What Fire Teaches Us About Large Language Models

Combustion, Containment, and the Case for a Fire Department

Elan Moritz

Eagles Perch Press

What Fire Teaches Us About Large Language Models
Combustion, Containment, and the Case for a Fire Department

Published by Eagles Perch Press, Philadelphia, Pennsylvania.
Frontiers of Intelligence series.

ORCID: 0000-0002-0167-4336

Set in EB Garamond with Latin Modern Sans and DejaVu Sans Mono.
Typeset with LuaLaTeX.

For Brélan
Forever missed

No civilization has ever abolished fire. Every civilization worth the name has built an institution whose sole business is arriving when fire goes where it was not invited.

— Working note, *Eternity's Pizza*, 2026

Contents

Preface

The argument of this short treatise began as an objection to itself. Analogies between artificial intelligence and earlier general-purpose technologies are cheap, abundant, and mostly bad. Electricity, the printing press, nuclear fission, the automobile: each has been pressed into service, and each has supplied a season's worth of op-eds before collapsing under the weight of its own disanalogies. I did not want to add fire to the pile.

What changed my mind was noticing that the interesting content of the fire analogy is not in the fire. It is in the fire *department*. The comparison that earns its keep is not "large language models are like fire" — a claim about the technology — but "our relationship to large language models will require something like the institution we built for fire" — a claim about us. The first is a metaphor. The second is a testable proposition about institutional form, and it can be wrong in specific ways.

Chapter 3 is therefore the load-bearing chapter, not the throat-clearing one. It catalogs where the analogy fails: fire does not optimize, does not model its extinguishers, and does not propagate through channels that also carry the alarm. Readers inclined to think the fire comparison is obviously apt should read that chapter first and treat the remainder as what survives it.

A word on method. This essay operates in the mode I have elsewhere called *operational constructivism*: a concept earns its ontological keep by the work it does in a validation loop, not by its resemblance to a prior concept. The fire analogy is admitted here only to the extent that it generates commitments one could act on and be embarrassed by. Where it generates only mood, I have tried to say so.

Philadelphia
July 2026

PART I

The Analogy and Its Limits

CHAPTER 1

The Thesis, Stated Narrowly

1.1 What is being claimed

Fire is the standing example of a technology that a civilization neither abandons nor fully controls. We do not consider fire a solved problem. We consider it a *managed* one, and the management is institutional rather than technical. The claim of this treatise is that large language models occupy, or will shortly occupy, the same category, and that the institutional repertoire developed for fire over roughly two centuries constitutes a usable design vocabulary — not a blueprint, but a vocabulary.

The claim is narrow in three deliberate ways.

1. It is a claim about *governance form*, not about capability, risk magnitude, or timeline. Nothing here depends on whether LLMs are close to or far from any particular threshold.
2. It is a claim about *diffuse, high-frequency, low-per-incident harm* — the regime in which fire actually lives — and not about catastrophic single-event risk, for which the fire analogy is actively misleading.
3. It is defeasible in stated ways. Chapter 3 names the conditions under which the analogy should be discarded rather than repaired.

The proposition

For any general-purpose capability that (i) society will not relinquish, (ii) is available to essentially all users, (iii) fails frequently but locally, and (iv) whose failures propagate through shared infrastructure, the historically stable governance response is a *standing, publicly funded, non-punitive response institution* paired with *ex ante material codes*. Fire is the paradigm case. LLMs satisfy (i)–(iv).

1.2 Why the usual analogies underperform

The electricity analogy (Brynjolfsson et al. 2021) is a claim about diffusion economics: general-purpose technologies take decades to show up in productivity statistics because complementary investments lag. It is a good analogy for *when* value appears and a poor one for *who cleans up*, because electrical harm was addressed largely through embedded standards rather than a response corps.

The electrical case is not silent on institutions — underwriter laboratories and the National Electrical Code matter enormously — but its response function is maintenance, not emergency dispatch.

The nuclear analogy is a claim about catastrophic tail risk and non-proliferation, and it imports a governance model — licensure, international inspection, restricted materials — built for a technology with a few hundred sites worldwide. LLMs have hundreds of millions of endpoints. Nuclear governance does not scale down; fire governance was built to scale down, because fire is everywhere and always has been.

The printing press analogy is a claim about epistemic disruption and is the best of the three for content harms (Eisenstein 1979). It has no response institution to offer, because none was built. That absence is itself informative: societies tolerated roughly two centuries of pamphlet chaos, and the eventual settlement was libel law, professional norms, and market structure rather than any dispatchable body.

1.3 The structure of the argument

Part I establishes the physical analogy and then attacks it. Chapter 2 maps the combustion triangle — fuel, oxidizer, ignition — onto a corresponding triangle for LLM harm, and identifies the mapping's genuinely load-bearing feature: *harm requires co-presence of independently innocuous conditions*. Chapter 3 lists seven disanalogies, three of which are fatal to naive versions of the argument.

Part II turns to the institution. Chapter 4 reconstructs why the municipal fire department took the form it did, with attention to the insurance-brigade period that preceded it — a period whose failure mode is being recapitulated in LLM trust and safety today. Chapter 5 treats building codes and the fire marshal as the *ex ante* complement to *ex post* response. Chapter 6 takes up prescribed burning and the suppression paradox, which yields the sharpest and least comfortable lesson in the essay.

Part III converts the analysis into design commitments (Chapter 7) and states what would falsify them (Chapter 8).

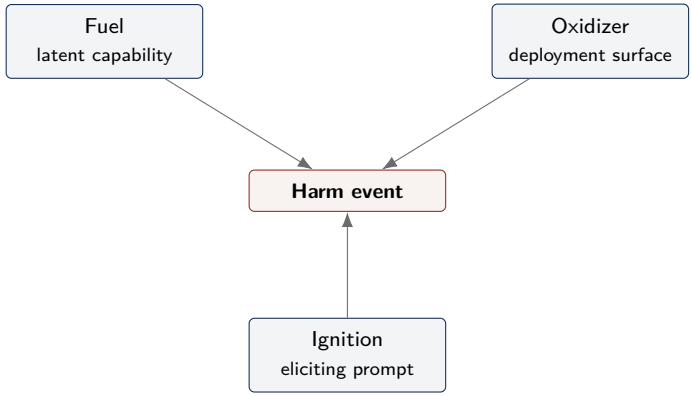
CHAPTER 2

Combustion as a Structural Model

2.1 The triangle and the tetrahedron

The elementary account of combustion requires three co-present conditions: fuel, an oxidizer, and an ignition source of sufficient energy. Fire suppression doctrine adds a fourth — the sustaining chain reaction — to yield the fire tetrahedron, and the addition is not pedantic. It is what licenses chemical suppression agents that interrupt radical propagation without removing any of the first three legs.

The structural content is this: none of the legs is dangerous alone. Fuel is inert. Oxygen is ambient and necessary for life. Ignition sources are ubiquitous and mostly useful. Harm is a property of the *configuration*, not of any component. This is why fire prevention is a discipline of separation rather than elimination, and why one cannot make a building fireproof by banning any single item.



2.2 The corresponding triangle for language models

The mapping I propose is deliberately conservative.

Fuel — latent capability.
The model's learned competences, including competences no one selected for. Fuel is not a defect; it is the reason the thing is worth having. A model with no dangerous capability is a model with little useful capability, for much the same reason that a fuel with no combustion enthalpy is a poor fuel.

Oxidizer — deployment surface.
Tool access, network reach, persistent memory, agentic autonomy, user population. This is the leg that is genuinely controllable and that receives, in my judgment, disproportionately little attention relative to the fuel leg.

Ignition — the eliciting context.
The prompt, the jailbreak, the unusual conversational trajectory, the adversarial document retrieved mid-task.

Chain reaction — autocatalytic propagation.
Model output entering training corpora, agent-to-agent delegation, and synthetic content that shapes the next round of elicitation.

What the mapping earns

Three consequences follow that do not follow from vaguer AI-risk framings:

1. **Single-leg interventions are structurally incomplete.** Alignment training addresses fuel. Content filtering addresses ignition. Neither is sufficient, and their sufficiency claims should be read as suppression-agent claims: effective on a class of fires, not on fire.
2. **The oxidizer leg is the cheapest intervention.** Reducing deployment surface — narrower tool scopes, shorter memory, constrained autonomy — lowers harm probability without touching capability. Fire

codes reflect this: we regulate ventilation and compartmentation far more aggressively than we regulate what materials may exist.

3. **The chain-reaction leg justifies interventions that look disproportionate.** Halon and its successors are expensive and environmentally costly, and were adopted anyway because interrupting propagation is worth more than removing any single leg. Provenance marking and training-corpus hygiene occupy this position.

2.3 Flashover and the nonlinearity of the regime

Structural firefighting distinguishes the growth phase from *flashover*, the point at which radiant heat brings all exposed combustible surfaces in a compartment to ignition temperature more or less simultaneously. Before flashover, a room is a fire problem. After flashover, it is a different physical regime, and tactics developed for the first are not merely less effective but actively lethal.

The relevant discipline here is not the observation that LLM deployment might have thresholds — everyone says that — but the operational practice built around it. Firefighters are trained to read *pre-flashover indicators*: rollover at the ceiling, darkening smoke under pressure, a sudden drop in visible flame as the compartment becomes fuel-rich. The indicators are cheap to observe and precede the transition by tens of seconds. The institutional insight is that where nonlinear transitions exist, the tractable investment is in *leading indicator literacy distributed to the people at the scene*, not in prediction from outside.

This is a direct criticism of the current evaluation regime, which concentrates threshold detection in pre-deployment testing conducted by parties who are, by construction, not at the scene. The fire service's answer to "when does this become dangerous" was never a better laboratory. It was training every person who enters the compartment to read the ceiling.

CHAPTER 3

Where the Analogy Fails

An analogy is worth using only if one can say what would break it. This chapter lists seven disanalogies in ascending order of severity. The first four are survivable — they narrow the claim. The last three are fatal to any version of the argument that treats LLM harm as *literally* a fire-management problem, and they are the reason Chapter 1 restricted the thesis to institutional form.

3.1 Survivable disanalogies

1. Fire is physically bounded; information is not

Fire propagates through matter at speeds set by thermodynamics and is stopped by the absence of fuel. Harmful information propagates at network speed and encounters no equivalent of a firebreak, because the medium is not consumed by transmission. Consequence: containment doctrine transfers, containment *physics* does not. Any proposal that assumes a spreading harm can be outrun by a response unit is importing the wrong half.

2. Fire has no interest in the fire department

Combustion does not model its suppressors. This disanalogy is survivable only because most LLM harm at present is not adversarially optimized against responders; it is misuse by ordinary users and error by ordinary systems. The disanalogy becomes fatal precisely when that ceases to be true, which is why it is listed here rather than below — it is a *time-indexed* defeater, and monitoring for its arrival is itself a design commitment (Chapter 7).

3. Fire damage is legible; model harm often is not

A burned building is visibly burned. Harm from a persuasive falsehood, a bad medical inference, or an eroded epistemic practice may be undetectable at the scene and may not localize to an incident at all. Dispatch models require alarms, and alarms require detectable events. Where harm is diffuse and slow, the fire department is the wrong institution and the public health service is the better one.

This is the strongest reason to treat the fire analogy as covering only a *subset* of llm harms — roughly, those with an identifiable incident.

4. Fire is a natural kind; llm harm is a heterogeneous list

We can define fire. "LLM harm" names a family with no shared mechanism: bioweapon uplift, defamation, dependency, labor displacement, and prompt injection have essentially nothing in common except a causal ancestor. A single institution addressing all of them would be an institution without a doctrine. The response, developed in Chapter 4, is that the fire service also lacks a single doctrine — structural, wildland, hazmat, and technical rescue are largely separate disciplines under one dispatch umbrella.

3.2 Fatal disanalogies

These break the literal reading

5. There is no analogue of extinguishment

A fire, once out, is out. There is no state of an LLM deployment corresponding to "extinguished." Model weights persist, copies proliferate, and capability once demonstrated is capability that others reproduce. The fire service's entire operational tempo — respond, suppress, overhaul, return to quarters — presupposes a terminal state that does not exist here. Any proposal for an "AI fire department" that imagines units returning to quarters after resolving an incident has mistaken the metaphor for the mechanism.

6. The response channel is the propagation channel

Fire departments do not arrive via the fire. They use roads, which fire does not travel. Responses to LLM-mediated harm travel the same networks, platforms, and often the same models as the harm. This collapses a separation that is structural in the fire case and creates failure modes with no fire analogue: a compromised response capability, a response that is itself generated content of contested provenance, and an adversary who observes the response in real time.

7. Fire departments do not build fires

The organizations best positioned to respond to model harm are the organizations that produced the models. In the fire case, this configuration existed — it was the insurance brigade era, discussed in Chapter 4 — and it was abandoned as unworkable within a few decades. That abandonment is the single most transferable fact in this essay, and it is transferable as a *warning*, not as reassurance.

3.3 What survives

Stripping out everything the disanalogies defeat leaves a considerably smaller claim than the one people usually mean when they invoke fire. What survives is not a theory of LLM risk. It is an observation about institutional form under four specific conditions — non-relinquishment, universal access, frequent local failure, and shared infrastructure — and a two-century record of what forms proved stable under those conditions.

The reduced claim

The fire analogy licenses reasoning about *who responds, funded how, with what authority, constrained by what codes*. It does not license reasoning about what LLM harm is, how fast it spreads, whether it can be stopped, or what a resolved incident looks like. Every subsequent chapter should be read as operating inside that restriction.

PART II

The Institutional Lesson

CHAPTER 4

Why the Fire Department Took the Form It Did

4.1 The insurance brigade and its collapse

Before municipal fire services, fire response in London and in several American cities was organized by property insurers. Insurers maintained brigades, and buildings insured by a given company carried a *firemark* — a metal plaque identifying the underwriter. The arrangement had an internal logic: the party bearing the loss bore the cost of preventing it, and incentives were, on paper, aligned.

The arrangement failed for reasons worth enumerating precisely, because each has a contemporary counterpart.

1. **Coverage gaps.** Uninsured buildings received no response. The party least able to afford protection received none, and the resulting fires threatened insured neighbors anyway.
2. **Externality blindness.** A brigade's duty ran to its policyholder, not to the block. Fire does not respect the boundary of a policy, so socially optimal response and contractually required response diverged systematically.
3. **Duplication and interference.** Multiple brigades arriving at one scene produced competition over water supply and, in the documented American case, outright brawling while structures burned.
4. **Reporting suppression.** An insurer's brigade generated information about losses that bore directly on the insurer's own rates and reputation. The party investigating the fire had a commercial interest in its characterization.

The municipal solution — publicly funded, universal-coverage, geographically rather than contractually assigned, with a separate

investigative function — was not adopted because anyone preferred public provision in the abstract. It was adopted because each of the four failures above is structural and none can be contracted around.

4.2 The contemporary recapitulation

Current LLM incident response is organized substantially as insurance brigades were. Each laboratory maintains a trust and safety function whose duty runs to its own deployment. The mapping is uncomfortably tight.

Brigade failure	Fire instance	LLM instance
Coverage gaps	Uninsured buildings unattended	Harms mediated by open-weight or unaffiliated deployments have no responder
Externality blindness	Duty to policyholder, not block	Duty to platform users, not to third parties harmed by outputs
Duplication and interference	Brigades fighting over hydrants	Incompatible incident taxonomies; no shared dispatch; findings not portable
Reporting suppression	Insurer investigates own loss	Producer characterizes severity of its own model's incident

The fourth row deserves emphasis. Fire investigation was separated from fire suppression and vested in a *fire marshal* with statutory authority precisely because the investigative function cannot be performed credibly by a party with a financial stake in the finding. This separation is, at present, entirely absent from the LLM landscape, and its absence is not a matter of insufficient goodwill. It is structural, in the way the insurance brigades' problem was structural.

4.3 Four properties worth transferring

Design properties of the municipal model

Universal coverage independent of contract.
Response is triggered by the incident's location, not by the victim's relationship to the responder. The LLM translation: incident response reachable by parties who are not customers of any laboratory.

Non-punitive response, separated from enforcement.
Firefighters do not issue citations. This separation is what makes calling them safe, and calling them early is worth more than any enforcement gained by merging the roles. The translation is a reporting channel with credible non-retaliation.

Standing capacity, not surge contracting.
Fire departments exist between fires. Capability is maintained during quiet periods at public expense because it cannot be assembled at the moment of need.

Mutual aid as default.
Neighboring departments respond across jurisdictional lines under standing agreements, using interoperable equipment and a common incident command system. The interoperability is the achievement; the willingness is easy.

4.4 The incident command system as the real export

If one thing from the fire service transfers cleanly, it is not the engine company. It is the Incident Command System — a role-based coordination protocol developed after the California wildfires of the early 1970s, in which the failure was not resources but the inability of many agencies to work a single incident (Bigley and Roberts 2001). ICS specifies roles rather than persons, a defined span of control, a single point of accountability per incident, and a common terminology mandated to the point of banning agency-specific radio codes.

Its transferable virtue is that it is *content-free*. ICS does not encode fire physics. It encodes how heterogeneous organizations coordinate under time pressure with incomplete information and no pre-existing command relationship — which is an exact description of a cross-platform LLM incident involving a model producer, a deploying company, an affected third party, and a regulator who have never worked together.

Notably, ICS survives every one of the seven disanalogies in Chapter 3. It makes no assumption about extinguishment, about the separation of response and propagation channels, or about the naturalness of the harm kind. This is what a genuinely transferable institutional artifact looks like: it transfers because it was never about fire in the first place.

CHAPTER 5

Codes, Marshals, and the Ex Ante Complement

5.1 Codes are written in loss

The uncomfortable regularity of fire regulation is that essentially every significant provision postdates a disaster that made it obvious. Exit doors swinging outward, unobstructed egress, sprinklers in high-occupancy buildings, compartmentation requirements: each traces to a specific event with a specific death toll. Fire codes are a sedimentary record of failures that were, in retrospect, foreseeable.

Two readings of this fact are available, and both are correct.

The pessimistic reading is that regulation is reactive by nature, that the political economy of prevention cannot mobilize without bodies, and that anyone expecting anticipatory LLM governance is expecting something without historical precedent in comparable domains.

The optimistic reading is subtler and, I think, more useful. The codes are reactive in *origin* but prospective in *operation*. Once written, a code applies to every building constructed thereafter, including buildings whose failure modes nobody has yet observed. The mechanism converts a particular past disaster into general future constraint. The design question is therefore not "how do we regulate without disasters" but "how do we ensure that when incidents occur, they are converted into general constraint rather than particular remediation." That is a question about incident investigation infrastructure, and it is answerable now.

The conversion function

The value of an incident is the generality of the constraint it produces. This requires three things the LLM field currently lacks: a mandatory reporting threshold, an investigative body with access and without stake, and a publication norm that makes findings binding on parties other than the one investigated. Aviation has all three (Dekker 2014); the fire service has approximations of all three; LLM deployment has none.

5.2 The marshal: authority without suppression duty

The fire marshal is a distinct office from the fire chief, and the distinction is instructive. The marshal investigates origin and cause, inspects for code compliance, and in most jurisdictions holds police powers including the authority to compel entry and, in extreme cases, to order a structure vacated. The marshal does not fight fires.

The separation accomplishes two things simultaneously. It preserves the suppression service's non-punitive character — one may call the fire department without summoning an inspector — while ensuring that enforcement authority exists somewhere and is not dependent on the cooperation of the party being investigated.

The translation is not difficult to state, though it is politically difficult to achieve: a body empowered to examine incidents involving deployed models, with access to logs and, where warranted, to weights and training data under confidentiality; with authority to compel that access; and organizationally separate from any body providing incident response assistance. The current arrangement, in which the party with access, the party with expertise, and the party with liability are the same party, has no analogue in any mature safety-critical domain.

5.3 Occupancy classification and the refusal of uniform rules

A useful and under-appreciated feature of fire codes is that they are not uniform. Requirements are indexed to *occupancy classification*:

assembly, business, educational, factory, high-hazard, institutional, mercantile, residential, storage. A chemical plant and a bookstore face different rules not because one is more valuable but because occupant density, egress capacity, fuel load, and occupant ability to self-evacuate differ.

The institutional occupancy category is the most instructive. Hospitals and prisons face the strictest requirements not because they burn more readily but because occupants *cannot self-evacuate*. Defend-in-place strategy, with heavy compartmentation and full sprinkler coverage, follows directly from that single fact about the occupants.

The corresponding LLM question is not "how capable is this model" but "can the affected party exit the interaction?" A user who can close the tab is in a mercantile occupancy. A patient whose triage was model-mediated, a defendant whose risk score was model-generated, an applicant screened by a model — these are institutional occupancies, and they warrant defend-in-place requirements regardless of the model's capability level.

Classification	**Governing fact**	**Deployment analogue**
Mercantile	Occupants transient, self-evacuating	Consumer chat; user may exit at will
Assembly	High density, egress-limited	Recommendation and ranking at population scale
High-hazard	Fuel load creates hazard to neighbors	Capability with third-party uplift potential
Institutional	Occupants cannot self-evacuate	Adjudicative, clinical, custodial, and screening uses

This reframing does real work. It shifts the regulatory question from model properties — which are contested, hard to measure, and change with each release — to deployment context, which is legible, stable, and known to the deployer. Fire codes regulate buildings, not chemistry.

CHAPTER 6

The Suppression Paradox

6.1 A century of getting it exactly wrong

For most of the twentieth century, United States wildland fire policy aimed at total suppression. The doctrine was codified after the 1910 Big Blowup and reached its clearest expression in the so-called 10 a.m. policy: every reported fire was to be controlled by ten o'clock the morning following its report. The policy was competently executed and enormously successful at its stated objective.

It was also the direct cause of the fire regime that now produces the largest and most destructive fires in the region's recorded history.

The mechanism is well established and not seriously disputed (Pyne 1982; Stephens and Ruth 2005). Many Western forest types evolved under frequent, low-intensity surface fire that consumed litter and small stems while leaving mature trees. Suppressing these fires allowed fuel to accumulate continuously across decades and permitted dense understory growth that provides ladder fuels connecting the surface to the canopy. The resulting stands do not burn as their predecessors burned. They burn as crown fires, at intensities against which suppression is ineffective and under which the ecosystem does not regenerate to its prior state.

The paradox, stated

Suppressing every small failure does not reduce total failure. It converts a distribution of many small events into a distribution of rare, uncontrollable ones — and it does so while producing continuous evidence of success, because the small fires that would have revealed the accumulating problem are precisely the ones being suppressed.

6.2 Fuel accumulation in deployment

The transfer is not automatic, and I want to be careful with it, because this is the point at which the fire analogy is most often invoked loosely.

The claim is *not* that LLM harms should be permitted to occur for their tonic effect. The claim is narrower and concerns what accumulates when failures are suppressed rather than surfaced.

Untested reliance.
Each patched failure that produces no visible incident permits users to extend reliance further. Reliance calibrated by observed failure is bounded; reliance calibrated by the absence of observed failure is unbounded. The accumulated quantity is unwarranted trust, and it is a ladder fuel.

Undiagnosed brittleness.
A failure suppressed by a filter is a failure whose mechanism was not investigated. The underlying disposition persists, correlated across every deployment of the same base model. Correlated latent failure across a monoculture is the dense even-aged stand.

Atrophied institutional response.
Organizations that have never handled an incident do not have incident response capability regardless of what their documentation asserts. Capability decays without exercise, which is why fire departments drill.

Deferred policy formation.
The absence of visible incidents removes the political basis for building the institutions of Chapters 4 and 5. Codes are written in loss; where loss is invisible, codes go unwritten while exposure grows.

6.3 Prescribed burning as institutional practice

The response of contemporary fire ecology is not to abandon suppression. It is to pair suppression with deliberate, bounded, expert-supervised introduction of the very process being suppressed. Prescribed burns are conducted under specified weather windows,

with prepared containment lines, by trained crews, with an explicit prescription defining acceptable intensity and defined abort conditions.

The features worth naming are the ones that distinguish a prescribed burn from arson: it is *deliberate*, *bounded*, *instrumented*, *reversible within a defined envelope*, and *conducted by parties accountable for the outcome*. Red-teaming and staged deployment are gestures in this direction. They fall short in a specific and diagnosable way: they are almost always conducted by, or under contract to, the party whose fuel load is being assessed, and the results are almost never published in a form that permits comparison across stands.

The deeper lesson is about the institution rather than the technique. It took the United States Forest Service roughly seventy years to accept that its foundational doctrine was producing the outcome it existed to prevent, and acceptance came only after the evidence had become impossible to characterize otherwise. The service was staffed by competent people acting in good faith on a doctrine that was correct in the small and catastrophically wrong in the aggregate. There is no reason to believe the current generation of LLM safety practice is less susceptible to a doctrine with the same shape — and the shape is recognizable: a metric that improves as the underlying condition worsens.

The diagnostic question

For any safety practice, ask: *if this practice were producing accumulating latent risk, would the metrics I track distinguish that case from success?* Where the answer is no, the practice is in the suppression regime, and the appropriate response is to instrument for the accumulating quantity rather than to trust the metric.

PART III

Design Consequences

CHAPTER 7

Firebreaks, Compartments, and Design Commitments

7.1 Compartmentation before suppression

A modern building's fire protection is layered, and the layers are ordered by reliability rather than by prominence. Passive measures come first: fire-rated assemblies, self-closing doors, shaft enclosures, draftstopping in concealed spaces. These require no detection, no power, no water, and no human decision. Active systems — detection, alarm, sprinklers — come second. Human response comes last.

The ordering is deliberate and reflects a hard-won judgment: the most reliable protection is the one that does not have to work, because it is simply the geometry of the building. A two-hour rated wall does not fail because a sensor was miscalibrated or a crew was delayed.

Contemporary LLM safety inverts this ordering. The dominant investment is in the model's behavioral disposition — an active, learned, distributional property that must correctly classify a situation it may never have seen, at inference time, under adversarial pressure. That is the sprinkler head, not the wall. It is valuable. It should not be first.

Commitment 1: Passive containment precedes behavioral alignment

Before asking whether a model will decline to do harm, ask what harm remains architecturally available to it. Scope tool permissions to the task rather than the session. Bound the blast radius of any single agent action. Make irreversible oper-

ations require an out-of-band confirmation that the model cannot itself supply. These constraints hold regardless of what the model decides, which is exactly the property a rated wall has and a disposition does not.

7.2 Defensible space and the third-party problem

Wildland-urban interface doctrine holds the property owner responsible for maintaining defensible space — clearance, ember-resistant venting, non-combustible roofing — in a zone extending well beyond the structure. The doctrine exists because the fire service cannot defend structures that have not been made defensible, and because one unmaintained property determines outcomes for its neighbors.

The LLM translation concerns parties who never chose to interact with the model at all: the person described in a generated document, the author whose work is paraphrased, the applicant screened without notice. Fire-code thinking treats hazard to neighbors as the primary object of regulation — high-hazard occupancy classification exists almost entirely to protect adjacent structures — while consent-based frameworks address only the policyholder.

Commitment 2: Regulate by hazard to non-parties

The relevant question for a deployment is not the consenting user's exposure but the exposure of parties with no relationship to the deployment and no means of exit. Where that exposure is material, institutional occupancy requirements from Chapter 5 apply: defend-in-place, recorded provenance, and an appeal path that does not route through the deploying party.

7.3 Interoperability as the unglamorous prerequisite

The Great Baltimore Fire of 1904 burned for over thirty hours across roughly seventy city blocks. Engines responded from Wash-

ington, Philadelphia, and New York. Many could not connect to Baltimore's hydrants, because hose couplings were not standardized — thread pitch and diameter varied by city. Crews stood beside working water supplies they could not use.

The subsequent standardization effort is the least glamorous and most consequential item in this treatise. The LLM equivalents are equally unglamorous: a shared incident taxonomy, common severity definitions, portable evaluation results, a provenance format that survives transit between platforms, and a disclosure channel that a researcher can use without learning each laboratory's bespoke process.

Commitment 3: Standardize couplings before building engines

Coordination infrastructure has no demo, generates no benchmark improvement, and is invisible until the moment it is decisive. It should nonetheless precede capability-specific safety investment, because its absence invalidates that investment at exactly the moment cross-organizational response is required.

7.4 Instrumenting for the time-indexed defeater

Disanalogy 2 of Chapter 3 — that fire has no interest in the fire department — was classified as survivable on the grounds that present LLM harm is not adversarially optimized against responders. That classification is a claim about the present, and claims about the present expire.

Commitment 4: Monitor for the defeater, not just the harm

Track leading indicators that the disanalogy is failing: outputs that adapt to the presence of monitoring, harm patterns that shift in response to mitigation faster than adversary iteration would explain, incidents in which the model's behavior differs measurably under observation. If these appear, the fire framework should be retired rather than extended — the

correct successor vocabulary is epidemiological or adversarial, not architectural.

7.5 Summary of commitments

#	Commitment	Fire-service source
1	Passive containment precedes behavioral alignment	Compartmentation ranked above suppression
2	Regulate by hazard to non-parties	High-hazard occupancy; defensible space
3	Standardize couplings before building engines	Baltimore 1904; hose thread standardization
4	Monitor for the defeater	Reading the ceiling; pre-flashover indicators
5	Separate investigation from response	Fire marshal distinct from fire chief
6	Instrument for accumulating latent risk	Prescribed burning; the suppression paradox

CHAPTER 8

What Would Falsify This

8.1 The argument in one paragraph

Fire is the standing case of a technology civilization keeps and does not control. What we built for it was not a better fire but an institution: a publicly funded, universal-coverage, non-punitive response service, paired with prospective material codes indexed to occupancy, with investigation separated from response, and coordination protocols that assume heterogeneous organizations meeting for the first time under time pressure. That institutional package is the transferable content of the fire analogy. The physics is not transferable, the extinguishment model is not transferable, and the current configuration — in which producers respond to their own incidents — recapitulates an arrangement that fire management abandoned as structurally unworkable in the nineteenth century.

8.2 Conditions under which the argument fails

Stating these is the point of the chapter.

If harm ceases to be incident-shaped.
The dispatch model requires detectable events with identifiable scenes. If the dominant harms prove to be diffuse epistemic and economic effects with no incident structure, the fire department is the wrong institution and public health is the right one. This is Disanalogy 3 becoming dominant, and I regard it as the most likely failure mode of the argument.

If the defeater arrives.
If LLM harm becomes systematically optimized against responders, architectural vocabulary — codes, compartments, occupancies — fails, because the environment stops being a static thing

to be engineered. Commitment 4 exists to detect this.

If capability concentration reverses the scaling premise. The fire model was chosen over the nuclear model on grounds of endpoint count. If frontier capability concentrates in a handful of sites while widely available systems become genuinely bounded, licensure governance becomes appropriate and this treatise's premise fails.

If the suppression paradox does not obtain. The Chapter 6 argument requires that suppressed failures leave an accumulating residue. If patched LLM failures genuinely remove the underlying disposition rather than masking it, there is no fuel load, and aggressive suppression is simply correct. This is an empirical question about the relationship between behavioral mitigation and latent capability, and it is answerable.

8.3 The residual

What remains after those conditions are granted is modest but not trivial. It is a claim that certain institutional forms are stable under certain structural conditions, that those conditions currently obtain, and that the forms in question were not obvious in advance — they were learned at considerable cost over two centuries by people who were not smarter than we are and who had the advantage of failures that were visible.

That last clause is the one I would leave a reader with. The fire service's institutional knowledge was purchased with burned cities. Its availability to us is a windfall, and windfalls have a characteristic failure mode: they are discounted precisely because they were not paid for. The Baltimore hose couplings are the emblem of the whole essay. Nothing about that failure was technically difficult, nothing about it was unforeseeable, and it required a city to burn before anyone standardized a thread pitch.

Closing

The question is not whether LLMs are like fire. They are not, in most of the ways that matter, and Chapter 3 says so at length. The question is whether we will need something like a fire department — and whether we will build it before or after the fire that makes the case unanswerable. The historical record on that question is not encouraging, but it is at least clear, and clarity about our own tendencies is the one resource that reliably precedes the loss rather than following it.

Concordance of Terms

The table below collects the mappings used throughout, with the chapter in which each is introduced and a note on how far it should be pressed. The *Load* column indicates whether the mapping is structural (S), heuristic (H), or explicitly limited (L).

Fire term	Deployment analogue	Load	Limit
Fuel	Latent capability	S	Cannot be removed without removing utility
Oxidizer	Deployment surface	S	The controllable leg; underweighted in practice
Ignition	Eliciting context	S	Unbounded source space; filtering is partial
Chain reaction	Autocatalytic propagation	H	No conserved quantity; propagation is not consumption
Flashover	Nonlinear capability transition	H	Threshold location unknown and possibly deployment-specific
Extinguishment	—	L	**No analogue.** Weights persist; capability reproduces
Firebreak	Architectural isolation	S	Information is not fuel-limited; geometric intuition misleads
Compartmentation	Scoped permissions; blast-radius bounds	S	Strongest transfer in the essay
Occupancy class	Deployment context class	S	Indexed to exit capability, not model capability
Defensible space	Non-party hazard mitigation	S	Addresses parties outside any consent relation

Fire term	Deployment analogue	Load	Limit
Fire marshal	Independent incident investigator	S	Requires compelled access; currently absent
Mutual aid	Cross-organization response agreements	S	Requires interoperability first
Hose coupling	Shared taxonomy, portable findings	S	Baltimore 1904; the emblem case
Prescribed burn	Instrumented deliberate failure	H	Fails when conducted by the party being assessed
Suppression paradox	Accumulating latent brittleness	H	Empirically contingent; see Ch. 8
Incident Command System	Cross-party coordination protocol	S	Survives all seven disanalogies; content-free by design
Insurance brigade	Producer-run trust and safety	S	Transfers as warning; four documented failure modes

References

Bigley, G. A. and K. H. Roberts (2001). "The Incident Command System: High-Reliability Organizing for Complex and Volatile Task Environments." In: *Academy of Management Journal* 44.6, pp. 1281–1299. DOI: 10.5465/3069401.

Brynjolfsson, E., D. Rock, and C. Syverson (2021). "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." In: *American Economic Journal: Macroeconomics* 13.1, pp. 333–372. DOI: 10.1257/mac.20180386.

Dekker, S. (2014). *The Field Guide to Understanding 'Human Error'*. 3rd ed. Farnham, Surrey: Ashgate.

Eisenstein, E. L. (1979). *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*. 2 vols. Cambridge: Cambridge University Press.

Pyne, S. J. (1982). *Fire in America: A Cultural History of Wildland and Rural Fire*. Princeton, New Jersey: Princeton University Press.

Stephens, S. L. and L. W. Ruth (2005). "Federal Forest-Fire Policy in the United States." In: *Ecological Applications* 15.2, pp. 532–542. DOI: 10.1890/04-0545.