at the very end are cut; the first word is brought to within 0.5 s of the start; the edit ends within 0.7 s of the last word unless sound continues; and any silence longer than 1.1 s inside a clip is closed, keeping 0.45 s of breath, unless the gap is filled with sound such as laughter. Runs of jump cuts alternate a held 12% zoom, so that a continuation of the same shot reads as intended and the change acts as a small orienting cue [22, 23].

Shared evidence. Both passes are given the companion report's principles in condensed form, with their evidence grades.

4. Senotel Vision-1.0

A cloud editor needs a network connection and a service that may be busy. Senotel Vision-1.0 is designed to run on a phone, offline, within a small download. These constraints rule out a large language model on the device today, so Vision-1.0 is a deterministic program whose every decision is a rule or a weighted judgement traceable to the evidence.

4.1 A curriculum of fourteen lessons

Vision-1.0 was taught in fourteen lessons, from what a thought is to planning a whole edit by simulating the viewer. Each lesson has the knowledge it carries, the skill that turns it into a decision, and an automated exam the skill must pass.

Lesson	Skill	Grounded in
1. What a thought is	Rebuild whole sentences from the recogniser's fragments; drop its overlapping repeats; treat music and laughter markers as non-speech; mark the line before a laugh as a punchline	[22, 24]
2. Reading a line	Features per thought: stakes, numbers, contrast, arousal, greetings, sign-offs, filler, meta-talk, dependence on the previous line, novelty, centrality to what the recording keeps returning to, vocal emphasis	[17, 19, 25–27]
3. Genre and length	Genre from cue words (a room that laughs about once a minute is comedy); a target length per genre	[1] RQ5, RQ6
4. The hook	Score every eligible thought with the hook weights and rank them	[1] RQ1
5. The payoff	The late thought that shares the hook's subject and carries payoff wording or a peak	[18, 20, 28]
6. Context	One plain line of who or what, unless the hook already starts there	[29]
7. What earns its seconds	Choose passages rather than sentences (below)	[19, 30–32]
8–9. Order and clips	Open on the hook, then follow the original order, payoff last; adjacent thoughts become one clip	[1] RQ4, RQ10
10. Hook text	The hook thought's best whole clause, ten words or fewer, filler removed	[17, 33]
11. Cutaways	Only for something being said, timed to its words; at most three; never over the hook or the payoff	[34, 35]
12–13. Sound and music	A sound only where the eye sees a change; none on sensitive footage; no comic stings on comedy; music by genre and mood, none when the footage has its own	[36] RQ8
14. Masterclass	Plan six complete edits (three best hooks at two lengths), score each with the proxy, keep the best that still tells the whole story	[1] RQ3

Table 2. The curriculum.

Passages, not sentences. An early version picked the best sentences one by one and produced edits of fifteen fragments. Viewers absorb a story through a plot they can follow [31] and a coherent order [32], and every jump is a cut across which the viewer must re-orient [30]. The body of the edit is therefore chosen by a dynamic programme over the thoughts in their original order that maximises $\sum \text{value}(\text{kept}) - \mu \cdot \text{seconds} - \kappa \cdot \text{jumps}$, with $\kappa = 0.6$ and μ , the price of a second, found by bisection so that the edit fits its length budget. A punchline always brings its setup.

4.2 Training

After the curriculum, the weights of two skills were fitted to the evidence-guided editor's decisions on the recordings of Section 5. The choice of opening thought is a choice of one among many, so it is fitted as a conditional logit with the editor's opener as the positive; whether a thought is kept is fitted as a logistic model, a thought counting as kept when at least half of it lies in the editor's clips. Both are maximum a posteriori fits with a Gaussian prior centred on the evidence-derived weights rather than on zero, so that thirteen videos can move a weight only as far as they provide evidence for. All accuracy figures are leave-one-video-out. Edit length was calibrated separately, on the development recordings only: there, the large editor's median edit was 1.29 times Vision-1.0's, so each genre's length target was multiplied by 1.29.

4.3 Size and speed

Vision-1.0 is a single 35 KB script (13 KB compressed) on top of a 33 KB module of rules that it shares with the large editor. Planning one edit took a median of 19 ms (maximum 30 ms) across the fourteen recordings on a desktop computer.

5. Experimental setup

Footage. Fourteen recordings from Wikimedia Commons, 52.9 minutes in total, chosen to span the genres of the companion report: two personal stories, two stand-up sets, three science explainers, an interview, three tutorials (one with almost no speech), a day-in-the-life vlog, a talk and a product review. All are public domain or licensed under Creative Commons licences permitting reuse with attribution (Appendix A). Each recording was transcribed once by Whisper tiny (English) [37], run locally through Transformers.js [38], and analysed for loudness and motion; all systems received exactly the same input, transcription errors included.

Systems. *Baseline*: the language-model editor with the baseline brief and the earlier post-processing. *Baseline + post-processor*: the same baseline plans through the new post-processor, isolating the effect of enforcement with the plan held fixed. *Evidence-guided*: the evidence-guided brief and post-processor, run twice on every recording (runs 1 and 2), because the model samples and one run is one draw. *Senotel Vision-1.0.0*: Vision as frozen before the held-out evaluation. *Senotel Vision-1.0.1*: one defect corrected after the held-out evaluation (Section 6.4).

Split. Seven recordings form the development set and six the test set (Appendix A). Vision's design changes and its length calibration used the development recordings only; the weight-fitting experiments used all thirteen recordings with speech, always leave-one-video-out; Vision-1.0.0 was frozen before its test-set edits were judged. The recording with almost no speech is outside Vision's scope, as it declines to plan footage without speech.

Measures. Length; the six checks; the proxy, with the longest lull and the time of the first payoff as it estimates them; and agreement with the evidence-guided editor's second run, as the F1 score of the source seconds both edits keep and whether both open on the same moment (first clips starting within 2 s of each other, or overlapping).

Statistics. Paired comparisons over the same recordings use an exact sign-flip permutation test on the mean paired difference, with Holm adjustment across the four measures of each comparison. With fourteen recordings these tests detect only large effects.

Blind judge. For an independent view of quality, each pair of edits of the same recording was rendered as text exactly as a viewer would meet it (the opening text, every spoken line in order, the cutaways, the music and the length) and given to gpt-oss-120b [39], an open-weight model from a different developer than the editor, served by GroqCloud [40]. The judge was asked which edit a typical viewer scrolling a feed would be more likely to watch to the end and be glad to have watched, told that both editors had the same material and not to prefer an edit for its length, and allowed to call a tie (Appendix D). Every pair was judged in both orders at temperature 0; a win counts only when both orders agree, and a split counts as a tie. A second judge from another developer was planned but could not be run, because its hosted service refused every request with rate-limit errors.

6. Results

6.1 What the evidence changed in the large editor

System	Recordings	Length, s (SD)	Median, s	Proxy	Checks of 6	All 6	Longest lull, s	First payoff, s
Baseline	14	83.4 (18.8)	82.8	33.0	4.07	1/14	12.0	24.6
Baseline + post-processor	14	80.1 (18.7)	78.6	36.3	4.57	2/14	11.1	23.8
Evidence-guided, run 1	14	61.2 (30.2)	59.1	42.1	4.71	1/14	9.9	13.8
Evidence-guided, run 2	14	64.8 (35.5)	54.5	42.9	4.36	2/14	16.6	16.1
Senotel Vision-1.0.0	13	47.5 (9.9)	47.9	58.5	5.00	4/13	7.7	13.8
Senotel Vision-1.0.1	13	47.5 (9.9)	47.9	58.3	5.00	4/13	7.7	13.8

Table 3. Means over the recordings each system planned. Proxy: the value-of-continuing score (0 to 100, uncalibrated). Checks: of the six in Table 1. Longest lull and first payoff are the proxy's estimates.

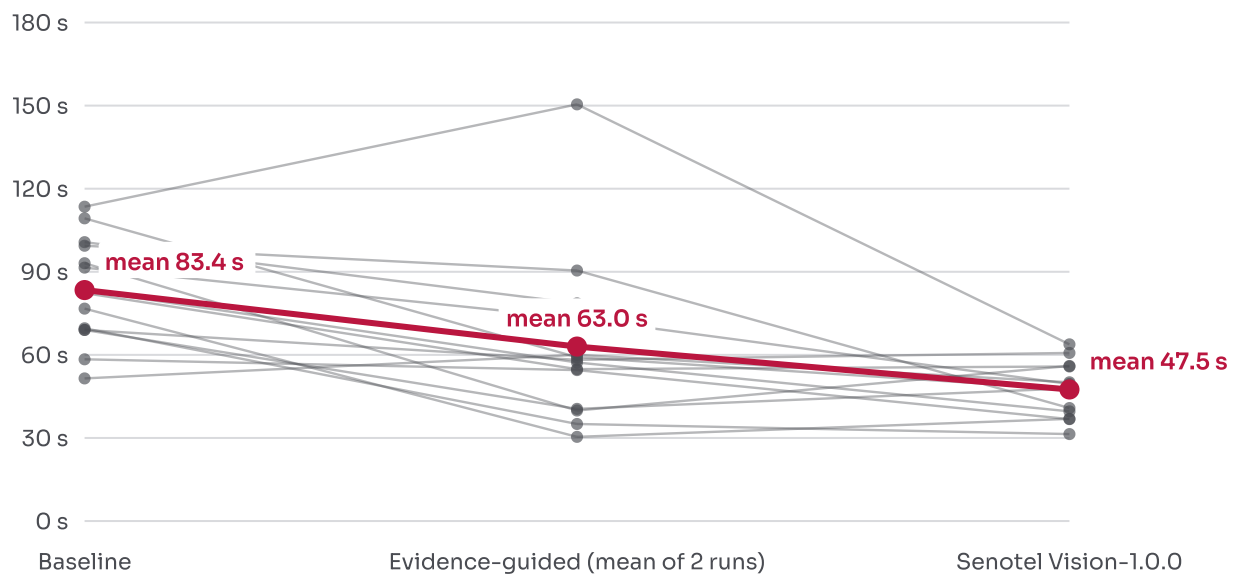


Figure 1. Edit length per recording. Grey: one line per recording. Crimson: the means. The evidence-guided editor (mean of its two runs) cut most edits shorter; Senotel Vision-1.0 holds a much narrower range of lengths.

Averaged over its two runs, the evidence-guided editor made edits 20.3 seconds shorter than the baseline ($p = .010$, Holm-adjusted = .040), passed 0.46 more checks ($p = .014$, adjusted = .042) and scored 9.5 points higher on the proxy ($p = .020$, adjusted = .042). The longest lull did not change ($p = .964$). The two runs differ: run 1 improved on all four measures after adjustment, run 2 only on length and proxy and only before adjustment, and run 2 produced the longest edit in the study (a 152-second tutorial). The first payoff moved earlier, from 24.6 s to 13.8 s and 16.1 s.

The post-processor alone, applied to the baseline's unchanged plans, shortened edits by 3.3 seconds ($p = .004$), raised the proxy by 3.3 ($p = .008$), brought the first word to within 0.02 s on average, and made every edit pass H1, H1b and R10 (Table 4). Most of the change in length, however, comes from the editor's own choices: the evidence changed what it decided to keep.

System	H1	H1b	H5	RQ2	R1	R10
Baseline	13/14	13/14	14/14	3/14	2/14	12/14
Baseline + post-processor	14/14	14/14	14/14	3/14	5/14	14/14
Evidence-guided, run 1	14/14	14/14	14/14	7/14	3/14	14/14
Evidence-guided, run 2	13/14	14/14	14/14	3/14	3/14	14/14
Senotel Vision-1.0.0	13/13	13/13	13/13	7/13	6/13	13/13
Senotel Vision-1.0.1	13/13	13/13	13/13	7/13	6/13	13/13

Table 4. Recordings passing each check. For every language-model configuration, the two checks failed most often are R1 (a long stretch without anything new) and RQ2 (an early payoff).

Comparison	Measure	n	Mean paired difference	Sign-flip test	Holm-adjusted
Evidence-guided (mean of runs) – baseline	Length, s	14	-20.34	p = .010	p = .040
	Proxy	14	+9.50	p = .020	p = .042
	Checks	14	+0.46	p = .014	p = .042
	Longest lull, s	14	+1.25	p = .964	p = .964
Evidence-guided run 1 – baseline	Length, s	14	-22.15	p = .009	p = .027
	Proxy	14	+9.14	p = .021	p = .042
	Checks	14	+0.64	p = .004	p = .016
	Longest lull, s	14	-2.08	p = .035	p = .042
Evidence-guided run 2 – baseline	Length, s	14	-18.54	p = .036	p = .144
	Proxy	14	+9.86	p = .036	p = .144
	Checks	14	+0.29	p = .313	p = .626
	Longest lull, s	14	+4.57	p = .946	p = .946
Baseline + post-processor – baseline	Length, s	14	-3.26	p = .004	p = .016
	Proxy	14	+3.29	p = .008	p = .024
	Checks	14	+0.50	p = .031	p = .062
	Longest lull, s	14	-0.93	p = .031	p = .062
Vision-1.0.0 – evidence-guided run 2	Length, s	13	-13.61	p = .125	p = .273
	Proxy	13	+12.77	p = .001	p = .004
	Checks	13	+0.54	p = .148	p = .273
	Longest lull, s	13	-0.99	p = .091	p = .273

Table 5. Paired comparisons. The mean paired difference is the first system minus the second; for length and lull, negative values are shorter.

6.2 What a blind judge preferred

The judge returned a verdict for one side in every one of its 81 individual judgements and gave the same verdict in both orders for 32 of 40 pairs (80%); the remainder count as ties. It chose the edit shown first in 51% of its verdicts, so position did not drive it, and in the 32 pairs it decided the shorter edit won 14 times, so it did not simply reward brevity, although some of its stated reasons mention it.

Pair (first vs second)	Recordings	First wins	Second wins	Ties	of which split	Sign test
Baseline vs Evidence-guided, run 2	14	3	7	4	4	$p = .344$
Evidence-guided, run 2 vs Senotel Vision-1.0.0	13	11	1	1	1	$p = .006$
Evidence-guided, run 2 vs Senotel Vision-1.0.1	13	9	1	3	3	$p = .021$

Table 6. Blind pairwise judgements. A win requires the same verdict in both orders; "split" counts pairs on which the two orders disagreed, scored as ties. Exact two-sided sign test on decided pairs.

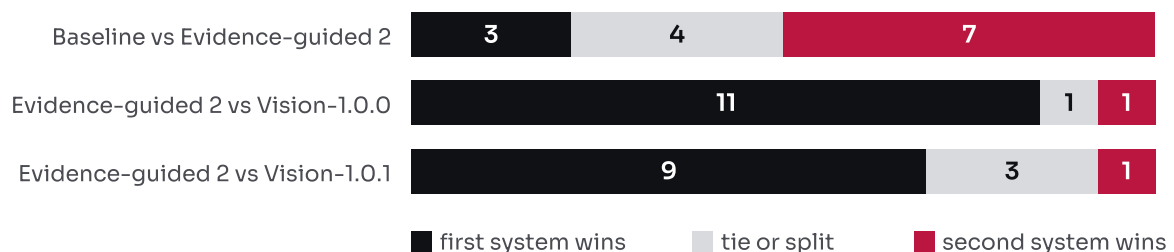


Figure 2. Judge outcomes per pair of systems, over the recordings both systems planned.

Evidence-guided against baseline. Comparing the two complete systems on all fourteen recordings, the judge preferred the evidence-guided edit on 7, the baseline on 3, and tied 4 ($p = .344$). The direction agrees with the checks; the sample does not allow a stronger statement. The judge's reasons for the wins named tighter pacing and the removal of filler. Of the three losses, one was the 152-second tutorial produced by run 2, where the judge preferred the baseline's 114-second cut; one was a personal story, where it preferred the baseline's "clearer, more uplifting narrative"; and one was a science explainer, where it preferred the baseline's question-led opening, a preference that runs against randomized evidence on question headlines [17].

Large editor against Senotel Vision-1.0. The judge preferred the evidence-guided editor on 11 of 13 recordings and Vision-1.0.0 on 1 ($p = .006$). Against Vision-1.0.1 the count was 9 to 1 with 3 ties ($p = .021$): the correction in Section 6.4 narrowed the margin without changing the conclusion. Vision's single win was a science explainer. Reading the losing edits shows where Vision went wrong. On two personal stories it withheld music and cutaways under its rule for sensitive footage, where the large editor scored both with piano, and the judge named the music and the cutaways in its reasons; on a guitar lesson it laid music over the instrument being taught; and most often the judge found the large editor's edit clearer and more coherent.

6.3 How far Senotel Vision-1.0 got

Vision-1.0 passed the checks at least as often as the large editor (5.00 of 6 against 4.36; all six on 4 of 13 recordings) and scored higher on the proxy (58.5 against 42.9). Neither number shows that it edits well: it selects its edit by maximising that same proxy among six candidates (Lesson 14), and it was taught the same principles the checks test. What it does not do is choose the material the large editor chooses.

Compared with evidence-guided run 2	Shared material (F1)	Same opening	Dev F1	Dev opening	Test F1	Test opening
Evidence-guided run 1 (same editor, second sample)	0.59	10/14	0.55	5/7	0.68	4/6
Baseline	0.59	8/14	0.62	4/7	0.54	3/6
Senotel Vision-1.0.0	0.33	3/13	0.35	3/7	0.31	0/6
Senotel Vision-1.0.1	0.33	3/13	0.35	3/7	0.31	0/6

Table 7. Agreement with the evidence-guided editor's second run. The first row is the ceiling set by sampling: two runs of the same editor on the same footage.

Two runs of the same large editor share 0.59 of their material and open on the same moment in 10 of 14 recordings. Vision-1.0 shares 0.33 and opens on the same moment in 3 of 13; on the held-out recordings, 0 of 6. The baseline agrees with the evidence-guided editor about as much as the latter agrees with itself (0.59), suggesting that language models converge on the material of a story even when they cut it differently; Vision-1.0 does not reach that material.

Training did not help. Fitted to the large editor's choices, the hook model ranked the editor's opener first in 1/13 recordings and within its top three in 3/13; the evidence-derived priors alone placed it in the top three in 6/13 (Table 8). Across prior strengths, held-out ranking improved steadily as the fitted weights were pulled back toward the priors and never surpassed them (Table 9). The keep model's held-out discrimination (AUC) moved from 0.52 to 0.58 against one run and from 0.58 to 0.56 against the other: no reliable gain. Vision-1.0 therefore retains the evidence-derived priors for both skills, and only the length calibration.

Teacher	Weights	Opener ranked 1st	Opener in top 3	Keep AUC	Whole-edit F1	Same opening (share)
Evidence-guided run 1	Evidence-derived priors	0/13	5/13	0.58	0.37	0.23
	Fitted (leave-one-video-out)	0/13	3/13	0.56	0.26	0.08
Evidence-guided run 2	Evidence-derived priors	1/13	6/13	0.52	0.33	0.23
	Fitted (leave-one-video-out)	1/13	3/13	0.58	0.29	0.00

Table 8. Learnability, leave-one-video-out, with each run of the evidence-guided editor as teacher. "Same opening (share)" is the fraction of the thirteen recordings on which Vision's whole edit opens on the teacher's moment.

Weights	Prior strength	Opener ranked 1st	Opener in top 3	Keep AUC
Evidence-derived priors	no fitting	1/13	6/13	0.52
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 0.1, \lambda_{\text{keep}} = 0.3$	0/13	1/13	0.58
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 0.6, \lambda_{\text{keep}} = 0.3$	1/13	3/13	0.58
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 3, \lambda_{\text{keep}} = 0.3$	1/13	4/13	0.58
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 10, \lambda_{\text{keep}} = 0.3$	1/13	5/13	0.58
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 0.6, \lambda_{\text{keep}} = 0.05$	1/13	3/13	0.58
Fitted (leave-one-video-out)	$\lambda_{\text{hook}} = 0.6, \lambda_{\text{keep}} = 3$	1/13	3/13	0.56

Table 9. Strength of the prior (λ , the weight of the pull toward the evidence-derived weights) against held-out accuracy, with run 2 as teacher.

6.4 A defect found in the held-out evaluation

Reading Vision-1.0.0's held-out edits revealed a defect the metrics did not: its opening text was sometimes a fragment, because it cut a long clause down to its densest eight words ("Usually steam on high for around six") or joined two recognised lines ("Feel that beat To celebrate your birthday"). Version 1.0.1 prefers a whole clause of up to ten words, then the clause without its filler, article and softeners, and only then the clause up to a phrase boundary; it never crosses two recognised lines and never ends on a number separated from its range. The change affects only the opening text, not which material is chosen, so lengths and agreement are unchanged. Because it was made after seeing held-out output, results for 1.0.1 are reported separately and are not held-out evidence. Appendix B lists the opening text of every system on every recording.

7. Discussion

What the integration shows. Giving a language-model editor the evidence as an ordered reasoning brief changed what it kept: edits became shorter, the first payoff arrived earlier and more checks passed. A deterministic post-processor that enforces the mechanical principles improves any plan, including one made without the evidence. The blind judge points the same way, seven recordings to three, but fourteen recordings cannot confirm the direction, and the judge is a language model reading text, not an audience.

Why the proxy cannot settle quality. The proxy was written from the same findings that the editor was given and that Vision was taught, and Vision selects its edit by maximising it. A system that follows the principles will score well on a model of those principles whether or not viewers prefer its edit. This is why we never use the proxy here as evidence that one system edits better than another, why we ran the blind judge, and why calibrating the proxy against real audience retention is the first item of future work.

Why Senotel Vision-1.0 falls short. The gap lies in one decision more than any other: which moment is the story. The large editor reads meaning; Vision reads words, timing, delivery and what a recording keeps returning to. That is often enough to find the subject but rarely enough to find the moment: the large editor's opener was among Vision's three strongest hook candidates in 6/13 recordings but ranked first in only 1/13. In the one recording where it ranked first, Vision's simulation step still chose another hook, because the proxy preferred that edit. Thirteen decisions by a teacher cannot teach semantics to a linear model over hand-made features; Table 9 shows the data pulling the weights away from the evidence rather than toward the teacher. Closing the gap requires far more teacher decisions and a representation of meaning, such as a compact sentence-embedding model, and preferably both.

A note on the teacher. An editor that a student is trained to imitate should itself be shown to be good by people, not only by proxies. The evidence-guided editor improved on its baseline, but in informal use its edits still showed weaknesses that our measures do not capture: lines cut where a speaker's thought was not finished, context too thin for a newcomer to follow, and cutaways that illustrated a phrase literally rather than its meaning. A student can be no better than the examples it learns from, so the choice of teacher deserves its own evaluation with human viewers before any distillation at scale.

8. Limitations

- **No real retention data.** No result here measures real viewers. The proxy is uncalibrated, and the judge is a language model reading text, not a viewer watching video: it cannot see faces, motion or timing, and it may share biases with the editor.
- **Small sample.** Fourteen recordings, each run once or twice. Sign-flip tests on fourteen pairs detect only large effects, and the two runs of the same editor differ visibly. Genre labels were assigned by us, and several genres have one recording.

- **One judge.** The second judge could not be run. A single judge, even order-balanced, carries its own preferences, and its reasons sometimes mention brevity despite the instruction.
- **Design choices made after looking.** Several of Vision's features were added after reading the large editor's first hooks on all fourteen recordings, before the development and test sets were fixed. The split protects the length calibration and the fitted weights, not those design choices. The 1.0.1 correction was made after the held-out evaluation and is reported separately.
- **Transcription errors.** All systems saw the same automatic transcripts from the smallest Whisper model, including their mistakes, which affect opening text most.
- **The editor model.** The language-model editor was a fast, low-cost model whose exact version could change between runs; stronger models may behave differently.

9. Conclusion and future work

Evidence about attention can be given to an automated editor in a form it follows, and doing so moved its edits in the direction the evidence recommends. A small model taught the same evidence follows the rules but does not yet choose the right moments; that requires a sense of meaning which rules over words do not provide. Our next steps are, in order: to choose the teacher by human evaluation on creators' own footage; to collect a much larger set of the chosen teacher's decisions on openly licensed recordings; to give Senotel Vision a compact representation of meaning that runs on the device; and to calibrate the value-of-continuing proxy against real audience-retention curves shared with consent, so that future comparisons can rest on viewers rather than on models.

Ethics, licensing and disclosure. All footage is in the public domain or under Creative Commons licences permitting reuse with attribution; authors and licences are listed in Appendix A. The footage was used for evaluation only and is not redistributed in edited form. No human participants were involved. One recording is a personal account that includes bullying and suicidal thoughts; it is referred to here only by its licence information, and the editors studied are designed to treat such footage with restraint (no comic sounds, no music by default). The research, code, experiments and this report were produced with the assistance of an AI system (Claude, by Anthropic) under the direction of Senotel. The editor in the experiments used Google's Gemini models [21]; the judge used OpenAI's open-weight gpt-oss-120b [39] served by Groq [40]. Senotel develops automated video-editing tools; no external funding was received.

References

1. Senotel Research. (2026). *What keeps viewers watching short-form video? An evidence-graded synthesis of hooks, structure, pacing and endings*. Senotel research report.
2. Berthouzoz, F., Li, W., & Agrawala, M. (2012). Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics*, 31(4) (SIGGRAPH 2012).
3. Truong, A., Berthouzoz, F., Li, W., & Agrawala, M. (2016). QuickCut: An interactive tool for editing narrated video. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '16)*. ACM.
4. Leake, M., Davis, A., Truong, A., & Agrawala, M. (2017). Computational video editing for dialogue-driven scenes. *ACM Transactions on Graphics*, 36(4), 130:1–130:14.
5. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D. B., Genova, K., Jin, Z., Theobalt, C., & Agrawala, M. (2019). Text-based editing of talking-head video. *ACM Transactions on Graphics*, 38(4), 68. <https://doi.org/10.1145/3306346.3323028>
6. Wang, B., Li, Y., Lv, Z., Xia, H., Xu, Y., & Sodhi, R. (2024). LAVE: LLM-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. ACM. <https://doi.org/10.1145/3640543.3645143>
7. Wu, S., Rizoio, M.-A., & Xie, L. (2018). Beyond views: Measuring and predicting engagement in online videos. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 12(1). arXiv:1709.02541
8. Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014). Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the First ACM Conference on Learning @ Scale* (pp. 31–40). ACM.

9. Krishnan, S. S., & Sitaraman, R. K. (2012). Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *Proceedings of the ACM Internet Measurement Conference (IMC '12)*. ACM.
10. Quan, Y., Ding, J., Gao, C., Li, N., Yi, L., Jin, D., & Li, Y. (2023). Alleviating video-length effect for micro-video recommendation. *ACM Transactions on Information Systems*. arXiv:2308.14276
11. Zhan, R., Pei, C., Su, Q., Wen, J., Wang, X., Mu, G., Zheng, D., & Jiang, P. (2022). Deconfounding duration bias in watch-time prediction for video recommendation. In *Proceedings of KDD 2022*. arXiv:2206.06003
12. Zhao, H., Cai, G., Zhu, J., Dong, Z., Xu, J., & Wen, J.-R. (2024). Counteracting duration bias in video recommendation via counterfactual watch time. In *Proceedings of KDD 2024*. arXiv:2406.07932
13. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*. arXiv:2306.05685
14. YouTube. (n.d.). Shorts analytics: Viewed versus swiped away. YouTube Studio Help.
15. Teixeira, T., Wedel, M., & Pieters, R. (2010). Moment-to-moment optimal branding in TV commercials: Preventing avoidance by pulsing. *Marketing Science*, 29(5), 783–804. <https://doi.org/10.1287/mksc.1100.0567>
16. Google, & Kantar. (2021). The short and the long of ABCDs effectiveness. Think with Google.
17. Fang, D., & Wheeler, S. C. (2026). Titles framed as questions reduce reader engagement. *Journal of Consumer Psychology*. Advance online publication. <https://doi.org/10.1002/jcpy.70031>
18. YouTube. (n.d.). Measure key moments for audience retention. YouTube Help. support.google.com/youtube/answer/9314415
19. Woltman Elpers, J. L. C. M., Wedel, M., & Pieters, R. G. M. (2003). Why do consumers stop viewing television commercials? Two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research*, 40(4), 437–453.
20. Baumgartner, H., Sujan, M., & Padgett, D. (1997). Patterns of affective reactions to advertisements: The integration of moment-to-moment responses into overall judgments. *Journal of Marketing Research*, 34(2), 219–232.
21. Google. (2026). *Gemini API: Models* [Documentation; Gemini Flash-family models, accessed October 2026]. ai.google.dev/gemini-api/docs/models
22. Smith, T. J., & Henderson, J. M. (2008). Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of Eye Movement Research*, 2(2), 1–17. <https://doi.org/10.16910/jemr.2.2.6>
23. Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & Psychophysics*, 65(7), 999–1010.
24. Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, 35(8), 1489–1517. <https://doi.org/10.1111/j.1551-6709.2011.01202.x>
25. Teixeira, T., Wedel, M., & Pieters, R. (2012). Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research*, 49(2), 144–159. <https://doi.org/10.1509/jmr.10.0207>
26. Berger, J., Moe, W. W., & Schweidel, D. A. (2023). What holds attention? Linguistic drivers of engagement. *Journal of Marketing*, 87(5), 793–809. <https://doi.org/10.1177/00222429231152880>
27. Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10), 1295–1306.
28. Molyneux, L., & Coddington, M. (2020). Aggregation, clickbait and their effect on perceptions of journalistic credibility and quality. *Journalism Practice*, 14(4), 429–446. <https://doi.org/10.1080/17512786.2019.1628658>
29. Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1), 75–98. <https://doi.org/10.1037/0033-2909.116.1.75>
30. Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, 50(1), 46–70. <https://doi.org/10.1111/j.1460-2466.2000.tb02833.x>
31. van Laer, T., de Ruyter, K., Visconti, L. M., & Wetzels, M. (2014). The extended transportation-imagery model: A meta-analysis of the antecedents and consequences of consumers' narrative transportation. *Journal of Consumer Research*, 40(5), 797–817. <https://doi.org/10.1086/673383>
32. Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. *Projections*, 2(1), 1–26.
33. Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review*, 23, 1–24.
34. Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review*, 7(3), 216–237. <https://doi.org/10.1016/j.edurev.2012.05.003>

35. Sundararajan, N. K., & Adesope, O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review*, 32(3), 707–734. <https://doi.org/10.1007/s10648-020-09522-4>
36. Husain, G., Thompson, W. F., & Schellenberg, E. G. (2002). Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music Perception*, 20(2), 151–171.
37. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 28492–28518.
38. Hugging Face. (2024). *Transformers.js* (Version 2.17.2) [Software]. github.com/huggingface/transformers.js
39. OpenAI. (2025). *gpt-oss-120b & gpt-oss-20b model card*. [arXiv:2508.10925](https://arxiv.org/abs/2508.10925)
40. Groq. (2026). *GroqCloud API documentation* [Hosted inference service, accessed October 2026]. console.groq.com/docs

Appendix A. Footage

Key	Label	Length	Split	Licence	Author as credited
smoker	story	2:50	dev	Public domain	Centers for Disease Control and Prevention
better	story	4:21	test	CC BY-SA 4.0	oetimmy
moms	comedy	3:05	dev	CC BY 3.0	Maz Jobrani
lycett	comedy	5:07	test	CC BY-SA 4.0	Stuart Barrett
blackhole	explainer	1:33	dev	Public domain	NASA Langley Research Center
potty	explainer	7:49	test	Public domain	NASA
redmoon	interview	7:58	dev	Public domain	NASA's Scientific Visualization Studio - eMITS/Michelle Handleman, NASA Interns/Alexa Figueroa, NASA/GSFC/Angel Kumari, eMITS/Gloriana Wills, ADNET Systems, Inc./Aaron E. Lepsch, Telophase/Maria-Jose Vinas Garcia, NASA Intern/Charles Connell
fish	tutorial	4:57	test	CC BY 3.0	SETHLUI.com
palmmute	tutorial	2:07	dev	CC BY 4.0	Blackstar Amplification
telegram	vlog	3:45	test	CC BY 3.0	Gothamist
tyson	talk	2:10	dev	Public domain	United States Department of Education
walabot	review	1:23	test	CC BY 3.0	Walabot DIY
smell	explainer	2:32	dev	Public domain	NASA Johnson
punch	tutorial	3:17	neither	CC BY 3.0	JustBeCreative

Table A1. The fourteen recordings, linked to their Wikimedia Commons pages, with licence and author as credited there. "Label" is the genre assigned when the footage was chosen; the systems detect genre themselves.

Appendix B. Opening text by system

Recording	Baseline	Evidence-guided 2	Vision-1.0.0	Vision-1.0.1
smoker	"The doctor said: 'Go to hospice until you die.'"	"Go to hospice until you die."	"Rest of your days there until you die."	"You need to go to hospice"
better	"I'm 58. I could have died at 17."	"I wanted to kill him."	"I could have killed myself at 17"	"I could have killed myself at 17"
moms	"Go back to your country"	"Why don't you go back?"	"Why don't you go back to your country?"	"Why don't you go back to your country?"
lycett	"Taking on homophobes... 🏳️"	"YOU'RE ALL AT RISK"	"Famously don't like public displays of emotion"	"London people famously was famously don't like getting involved"
blackhole	"They Aren't Actually Holes"	"Black holes aren't actually holes"	"A stellar mass black hole is born"	"That's star that's larger than our own Sun"
potty	"How do astronauts pee in space?"	"Living inside zero-g"	"I actually had to remind myself"	"I actually had to remind myself"
redmoon	"Every sunrise and sunset at once"	"The moon remembers everything."	"Once every six months it actually passes through"	"About once every six months it passes"

Recording	Baseline	Evidence-guided 2	Vision-1.0.0	Vision-1.0.1
fish	"Hate fishy smell? Try this!"	"NO MORE FISHY SMELL"	"Usually steam on high for around six"	"Usually steam on high for around six to eight minutes"
palmmute	"The #1 Guitar Technique"	"Palm muting is essential for every rock player"	"Dead and if it's wrong you'll hear"	"You'll know if you're this right or wrong"
telegram	"HIP HOP TWERKING GORILLA?!"	"I am the hip hop twerking gorilla!"	"Feel that beat To celebrate your birthday"	"It's always crazy"
tyson	"What makes a truly great teacher?"	"Two teachers who mattered"	"Maybe the great tragedy is not all teachers"	"Maybe the great tragedy is not all teachers"
walabot	"Detect Pipes & Wires Like a Pro"	"Tracking curved pipes behind walls"	"Won't show two images on the screen simultaneously"	"Wall about won't show two images on the screen simultaneously"
smell	"What Space Actually Smells Like"	"The unique smell of space"	"There's one smell up here that's really unique"	"There's one smell up here that's really unique"
punch	"DIY Cute Punch Needle Patch 🧵"	"Make your own punch needle badge"	no plan	no plan

Table B1. The text each system placed on screen at the start. Transcription errors are reproduced as the systems saw them.

Appendix C. Per-recording results

Recording	Split	Baseline	Evidence-guided 1	Evidence-guided 2	Vision-1.0.0	Vision-1.0.1
smoker	dev	109.3 / 43 / 5	60.8 / 54 / 5	57.5 / 59 / 5	47.6 / 62 / 5	47.6 / 59 / 5
better	test	51.5 / 57 / 5	60.6 / 45 / 5	59.4 / 39 / 4	50.1 / 64 / 6	50.1 / 64 / 6
moms	dev	82.3 / 39 / 4	68.3 / 44 / 5	41.4 / 59 / 5	55.8 / 63 / 6	55.8 / 63 / 6
lycett	test	91.5 / 45 / 6	72.9 / 49 / 6	73.5 / 42 / 6	49.5 / 50 / 5	49.5 / 50 / 5
blackhole	dev	83.3 / 31 / 4	78.7 / 30 / 5	35.8 / 54 / 4	39.6 / 51 / 4	39.6 / 52 / 4
potty	test	58.4 / 41 / 4	57.5 / 48 / 5	51.5 / 57 / 6	36.8 / 67 / 5	36.8 / 67 / 5
redmoon	dev	69.0 / 36 / 4	42.0 / 48 / 4	74.4 / 27 / 4	60.7 / 58 / 5	60.7 / 58 / 5
fish	test	113.5 / 24 / 4	148.5 / 11 / 5	152.4 / 14 / 4	63.8 / 47 / 4	63.8 / 47 / 4
palmmute	dev	93.1 / 26 / 3	37.5 / 59 / 4	42.3 / 53 / 4	56.0 / 55 / 6	56.0 / 55 / 6
telegram	test	69.0 / 32 / 4	37.0 / 56 / 4	44.0 / 55 / 4	47.9 / 56 / 5	47.9 / 56 / 5
tyson	dev	99.4 / 17 / 4	81.3 / 28 / 5	99.6 / 24 / 4	40.8 / 58 / 4	40.8 / 58 / 4
walabot	test	69.6 / 22 / 3	37.5 / 43 / 4	32.6 / 53 / 4	31.4 / 60 / 4	31.4 / 60 / 4
smell	dev	76.7 / 43 / 4	31.4 / 62 / 5	29.4 / 58 / 4	36.9 / 69 / 6	36.9 / 69 / 6
punch	neither	100.7 / 6 / 3	43.2 / 13 / 4	114.0 / 6 / 3	no plan	no plan

Table C1. Length (s) / proxy score / checks passed (of 6), per recording and system. "No plan": Vision declines footage without speech (Section 5).

Recording	Baseline vs Evidence-guided 2	Evidence-guided 2 vs Vision-1.0.0	Evidence-guided 2 vs Vision-1.0.1
smoker	Evidence-guided 2	Evidence-guided 2	split
better	Baseline	Evidence-guided 2	Evidence-guided 2
moms	Evidence-guided 2	Evidence-guided 2	Evidence-guided 2

Recording	Baseline vs Evidence-guided 2	Evidence-guided 2 vs Vision-1.0.0	Evidence-guided 2 vs Vision-1.0.1
lycett	split	Evidence-guided 2	Evidence-guided 2
blackhole	split	Vision-1.0.0	Vision-1.0.1
potty	Baseline	Evidence-guided 2	Evidence-guided 2
redmoon	Evidence-guided 2	Evidence-guided 2	Evidence-guided 2
fish	Baseline	split	split
palmmute	Evidence-guided 2	Evidence-guided 2	Evidence-guided 2
telegram	Evidence-guided 2	Evidence-guided 2	Evidence-guided 2
tyson	split	Evidence-guided 2	Evidence-guided 2
walabot	Evidence-guided 2	Evidence-guided 2	Evidence-guided 2
smell	split	Evidence-guided 2	split
punch	Evidence-guided 2	not planned	not planned

Table C2. The judge's verdict per recording ("split": the two orders disagreed).

Appendix D. The judge's instructions

Two editors each cut one short vertical video (for TikTok, Instagram Reels or YouTube Shorts) from the SAME raw footage. You cannot see the videos, so each is described as a viewer would experience it: the text on screen at the start, every spoken line in the order it is heard, the cutaway pictures, the music and the length. The spoken lines come from automatic transcription, so expect some misheard words in both.

Picture a typical viewer who comes across each video while scrolling. Which edit are they more likely to keep watching to the end, and be glad they watched?

Judge the editing, not the footage: both editors had exactly the same material. Do not prefer an edit just because it is longer or shorter. If neither is clearly better, answer tie.

EDIT A

[the first edit]

EDIT B

[the second edit]

Reply with JSON only, nothing else, in this shape: {"winner": "A" or "B" or "tie", "reason": "one sentence"}

Each edit was given as: Length; Text on screen at the start; What is said, in order (numbered lines); Cutaway pictures; Music; Sound effects (a count).

Reproduced verbatim except for its one-line title, which names an internal service and is omitted.