

Packaging a Podcast Network with AI

What we built for a podcast network, what it produced, and what it does not do.

Grant Katanick, Katanick Creative Solutions August 2026

Summary

Between June and August 2026 we replaced a manual thumbnail and title process with a documented system that turns an episode transcript into publish-ready packaging. It ran across three shows, produced 535 images over 82 dated video folders, across 66 delivery dates between 10 June and 26 August. Every episode that arrived in time to work on was packaged.

The system is not a model. The model is the easy part and it is interchangeable. What took three months to build is everything around it: the rules the process follows, the reference data it reads, the three points where a person signs off, and the way a correction from someone on the team becomes a permanent change rather than a note that gets forgotten.

This paper describes that layer, because it is the part that transfers.

1. The problem we started with

The old process worked, and it did not scale.

Titles and thumbnails were produced by hand across separate AI chats, one per show, each trained by a different person with different instructions. That has three costs that only appear at volume:

- **The rules lived in people's heads.** A guest who did not want their face used was remembered by whoever remembered. There was no place to write it down that the next run would read.
- **Quality moved with whoever did the work.** The same show could get a strong title one week and a vague one the next, and nobody could say why.
- **Every improvement was lost.** When someone worked out that a title should name the person rather than say "he says", that lesson stayed in one chat window.

Volume made all three worse. A network publishing daily cannot carry rules in memory.

2. What the system does

A transcript goes in. What comes out is three titles, three overlay texts, three image prompts with the correct speaker reference resolved, and three finished images at 2752x1536, uploaded for review with the board updated.

Five stages, three of them stopping for a person.

One. Read the episode. The system reads the full transcript and proposes the three strongest hooks. It matches the video's main focus rather than the most vivid tangent, which is a rule we added after a thumbnail promoted a thirty-second aside.

Two. Write titles and overlays together, and stop. These are never split. A title and the overlay text that sits on the image are one decision, because they have to work as a pair. A person approves both before anything is generated.

Three. Resolve the reference. Every speaker maps to a verified photograph through an alias map that never trusts the transcript's spelling of a name. Automatic transcripts get names wrong constantly, and a wrong face is the most expensive error this system can make.

Four. Generate and measure. Images are produced, then measured from the file rather than judged by eye. Overlay height as a percentage of frame, right-edge position, spelling character by character.

Five. Deliver. Upload, update the board, set the status.

3. The part that actually took three months

Anyone can write the five stages above in an afternoon. The value is in the rules underneath them, and every one of them cost a real mistake to learn.

Titles

Attribute by name. Never "He Says" or "They Claim". Use the person's name. A title that hides who said something reads as evasive and performs worse.

The "what X?" test. Read each noun in the title and ask "what X?" If the title cannot answer, it is vague. "The arrests", "the ships", "the discovery" are the usual offenders. This single test removed most of the weak titles we were producing.

Every option a different shape. Three options are worthless if all three open the same way. No two may share an opening construction, which forces genuinely different angles rather than three rewordings of one idea.

Overlay text

Two tests run on every overlay before a person sees it, and both came from thumbnails that shipped and should not have.

The airtime test. Find the phrase in the transcript and count how often it appears. One overlay used a phrase that occurred once in 162 lines, and a viewer called the thumbnail clickbait. Flagging the risk in a notes file is not a fix. That set was flagged and shipped anyway, so the test is now a gate rather than a warning.

The say-it-alone test. Cover the image. Does the overlay name the subject? One overlay read as a single place-name, which was true and told the viewer nothing about the episode.

A measured rule that changed the output

Overlay height is set by the empty band the composition leaves and by the number of lines put into it. The renderer fits the whole text block into that band, then divides the band between the lines. Measured across 32 shipped overlays, a single line averages 22.9 percent of frame height and never came back below 19.1 percent. Two lines average 17.1 percent and drop as low as 11.5 percent when portraits or insets crowd the band.

Character count is not the lever. Five overlays with a nine-character line rendered between 21.5 and 24.2 percent. A sixth, with the same nine characters broken across two lines under five portraits, rendered at 12.2 percent. Twelve overlays put eight or more characters on one line and none of them came back under 19.1 percent.

That means you decide the line count and reserve the empty band when you **write** the overlay, not after you generate it. Before we measured this, the response to a small overlay was to ask the renderer for a bigger one, which does not work: past roughly 35 to 40 percent the renderer clamps rather than scales.

This rule was itself corrected. An earlier version of it counted characters in the longest line. Re-measuring 32 shipped overlays showed character count barely predicts height and line count predicts it strongly. The wrong version pointed the fix in the wrong direction, at the spelling rather than at the composition. The measurements are in the exhibit that accompanies this paper.

Legal guardrails

Four rules apply to every show and are not relaxable. No named private person shown as guilty. No fabricated quotes. No real logos or seals. No health-cure claims. Contested claims are attributed, hedged, or asked as a question, never stated as fact.

There is a subtler rule underneath them. Naming a forbidden thing inside a "do not" instruction can produce exactly the thing you named, or get the request refused outright. Describing the correct result instead gets the same outcome without tripping the filter. That distinction sounds small until you have watched a run fail on it.

The reference map

The most valuable file in the system is not a prompt. It is a map of 59 people that records what the automatic transcripts get wrong about each name, which photograph to use, the written reason for that

choice, and which photographs have already failed and must never be used again.

That last column took two months of production to accumulate. It is the difference between a workflow and a demonstration, and it is the piece a new tool cannot start with.

4. Where the humans are, and why there

Three stops, chosen deliberately.

1. **Hooks.** A person confirms which three topics are worth packaging.
2. **Titles and overlays.** A person approves the exact words before any image is made.
3. **Winners.** A person picks which of A, B and C ships.

They sit at the moments that carry real risk. A process that asks for approval at every step gets bypassed inside a month, and a process that asks for none produces the failure everybody fears.

The correction path matters as much as the gates. When the team flagged that an episode had shipped without the presenter's face on any thumbnail, that correction went into the configuration rather than into a message. A face quota now applies to every set. The rules are better today than the day they were written, and they improve without the person who wrote them.

5. What this record proves, and what it does not

It proves the system runs. 82 dated folders between 10 June and 26 August 2026, across three shows, 535 images delivered, with a documented process that a different person can follow.

It does not prove a business outcome, and I am not going to invent one. Nobody timed the old process before we replaced it, so any percentage I quoted for time saved would be a guess dressed up as a result. If a return-on-investment number matters, the measurement has to be built in from day one and read at the end. That is a decision to make at the start of the next system, not a number to reconstruct from this one.

Two steps remain manual by choice. The image download and the final logo pass are done by hand, because automating them would have cost more than it saved. Choosing not to automate something is a design decision, not an unfinished edge.

6. What transfers to the next thing

The specific rules here are about thumbnails. The shape is not.

Any recurring process where a model produces something a customer sees needs the same four things, and they are the four that the failed pilots skip.

1. **Reference data that is written down**, including what has already failed.
2. **Rules stated positively**, in language anyone on the team can check output against.
3. **Human gates at the two or three moments that carry risk**, and nowhere else.
4. **A correction path** that turns feedback into a permanent change rather than a repeated fix.

The model was never the hard part. The hard parts were resolving identity from dirty input, keeping a person at the moments that matter, writing every failure down in a form the next run can read, and giving people who are not the builder a way to correct the system rather than only its output.

If a pilot stalled, it almost certainly stalled on one of those four.

Prepared for a podcast network, August 2026.