SANSKRIT AND AI / PART 04 / STUDY NOTES

Sanskrit and AI, Part 4: The Hidden Layers

Study notes by Sanskrit AI Universe, based on this video

Video by Dr P.R.K. Varma, Founding Editor · DrVARMA PRK · 15 Nov 2026

Not the creator's own slides

About these notes

— Official title: Sanskrit and AI Part 4 :The Hidden Layers | Dr.P.R.K.Varma.,M.D.,D.M.

— Speaker and channel: Dr P.R.K. Varma, Founding Editor · DrVARMA PRK

— Published 15 Nov 2026 · Running time 24 min 31 s · https://www.youtube.com/watch?v=YcaTkGbZw0s

— These pages are study notes drawn from the video's own title, description and captions, plus the journal's already published Watch note. They are not a transcript and not a substitute for watching the film.

— The journal treats videos as leads, not as sources you should cite on their own. Check the papers and editions behind a claim.

— Caption source: English auto-generated subtitles. Wording may be imperfect. Distinctive phrases follow the captions closely.

This part in brief

- Everything the model has learned is in there and almost none of it can be read.
- So this part is about two things that work well and are hard to see into.
- What does it mean to say that the model has learned something?
- During training, each example nudges billions of weights by a tiny amount.
- When training ends, there is no list of what was learned.

Chapter one, what is learned

— Three questions, rule or memory, same or different, near or far.

Problem one, memorize or generalize

— Researchers showed that a network which learns real images can also learn pure noise, labels assigned at random.

— With the right prompt, they can be made to repeat passages of their training text word for word.

— And sometimes a network memorizes first and finds the rule much later, long after its training score stopped improving.

— What is needed, Patañjali says, is a lakshana, a rule with a general part and a special part by which, with little effort, one reaches great floods of words.

Problem two, same thing, different clothes

— To a language model, those can be three unrelated strings of numbers.

— In one study, changing only the layout of a prompt, a colon, a space, a capital letter moved a model's accuracy by tens of points.

— Pāṇini defines sameness in the first quarter of his first chapter.

— When a rule names the short vowel a, it means every member of the class.

Problem three, what counts as near

— Inside a model, every word and every passage becomes a point in space.

— A passage can be near because it is on the same subject uses the same words or has the same tone or says the exact opposite.

— One study showed that the usual measure can, in some models, give results that are arbitrary.

— So the system fetches what is near and nobody can say what near means.

Chapter two, what is inside

— Four questions about layers, units, pieces, and explanations.

Problem four, layers without names

- The stages exist, their names and their borders do not.
- Pāṇini's grammar also works in layers, but each one can be named.
- Modern scholars count four levels, meaning roles and tenses, abstract word forms, and finally sound.
- Three sets of rules carry a sentence across from meaning to roles, from roles to forms, from forms to sound.

Problem five, one unit, many meanings

— A single neuron in an image model was found to respond to cat faces, to the fronts of cars and to cats' legs. cats' legs.

— A network has far more ideas to stall than it has neurons, so it packs many ideas into each neuron and spreads each idea across many. across many.

— It also means that no single unit can be read on its own.

— Take the root add a prefix and watch. a blow, a blow, food, destruction, food, destruction, a pleasure walk, avoidance.

Problem six, how far the split

- So researchers trained a second system to break a network's activity into separate features, each with one meaning.
- In one large model they found millions, a feature for a famous bridge, a feature for errors in computer code.
- So the question changed from what are the parts to when do we stop cutting?
- If a horse is called ashva because it covers the road, then whatever covers the road should be called a horse.

Problem seven, it cannot explain itself

- Ask a model how it reached an answer and it will tell you fluently.
- Researchers asked one model to add 36 and 59 and watched its internal steps.
- So if a word in a rule seems to do no work, it must be there for a reason.
- The list of sounds that opens the grammar uses one of its marker letters twice.

Chapter three, the source

— What the data lost, who sets the standard, and how far a model can travel. can travel.

Problem eight, what the data left out

- A model learns from text, but text is what remains after a great deal has been thrown away. thrown away.
- One audit of nearly 2,000 training sets found the license missing for most of them as they are shared online.
- His grammar was composed to be recited, not written, and some of its signals lived only in the voice.
- A heading that governs the rules after it was spoken with a special pitch.

Problem nine, models trained on models

- A study in the journal Nature tested what happens when each model is trained on the output of the one before.
- Sanskrit grammar faced the same danger, a standard that feeds on itself.
- If correct Sanskrit is whatever the grammar produces, then the grammar can never be wrong and never be checked.
- The true model speaker uses the forms the grammar teaches without having studied the grammar.

Problem 10, a new language, little data

— A model trained on English and a few dozen others reaches them only by transfer, by carrying over what it learned elsewhere.

— They freeze the large model and train a thin layer of new weights for the new language that carries over what the languages share.

— Indian grammarians answered this many centuries ago when they began to describe the other languages of India.

— Keep a source the model did not write and sort what transfers from what must be collected.

Sources and credit

— Study notes by Sanskrit AI Universe, based on the video “Sanskrit and AI Part 4 :The Hidden Layers | Dr.P.R.K.Varma.,M.D.,D.M.”. These are not the creator's own slides.

— Video by Dr P.R.K. Varma, Founding Editor, posted on the YouTube channel DrVARMA PRK.

— Watch the video: https://www.youtube.com/watch?v=YcaTkGbZw0s

— Captions used: English (auto-generated) (en), retrieved on 24 November 2026 via a public Invidious instance. Auto-captions can mis-hear words; obvious proper-name spellings were given in IAST where the journal already uses them.

— Also used: the official video title and description, and the Watch catalogue note already published in this journal (presenter and why the video is listed).

— If the Founding Editor later supplies his own deck for this part, replace the PPTX and PDF in src_assets/slides/ (see the journal README, “How to swap in RK's own slides”).