

OCTA RESEARCH – 365 Days of Learning, Theorizing, and Building AGI

Regularization as Singular Geometry Shaping in Transformers

Week 1 · Day 7 · January 7, 2026

OCTA Research Internal Theory Program

Version 1.2 – Living Document Series (Geometry-Expanded)

OCTA Research 365 Program Note (Week 1, Day 7).

Days 1–6 established:

- singular geometry as the right lens for Transformers (Days 1–3),
- SLT/RLCT, Fisher geometry, and probes (Days 3–4),
- singular scaling surfaces and capability loci over (N, P) (Day 5),
- training dynamics as stochastic flow on singular manifolds (Day 6).

Day 7 asks:

How do regularizers (weight decay, dropout, architectural priors) reshape the singular geometry and dynamics of training, and how can OCTA use that as a control surface?

We treat regularization not as an ad hoc trick, but as an explicit *geometry-shaping operator* that:

- bends minimizer manifolds and modifies their dimensions,
- shifts RLCT and Fisher spectra,
- gates training-time singular events and capability transitions,
- and opens a design channel for safety windows, robustness, and OCTA-native priors.

Executive Summary.

- Regularized Transformers are still singular models, but their singular manifolds and RLCT are *engineered objects*.
- L2, L1, spectral, dropout, and OCTA-specific regularizers act as *operators on the geometry* of Θ^* and on SGD dynamics.
- The right view is: “choose regularization to shape which manifolds exist, which are explored, and how capabilities emerge along them.”

Contents

1	Regularization as Loss and Geometry Modifier	3
1.1	Minimizer sets, strata, and deformation	3
1.2	Bayesian view: R as log-prior	4
2	Local Geometry, Flat Directions, and RLCT Deformation	4
2.1	Hessian perturbation	4
2.2	Approximate RLCT deformation	5
3	Canonical Regularizers as Geometry Operators	5
3.1	L2 weight decay: global quadratic bending	5
3.2	L1 penalties and singular corners	6
3.3	Spectral and path-norm regularization	6
3.4	Dropout as stochastic geometry operator	7
3.5	Toy example: L2 on a redundant parameterization	7
4	Transformer-Specific Geometry Shaping	8
4.1	Head entropy and diversity regularization	8
4.2	Attention entropy regularization	8
4.3	Layer norm and orthogonality priors	9
5	Interaction with Training Phases	10
5.1	Phase-dependent roles	10
5.2	Regularization and event gating	10
6	Geometry-Aware Regularization Protocols	11
6.1	Protocol D7.1: RLCT-vs- λ sweep	11
6.2	Protocol D7.2: Head-geometry regularizer sweep	11
6.3	Protocol D7.3: Curvature-aware regularization schedules	12
6.4	Protocol D7.4: Ablation across regularizer families	13
7	Safety Windows, Control Surface, and OCTA View	13
8	Synthesis and Outlook	14

List of Figures

1	Qualitative effect of regularization: a broad, flat minimizer manifold Θ_0^* is deformed into a narrower, more localized set Θ_λ^* . The singular model remains singular, but its geometry is reshaped.	4
2	Conceptual behavior of an RLCT proxy as a function of regularization strength λ . OCTA aims to operate in a band where RLCT is neither too low (degenerate) nor too high (overconstrained).	6
3	Dropout as stochastic geometry operator: minimizers must perform well across many masked subnetworks, selecting a more robust sub-manifold inside Θ_0^* .	7
4	Conceptual head-entropy trajectories with and without head-geometry regularization. R_{head} pulls $H_{\text{head}}(t)$ into a desired diversity band.	9

5	Orthogonality priors regularize representation geometry: from irregular ellipsoids to more isotropic shapes, simplifying Fisher structure and potentially RLCT.	10
6	Conceptual relationship between RLCT proxy and generalization as a function of regularization strength λ . Protocol D7.1 maps these curves empirically for a given architecture/dataset.	12
7	Extended safety window concept in $(N, \lambda, \text{Cap}_T)$ space: only certain combinations of model size and regularization strength fall into safe geometry and capability regimes. Regularization becomes a safety control axis.	14

1 Regularization as Loss and Geometry Modifier

Let $\theta \in \Theta \subset \mathbb{R}^d$ denote Transformer parameters and let $L_0(\theta)$ be the base empirical loss (e.g. negative log-likelihood). A generic regularized objective is

$$L(\theta) = L_0(\theta) + \lambda R(\theta), \quad (1)$$

where $\lambda \geq 0$ is a regularization strength and R is a penalty or prior term (weight decay, sparsity, orthogonality, head diversity, etc.).

1.1 Minimizer sets, strata, and deformation

Define the unregularized minimizer set

$$\Theta_0^* := \arg \min_{\theta} L_0(\theta),$$

and for each λ ,

$$\Theta_\lambda^* := \arg \min_{\theta} L(\theta).$$

In the overparameterized regime relevant to OCTA (and modern Transformers), Θ_0^* is typically a large, stratified, singular set:

$$\Theta_0^* = \bigcup_s \Theta_{0,s},$$

each $\Theta_{0,s}$ a smooth manifold (stratum) of dimension $d_{0,s}$, glued together along lower-dimensional intersections (Days 2–3).

Turning on λ induces a *deformation*:

$$(\Theta_0^*, L_0, F_0) \mapsto (\Theta_\lambda^*, L, F_\lambda),$$

where F_0, F_λ denote Fisher geometries.

Definition 1.1 (Geometry-shaping operator). A regularizer R induces a *geometry-shaping operator*

$$\mathcal{G}_R(\lambda) : (\Theta_0^*, L_0, F_0) \longrightarrow (\Theta_\lambda^*, L_0 + \lambda R, F_\lambda),$$

mapping unregularized singular geometry to regularized singular geometry as a function of λ .

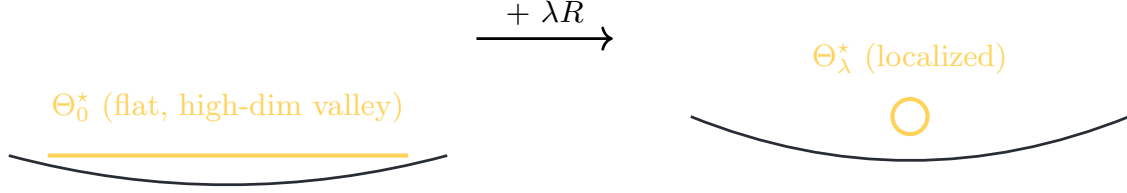


Figure 1: Qualitative effect of regularization: a broad, flat minimizer manifold Θ_0^* is deformed into a narrower, more localized set Θ_λ^* . The singular model remains singular, but its geometry is reshaped.

1.2 Bayesian view: R as log-prior

From a Bayesian viewpoint, (1) is equivalent to

$$L(\theta) = -\log p(\mathcal{D} \mid \theta) - \log p_\lambda(\theta),$$

with prior

$$p_\lambda(\theta) \propto \exp(-\lambda R(\theta)).$$

Thus:

- R defines an explicit *prior geometry* over parameters;
- the posterior geometry inherits singularities from both likelihood and prior;
- RLCT and Fisher geometry now depend on the interaction of model and prior.

For OCTA, this perspective is crucial: regularization becomes a channel to encode structural priors (OCTA neurons, mesh constraints, safety priors) directly as geometry.

2 Local Geometry, Flat Directions, and RLCT Deformation

We study how (1) changes local curvature and RLCT in a neighborhood of a minimizer.

2.1 Hessian perturbation

Let $H_0(\theta) := \nabla^2 L_0(\theta)$ and $H_R(\theta) := \nabla^2 R(\theta)$ (where defined). Then:

$$H_\lambda(\theta) := \nabla^2 L(\theta) = H_0(\theta) + \lambda H_R(\theta).$$

Let $\theta^* \in \Theta_0^*$ and consider its local linear algebra. Write

$$H_0(\theta^*) = U \begin{pmatrix} 0_{k \times k} & 0 \\ 0 & \Lambda_+ \end{pmatrix} U^\top,$$

with Λ_+ positive-definite and k zero eigenvalues (flat directions).

Definition 2.1 (Flat subspace and curved subspace). The *flat subspace* at θ^* is

$$\mathcal{F}_0 := \ker H_0(\theta^*),$$

of dimension k , and the *curved subspace* is its orthogonal complement.

Proposition 2.2 (L2 regularization lifts flat directions). *Suppose $R(\theta) = \frac{1}{2}\|\theta\|^2$ so that $H_R(\theta) = I_d$. Then*

$$H_\lambda(\theta^*) = H_0(\theta^*) + \lambda I_d$$

has eigenvalues

$$\lambda, \dots, \lambda \quad (k \text{ times}), \quad \lambda + \mu_1, \dots, \lambda + \mu_{d-k}$$

where $\mu_i > 0$ are eigenvalues of Λ_+ . In particular, the flat subspace \mathcal{F}_0 acquires curvature λ .

Proof. Immediate from spectral decomposition: adding λI_d shifts all eigenvalues by λ without changing eigenvectors. \square

Remark 2.3. More general R with $H_R(\theta^*)$ positive-definite on \mathcal{F}_0 will also lift flat directions. If H_R is only semidefinite, some flat directions may remain singular.

2.2 Approximate RLCT deformation

Exact RLCT under regularization is analytically challenging, but we can reason qualitatively:

- Flat directions in L_0 contribute to low RLCT λ_0 (high degeneracy).
- Lifting some flats with H_R raises effective RLCT λ_λ .
- Excessive lifting (large λ or too strong H_R) can overconstrain the model, harming fit and usable singular structure.

OCTA Principle 2.4 (OCTA Regularization Principle I – RLCT steering). Regularization strength λ and form R should be viewed as steering the effective RLCT λ_λ :

- *under-regularized*: many nearly-flat directions, low RLCT, high variance,
- *over-regularized*: excessively high RLCT, reduced expressivity,
- *OCTA regime*: intermediate RLCT where useful singular manifolds remain but harmful degeneracies are trimmed.

3 Canonical Regularizers as Geometry Operators

We now interpret standard regularizers as explicit operators on singular geometry and training dynamics.

3.1 L2 weight decay: global quadratic bending

For $R(\theta) = \frac{1}{2}\|\theta\|_2^2$ and small λ :

- flat directions of L_0 acquire curvature λ ;
- narrow minima become narrower (curvature shifts by $+\lambda$);
- the SGD stationary distribution (Day 6) becomes more strongly concentrated near the origin in parameter space.

From a symmetries viewpoint:

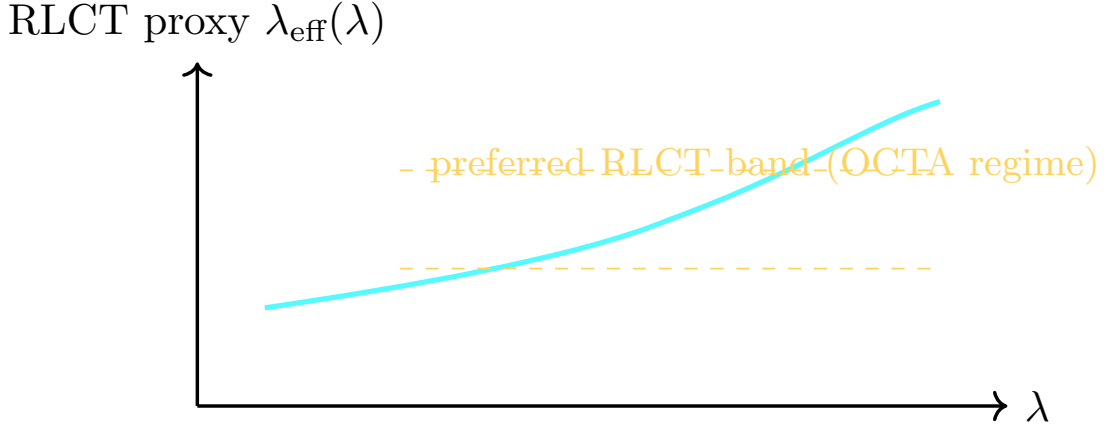


Figure 2: Conceptual behavior of an RLCT proxy as a function of regularization strength λ . OCTA aims to operate in a band where RLCT is neither too low (degenerate) nor too high (overconstrained).

- Θ_0^* often decomposes into orbits under scaling or reparametrization symmetries;
- Θ_λ^* selects orbit representatives with minimal norm.

OCTA Principle 3.1 (OCTA Regularization Principle II – Center-of-mass selection). In overparameterized Transformers, L2 weight decay approximately chooses *center-of-mass* representatives on each symmetry orbit in Θ_0^* , shrinking redundant degrees of freedom while (for moderate λ) preserving functional equivalence.

3.2 L1 penalties and singular corners

For $R(\theta) = \|\theta\|_1$, the gradient and Hessian are not smooth everywhere, but the geometric picture is:

- L1 introduces *corners* in the landscape, aligning singularities with coordinate axes;
- minimizer sets intersect coordinate subspaces more frequently (sparsity);
- singular manifolds may turn into piecewise-linear structures.

Remark 3.2. L1 can simultaneously increase interpretability (many exact zeros) and complicate Fisher geometry (non-differentiable points), which matters for OCTA’s curvature-based telemetry (Day 4).

3.3 Spectral and path-norm regularization

For matrices W (e.g. attention projections, MLP weights), spectral or path-norm regularizers such as:

$$R_\sigma(W) = \|W\|_{\text{op}}^2, \quad R_{\text{path}}(\theta) = \sum_{\text{paths}} \prod_{l \in \text{path}} |w_l|^2$$

shape the geometry at the level of functions rather than raw parameters:

- they penalize directions that strongly amplify inputs;
- they bound Lipschitz constants or path-wise amplification;

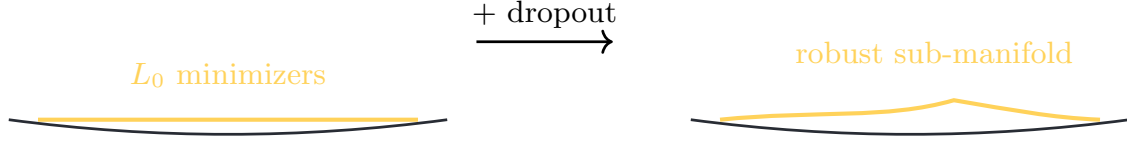


Figure 3: Dropout as stochastic geometry operator: minimizers must perform well across many masked subnetworks, selecting a more robust sub-manifold inside Θ_0^* .

- they smooth out extreme curvature in directions associated with exploding activations or high-gain circuits.

OCTA Principle 3.3 (OCTA Regularization Principle III – Operator-level shaping). Spectral and path-norm regularizers primarily act in function space: they shape the geometry of realized maps (Lipschitz, smoothness) rather than only parameter norms. For OCTA, they are key tools for defining safety windows around sensitivity and amplification.

3.4 Dropout as stochastic geometry operator

Dropout is not a deterministic penalty term but can be seen as a stochastic geometry operator:

- at each step, it randomly zeroes subsets of activations or weights;
- the effective loss is an expectation over subnetworks induced by masks;
- minimizers must perform well across many masked subnetworks.

Heuristically:

- dropout acts like a *smoothing* of the loss surface over a set of subnetworks;
- it favors directions that are robust under random pruning;
- it reduces reliance on precise interference between many parameters.

3.5 Toy example: L2 on a redundant parameterization

Consider the redundant scalar regression model with prediction $\hat{y} = uvx$ and squared loss

$$L_0(u, v) = \frac{1}{2}(uv - w^*)^2,$$

with minimizer curve $\Theta_0^* = \{(u, v) : uv = w^*\}$ (Day 6).

Add L2 regularization:

$$L(u, v) = \frac{1}{2}(uv - w^*)^2 + \frac{\lambda}{2}(u^2 + v^2).$$

Proposition 3.4 (Center-of-mass point on redundant curve). *For $w^* \neq 0$ and $\lambda > 0$, the global minimizer of $L(u, v)$ satisfies $u = v$ and hence:*

$$u = v = \text{sgn}(w^*)\sqrt{|w^*| + O(\lambda)}.$$

As $\lambda \rightarrow 0$, (u, v) approaches the point on $uv = w^*$ with $|u| = |v|$, i.e. the “balanced” parametrization.

Proof. For $u, v \neq 0$, impose the constraint $uv = w^*$ and minimize the L2 term $u^2 + v^2$ subject to this constraint. By symmetry (or by Lagrange multipliers), the minimum is attained at $|u| = |v|$, so $u = v$ with $uv = w^*$. A small $\lambda > 0$ perturbs this solution smoothly. \square

Geometrically:

- the entire curve $uv = w^*$ collapses to a single point as $\lambda > 0$;
- this point is the norm-minimizing representative of the equivalence class;
- in high-dimensional Transformers, analogous phenomena happen across many redundant parametrizations and head/layer symmetries.

4 Transformer-Specific Geometry Shaping

We now define regularizers tailored to Transformer singular geometry, building on Days 2, 4, and 6.

4.1 Head entropy and diversity regularization

Let $E_{t,h}$ be headwise Fisher energy (Day 4). Define normalized energies

$$\tilde{E}_{t,h} := \frac{E_{t,h}}{\sum_{h'} E_{t,h'}},$$

and head specialization entropy

$$H_{\text{head}}(t) := - \sum_h \tilde{E}_{t,h} \log \tilde{E}_{t,h}.$$

We introduce a regularizer that steers $H_{\text{head}}(t)$ toward a *target diversity*:

$$R_{\text{head}}(\theta) = \alpha (H_{\text{head}}(t) - H_{\text{target}})^2,$$

where H_{target} encodes a desired balance between specialization and redundancy.

OCTA Principle 4.1 (OCTA Regularization Principle IV – Head geometry shaping). Head entropy regularization shapes the singular geometry of the attention subsystem by:

- discouraging degenerate states with a single dominant head;
- discouraging pathological uniformity where all heads are nearly identical;
- steering training toward configurations with a diverse but structured set of specialized heads, associated with richer but controlled singular strata.

4.2 Attention entropy regularization

At token level, penalize overly peaked or overly flat attention distributions. For attention weights a_{ij} :

$$R_{\text{attn}}(\theta) = \beta \mathbb{E}_i \left[(H(a_{i\cdot}) - H_{\text{attn-target}})^2 \right],$$

where $H(a_{i\cdot})$ is the Shannon entropy of the attention distribution over keys j .

Geometrically:

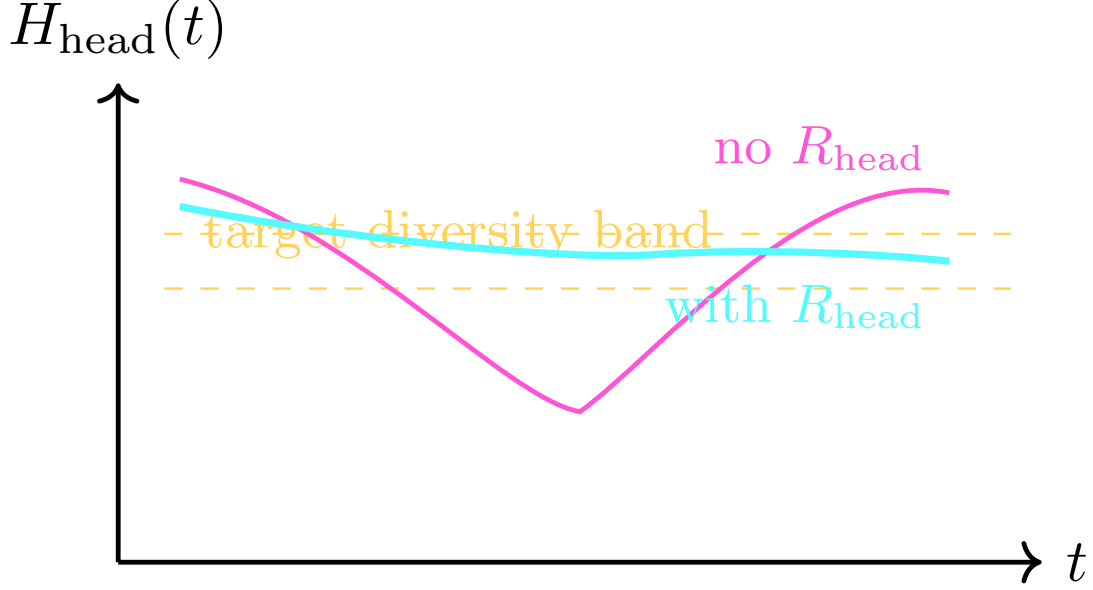


Figure 4: Conceptual head-entropy trajectories with and without head-geometry regularization. R_{head} pulls $H_{\text{head}}(t)$ into a desired diversity band.

- extreme hard attention corresponds to sharp, singular loci in attention space (Day 2);
- extreme soft attention underutilizes the architecture’s singular expressivity;
- R_{attn} shapes the typical location of attention maps in the stratified attention manifold.

4.3 Layer norm and orthogonality priors

Layer norms and residual connections already impose structural priors. We can further include:

- orthogonality regularizers on projection matrices W_Q, W_K, W_V :

$$R_{\text{orth}}(W) := \gamma \|W^\top W - I\|_F^2;$$

- penalties on deviations from identity in selected residual pathways;
- constraints encouraging near-isometries in representation subspaces.

OCTA Principle 4.2 (OCTA Regularization Principle V – Representation geometry shaping). Orthogonality and isometry priors define a smoother, more uniform geometry in representation space, which:

- stabilizes curvature across directions,
- avoids pathological “needle” manifolds with extreme amplification,
- supports more predictable singular events and safer capability transitions.

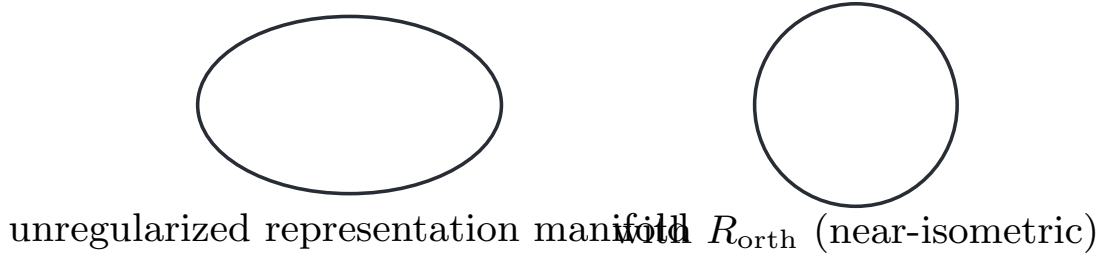


Figure 5: Orthogonality priors regularize representation geometry: from irregular ellipsoids to more isotropic shapes, simplifying Fisher structure and potentially RLCT.

5 Interaction with Training Phases

Regularization interacts strongly with the training phases T0–T3 introduced on Day 6.

5.1 Phase-dependent roles

- **T0 (exploration / chaotic descent):**
 - strong regularization can prematurely constrain exploration;
 - moderate weight decay and dropout encourage robust directions and prevent early divergence;
 - geometry is still fluid; we prefer gently shaping priors here.
- **T1 (valley entry / stabilization):**
 - regularizers start to sculpt the valleys being entered;
 - head entropy and attention entropy constraints promote good division of labor;
 - orthogonality and spectral regularizers control curvature.
- **T2 (plateaus + events):**
 - regularization modulates the frequency and magnitude of training-time singular events;
 - stronger regularization can smooth transitions but may suppress beneficial capability jumps;
 - weaker regularization can permit richer geometry but increase volatility and risk of unsafe transitions.
- **T3 (fine-tuning / saturation):**
 - regularization trades off long-run generalization vs. tailoring to specific downstream tasks;
 - geometry-informed scheduling (Section 6.3) is most impactful here.

5.2 Regularization and event gating

OCTA Principle 5.1 (OCTA Regularization Principle VI – Event gating). In T2, regularization acts as an *event gate*:

- it controls which singular exits (Day 6) are energetically accessible;

- it biases SGD toward transitions that preserve flatness and low RLCT;
- it can be tuned to avoid geometry regimes associated with unsafe or undesirable capability jumps.

6 Geometry-Aware Regularization Protocols

We now define explicit OCTA protocols to empirically study and deploy geometry-aware regularization.

6.1 Protocol D7.1: RLCT-vs- λ sweep

Experimental Protocol 6.1 (D7.1: RLCT-vs- λ mapping).

- Fix architecture, dataset, and optimizer.
- Choose a grid of regularization strengths $\lambda \in \{\lambda_1, \dots, \lambda_m\}$ for a given R (e.g. L2).
- For each λ_k :
 - train multiple runs with different seeds;
 - record RLCT proxies over time (Day 4 methods);
 - record final generalization metrics and capabilities (Days 5–6).
- Fit curves $\lambda \mapsto \lambda_{\text{eff}}(\lambda)$ and $\lambda \mapsto \text{Gen}(\lambda)$.
- Identify:
 - ranges where RLCT increases but generalization worsens (over-regularization),
 - ranges where RLCT is very low and generalization unstable (under-regularization),
 - intermediate ranges with best trade-offs (OCTA band).

6.2 Protocol D7.2: Head-geometry regularizer sweep

Experimental Protocol 6.2 (D7.2: Head geometry regularization).

- Implement R_{head} and R_{attn} from Section 4.
- Sweep the regularization weights (α, β) over a grid.
- For each configuration:
 - track $H_{\text{head}}(t)$, $E_{t,h}$, and attention-strata occupancy;
 - record training-time singular events and capability jumps (Day 6);
 - evaluate generalization and robustness metrics.
- Analyze:
 - how head specialization profiles change with (α, β) ;
 - whether certain geometry profiles correlate with desirable capabilities (e.g. better in-context learning, robustness to distribution shift);
 - which (α, β) regimes produce stable vs. chaotic singular events.

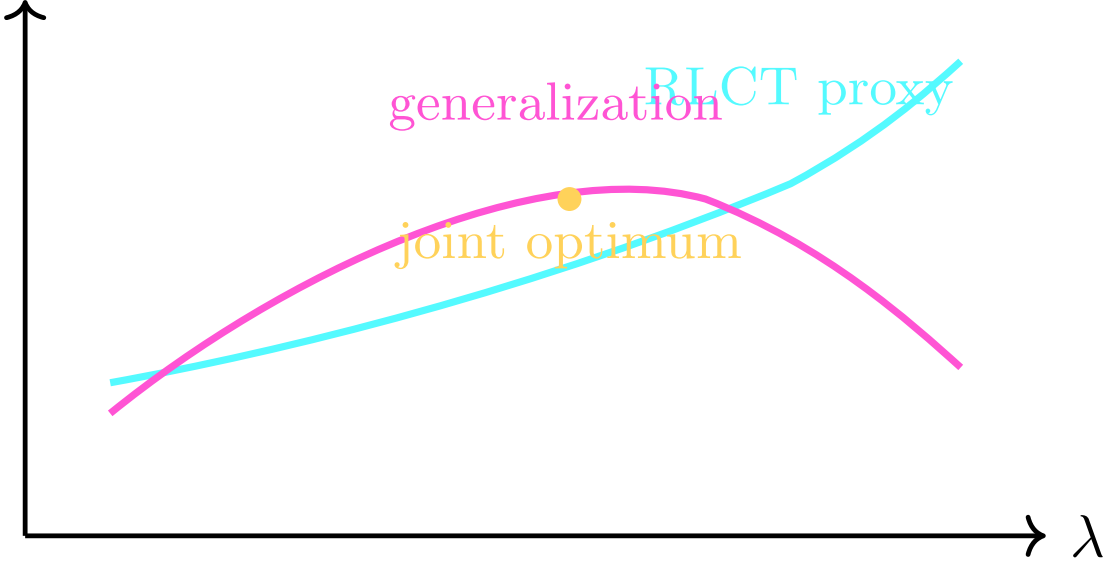


Figure 6: Conceptual relationship between RLCT proxy and generalization as a function of regularization strength λ . Protocol D7.1 maps these curves empirically for a given architecture/dataset.

6.3 Protocol D7.3: Curvature-aware regularization schedules

Combine Day 6 curvature-aware learning rate schedules with regularization schedules:

$$\lambda_{t+1} = \lambda_t \cdot g(\text{Tr}(F_t), \lambda_{t,\min}^+).$$

Experimental Protocol 6.3 (D7.3: Curvature-aware λ schedule).

- (a) At checkpoints, compute curvature indicators as in Day 6:

$$C_t := \alpha_1 \text{Tr}(F_t) + \alpha_2 \lambda_{t,\min}^+.$$

- (b) Define a target curvature range $[C_{\min}, C_{\max}]$.

- (c) Update learning rate and regularization strength:

$$\eta_{t+1} = \begin{cases} \eta_t \cdot (1 + \epsilon_\eta), & C_t < C_{\min}, \\ \eta_t, & C_{\min} \leq C_t \leq C_{\max}, \\ \eta_t \cdot (1 - \epsilon_\eta), & C_t > C_{\max}, \end{cases} \quad \lambda_{t+1} = \begin{cases} \lambda_t \cdot (1 - \epsilon_\lambda), & C_t < C_{\min}, \\ \lambda_t, & C_{\min} \leq C_t \leq C_{\max}, \\ \lambda_t \cdot (1 + \epsilon_\lambda), & C_t > C_{\max}, \end{cases}$$

with small $\epsilon_\eta, \epsilon_\lambda > 0$.

- (d) Evaluate whether this joint schedule:

- stabilizes training near high-curvature regions,
- preserves beneficial singular events,
- improves final performance and robustness.

6.4 Protocol D7.4: Ablation across regularizer families

Experimental Protocol 6.4 (D7.4: Regularizer family ablation).

- (a) Define a base configuration (architecture, dataset, optimizer, base λ).
- (b) Construct a matrix of conditions:

$$L2 \in \{0, \lambda\}, \quad \text{dropout} \in \{\text{off}, \text{on}\}, \quad R_{\text{head}} \in \{0, \alpha\}, \quad R_{\text{orth}} \in \{0, \gamma\}, \dots$$

- (c) For each combination:
 - run multiple seeds and record geometry (Fisher, RLCT proxies, head/layer metrics),
 - record training-time singular events,
 - compute capability metrics (Days 5–6).
- (d) Build a *regularizer influence map*:
 - which regularizers primarily affect RLCT,
 - which primarily affect head diversity,
 - which primarily affect event frequency and magnitude,
 - and how they interact (synergy vs. interference).

7 Safety Windows, Control Surface, and OCTA View

Day 5 introduced *safety windows* in (N, P) where capabilities and risk profiles are acceptable. Day 7 extends this to include regularization axes $(\lambda, \alpha, \beta, \gamma, \dots)$.

Definition 7.1 (Extended safety window). An *extended safety window* is a subset

$$\mathcal{S} \subset \{(N, P, \lambda, \alpha, \beta, \gamma, \dots)\}$$

such that:

- capabilities are within the desired envelope (neither too weak nor too strong),
- training geometry (RLCT, Fisher spectra, singular events) lies within acceptable ranges,
- robustness and safety metrics (e.g. adversarial robustness, misuse risk) meet predefined thresholds.

OCTA Principle 7.2 (OCTA Regularization Principle VII – Safety shaping). Regularization hyperparameters are part of the OCTA safety control system:

- they can be tuned to avoid risky geometry regimes (e.g. extremely sharp, high-gain representations);
- they can be combined with telemetry (Days 4–6) to define dynamic interventions during training;
- they allow OCTA to specify not just *what capabilities* emerge but *how those capabilities arise geometrically and how stable they are*.

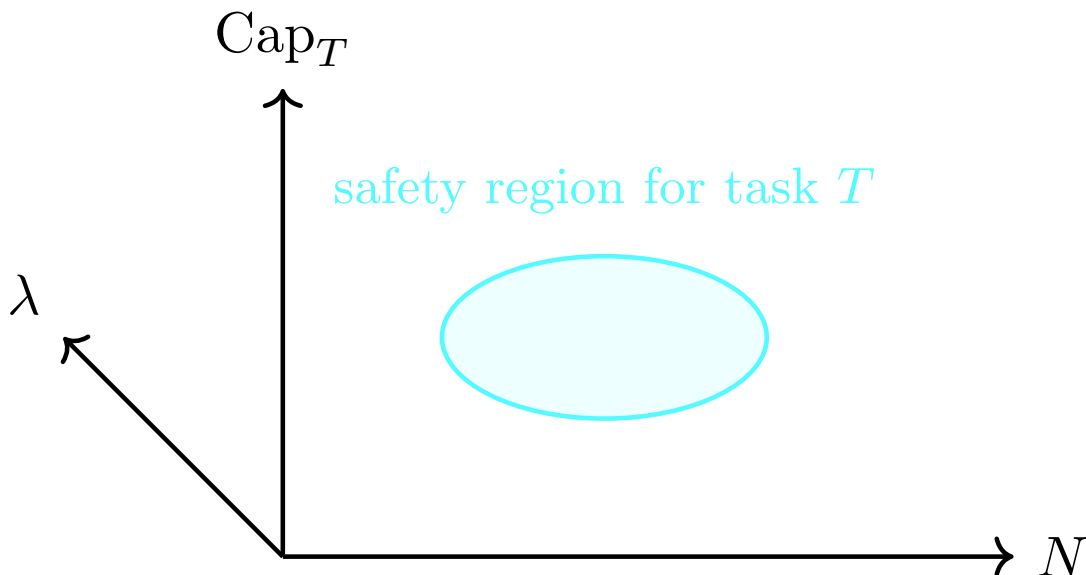


Figure 7: Extended safety window concept in $(N, \lambda, \text{Cap}_T)$ space: only certain combinations of model size and regularization strength fall into safe geometry and capability regimes. Regularization becomes a safety control axis.

8 Synthesis and Outlook

Day 7 integrates regularization into the Week 1 singular geometry picture:

- Days 1–3: Transformers are singular models with stratified minimizer sets.
- Day 3: SLT/RLCT provides the asymptotic language for these singularities.
- Day 4: Fisher and RLCT probes expose local geometry.
- Day 5: scaling surfaces and capability loci live on this geometry.
- Day 6: training is a stochastic flow on singular manifolds, with plateaus and training-time singular events as dynamical signatures.
- Day 7: regularization is a geometry-shaping operator that bends these manifolds, steers RLCT, gates singular events, and defines safety windows.

Compressed OCTA view:
Architecture \Rightarrow Singular Geometry
Regularization \Rightarrow Geometry Shaping
SGD Dynamics \Rightarrow Trajectories, Plateaus, Events
 \Rightarrow Capabilities & Safety Windows.

For OCTA as a 365-day program, the next steps are:

- **Day 8:** move beyond generic regularizers to *OCTA-native regularizers*: geometry-aware priors tied to OCTA neurons, mesh training, P3P-like decentralized learning, and Perfect Attractor structure.

- **Days 9–10:** tighten the link between regularization geometry and interpretability: map which regularizers produce which circuit motifs, attention motifs, and meso-scale OCTA structures.
- **Subsequent weeks:** fuse this with hardware, edge-device constraints, and distributed training (H200-class clusters + OCTA edge mesh), so that regularization, geometry, and compute topology are all treated as parts of a single control surface.

**Regularizers are not merely knobs for overfitting;
they are levers that sculpt the singular manifolds where AGI lives.**

References

- [1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 1998.
- [2] S. Mandt, M. D. Hoffman, D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 2017.
- [3] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.
- [4] S. Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.
- [5] B. Neyshabur, S. Bhojanapalli, D. McAllester, N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2017.
- [6] Y. Gal, Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.