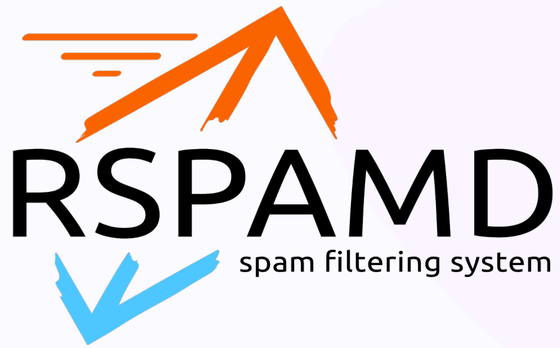
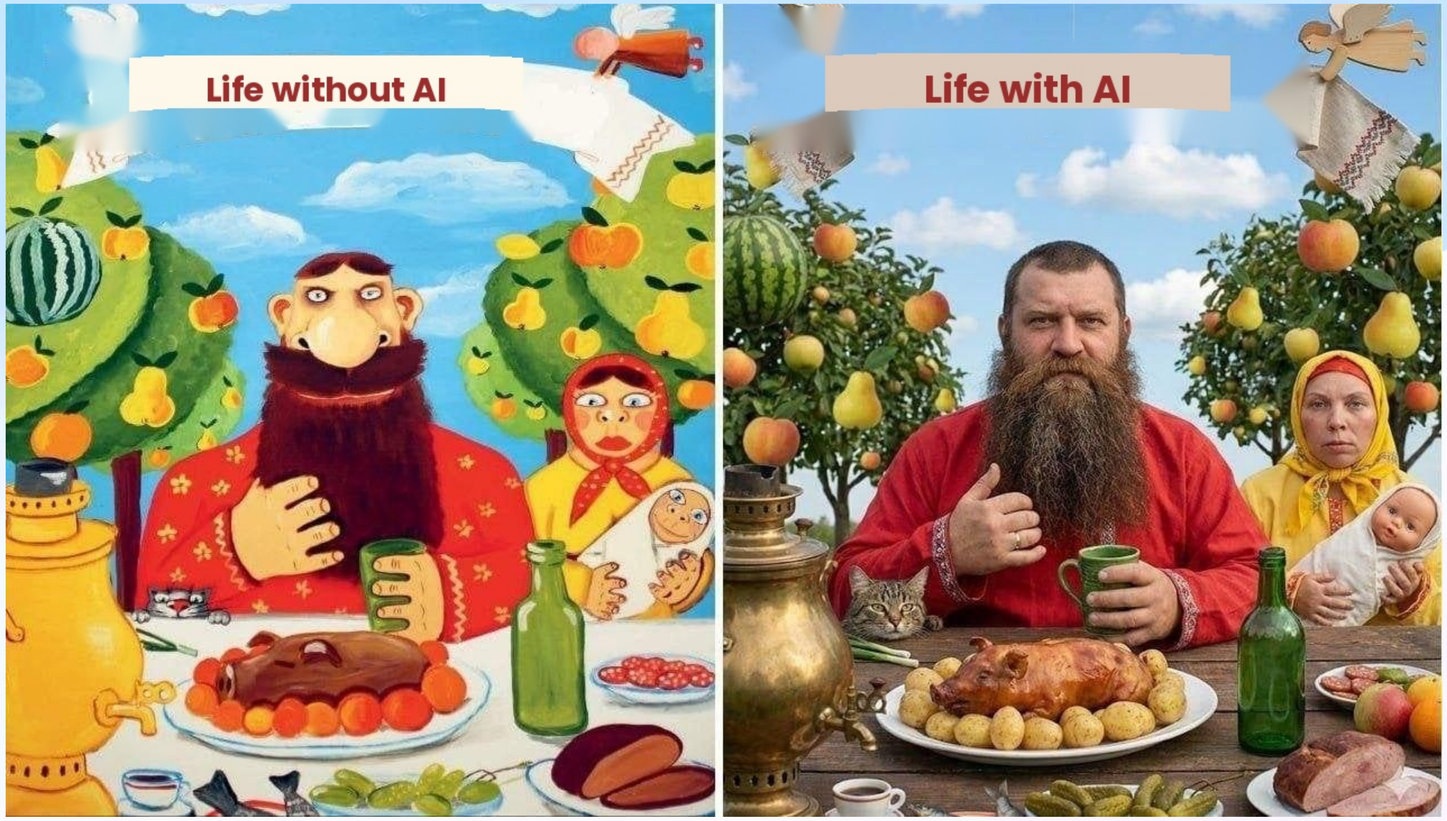
Rspamd
A Year of Features and LLM-Assisted Development

FOSDEM 2026 • Vsevolod Stakhov



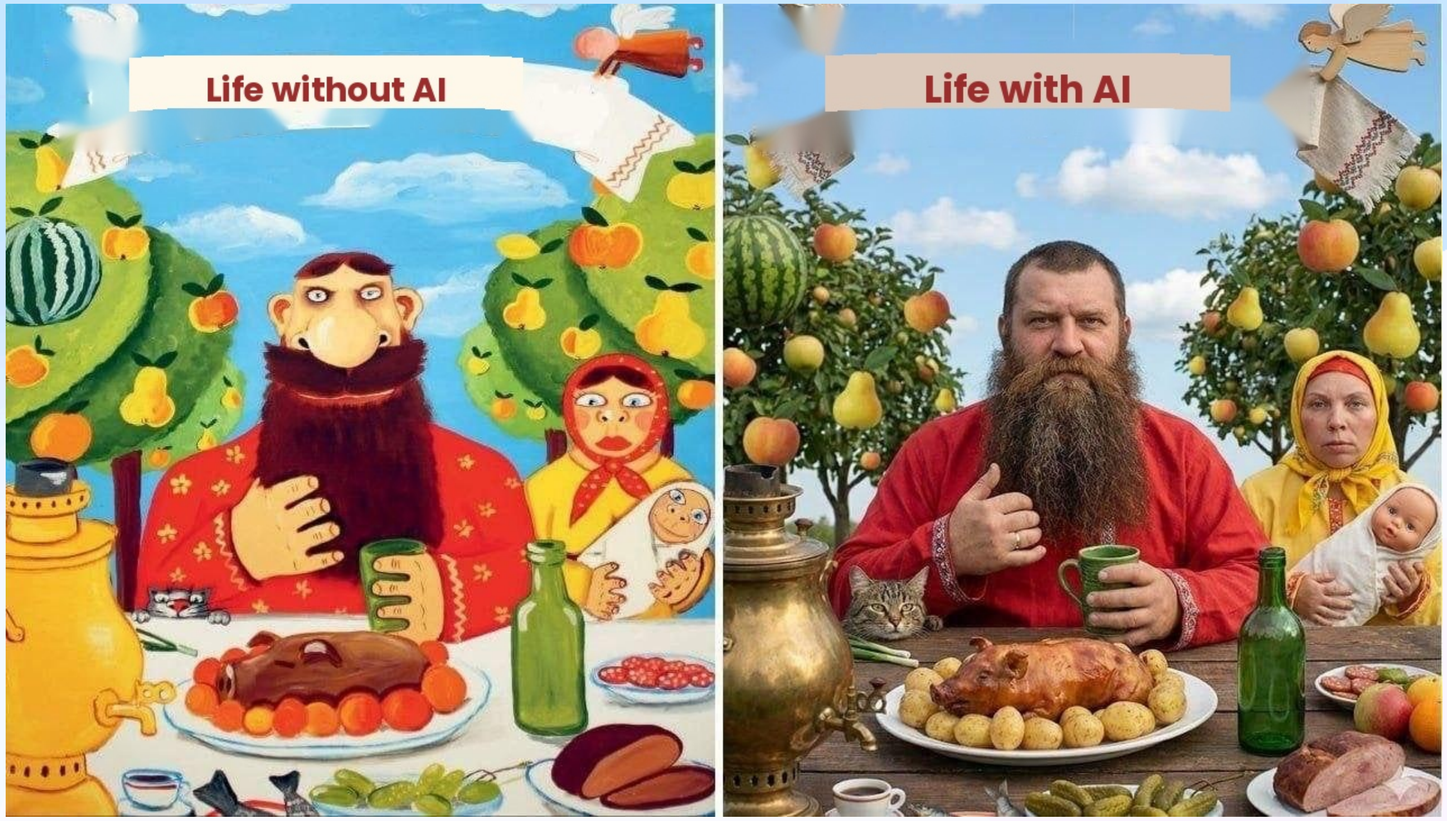
Part 1: LLM-Assisted Development

From Little Helper to Big Thing



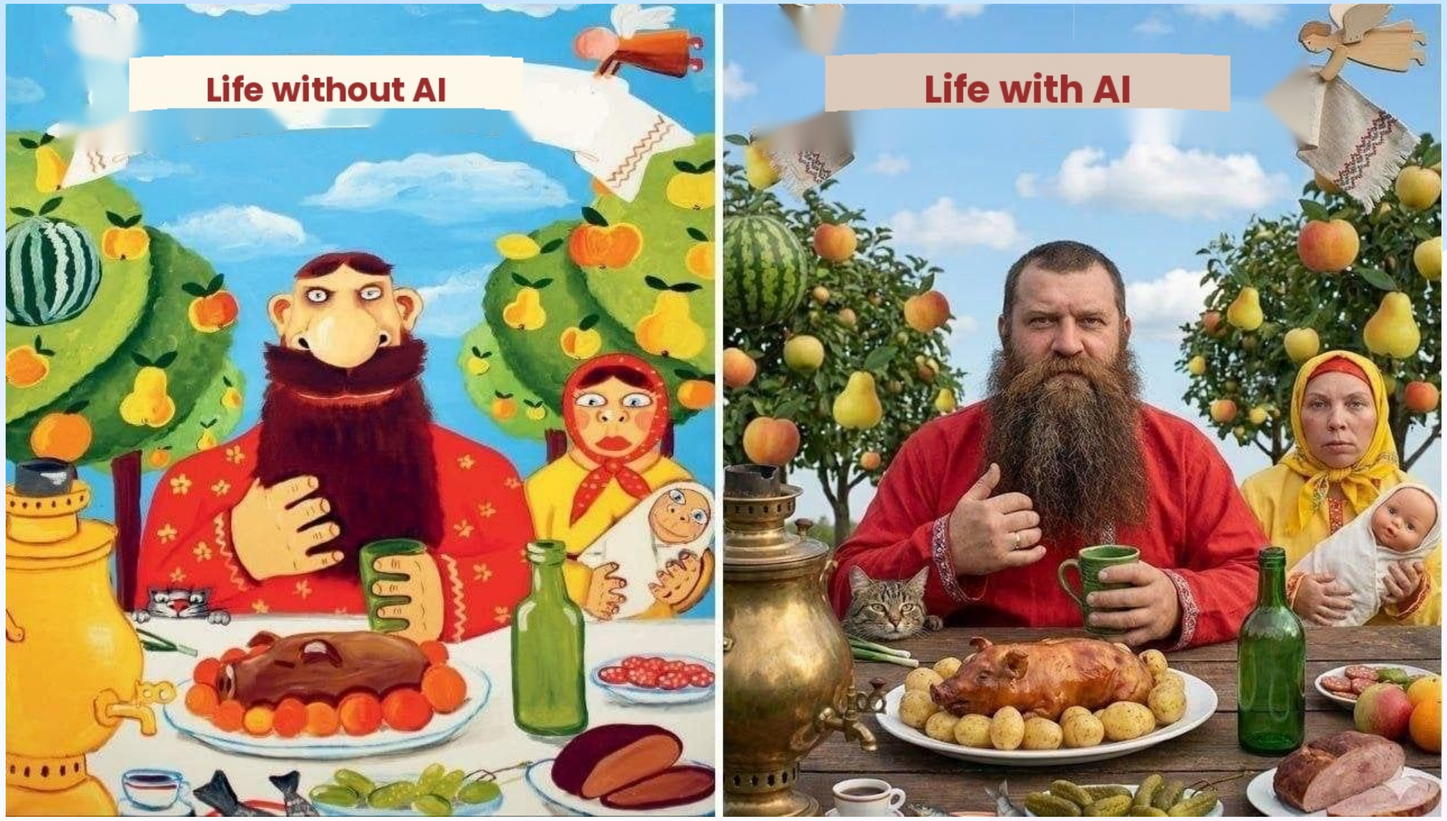
From Little Helper to Big Thing

- 2024: Code Completion



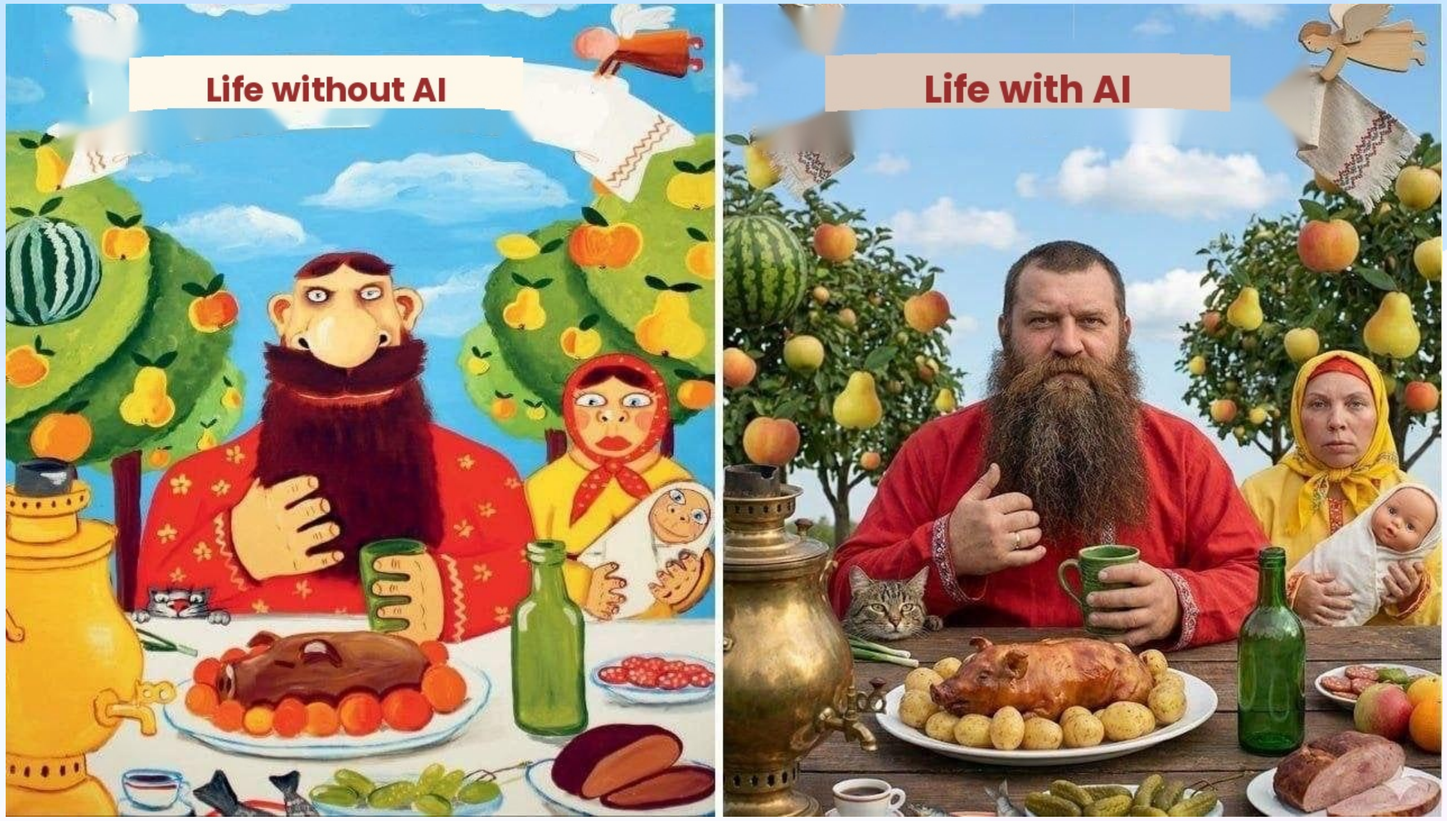
From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions



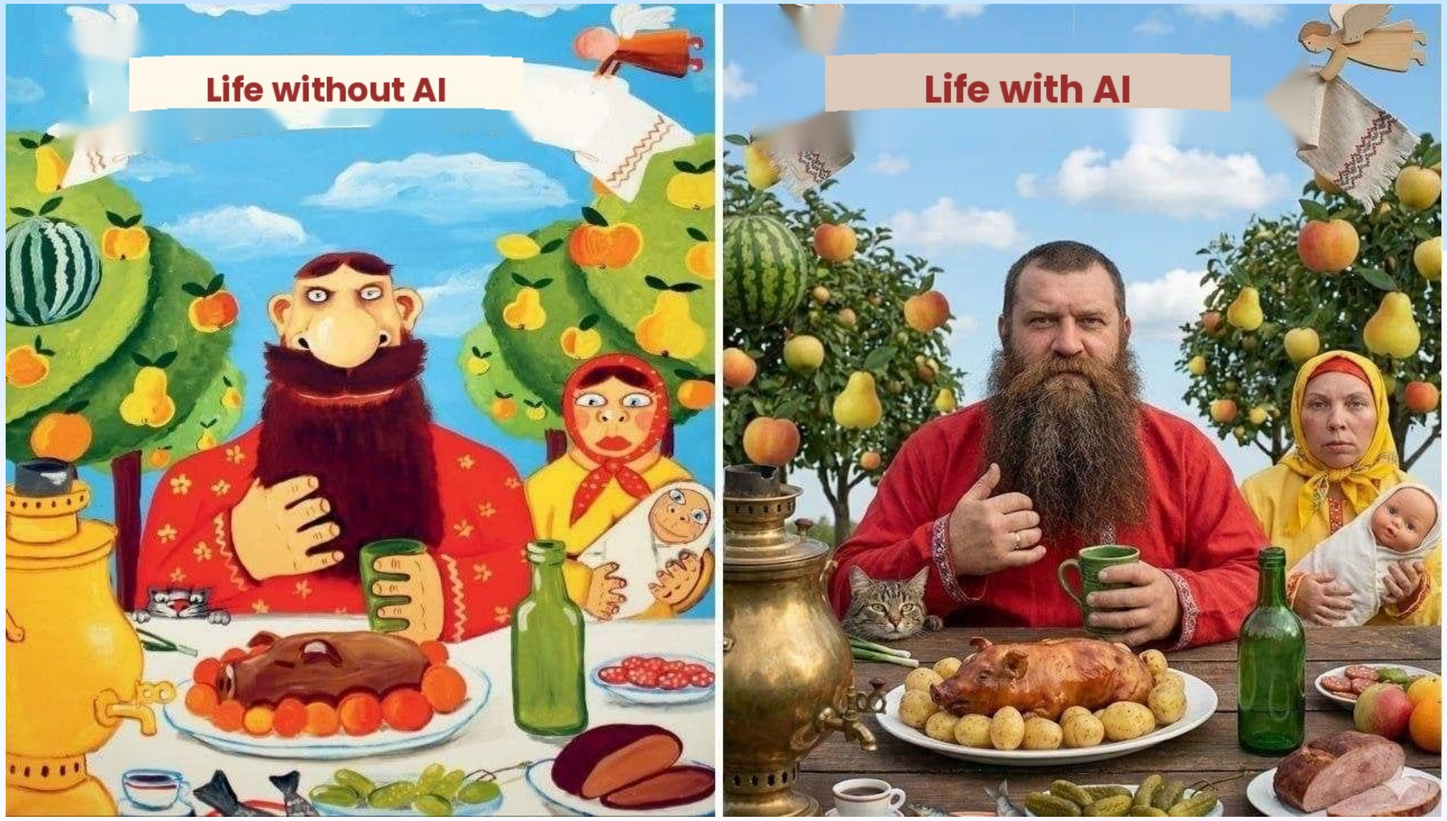
From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits



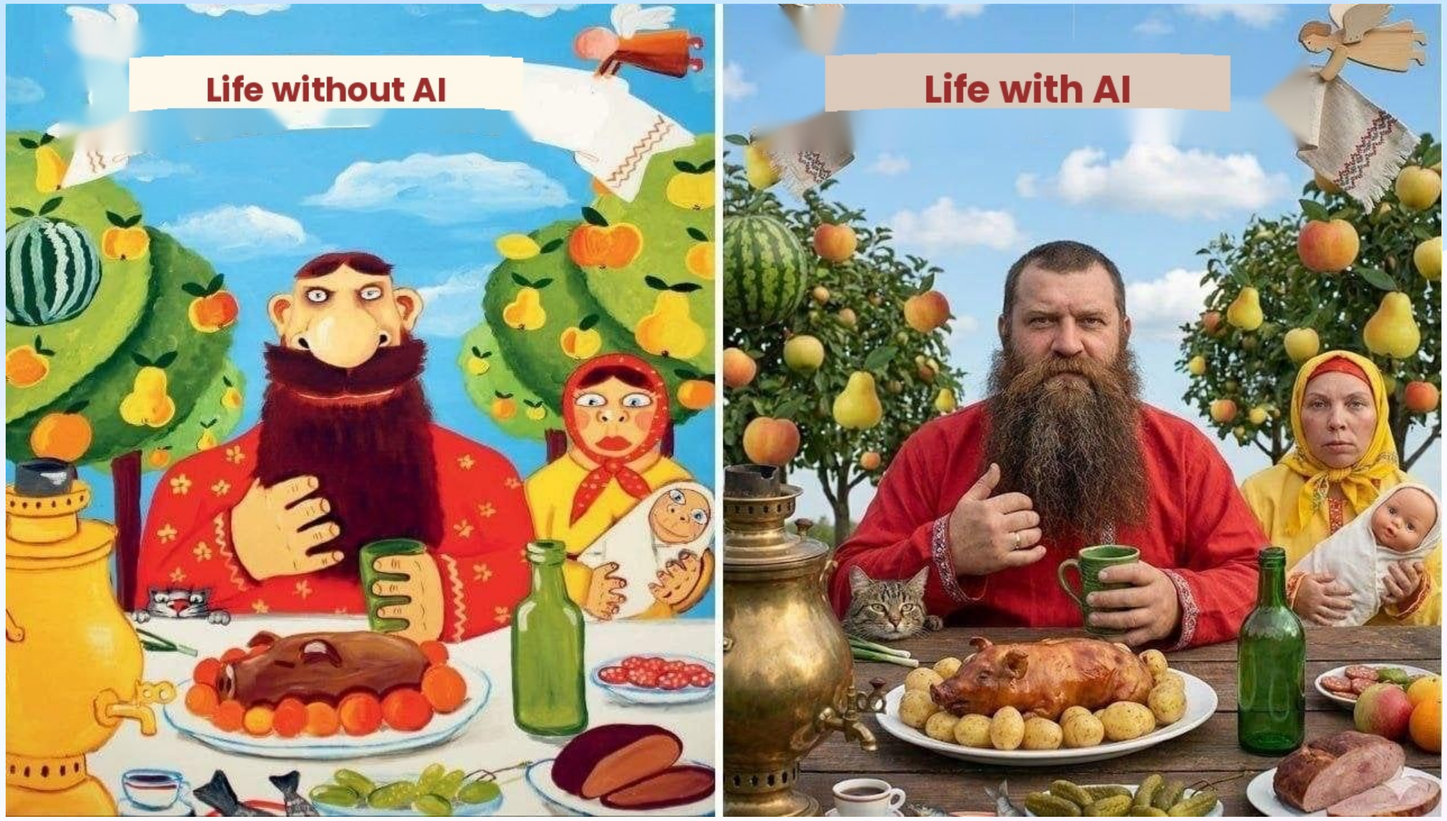
From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup



From Little Helper to Big Thing

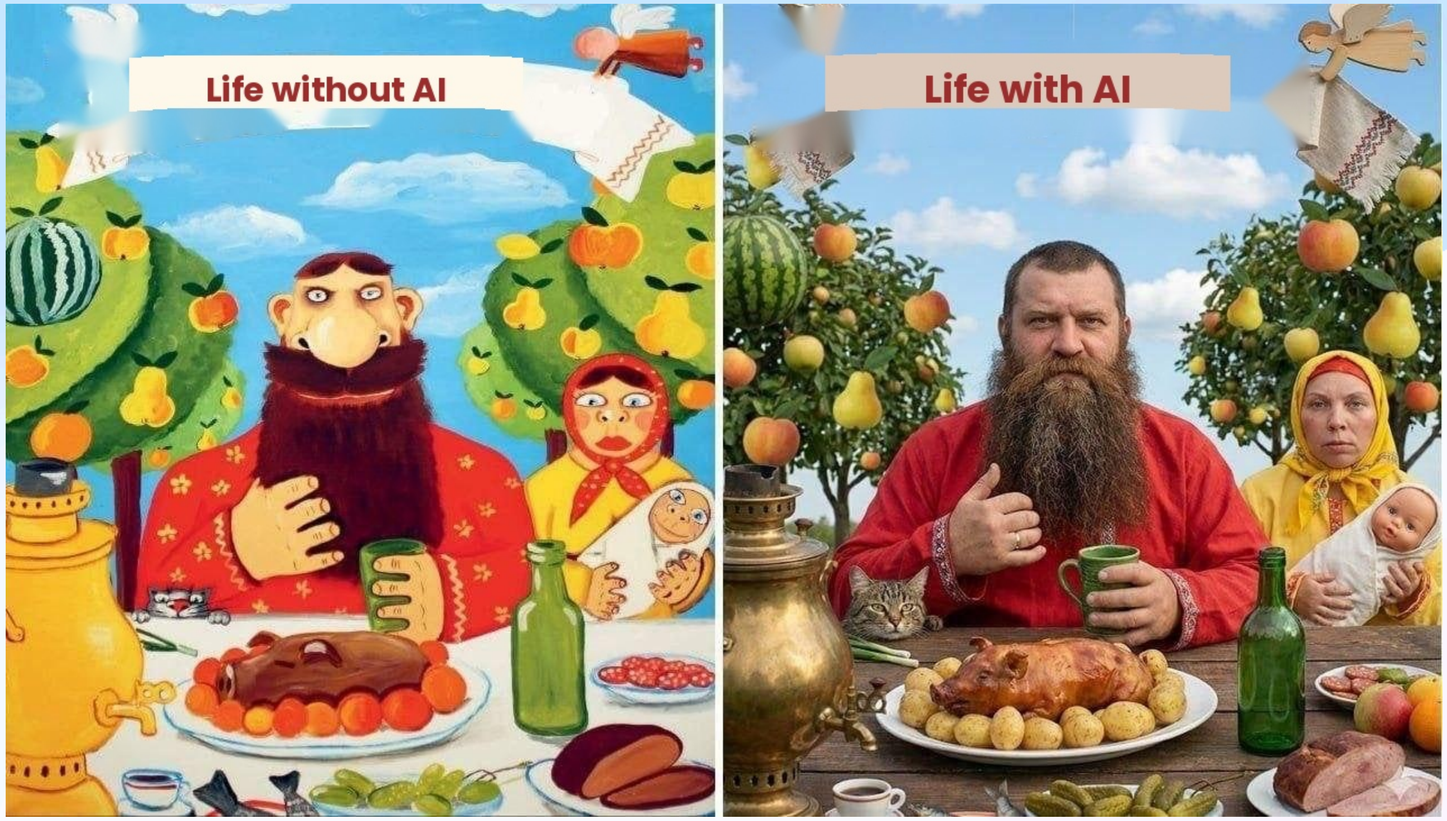
- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup



From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup

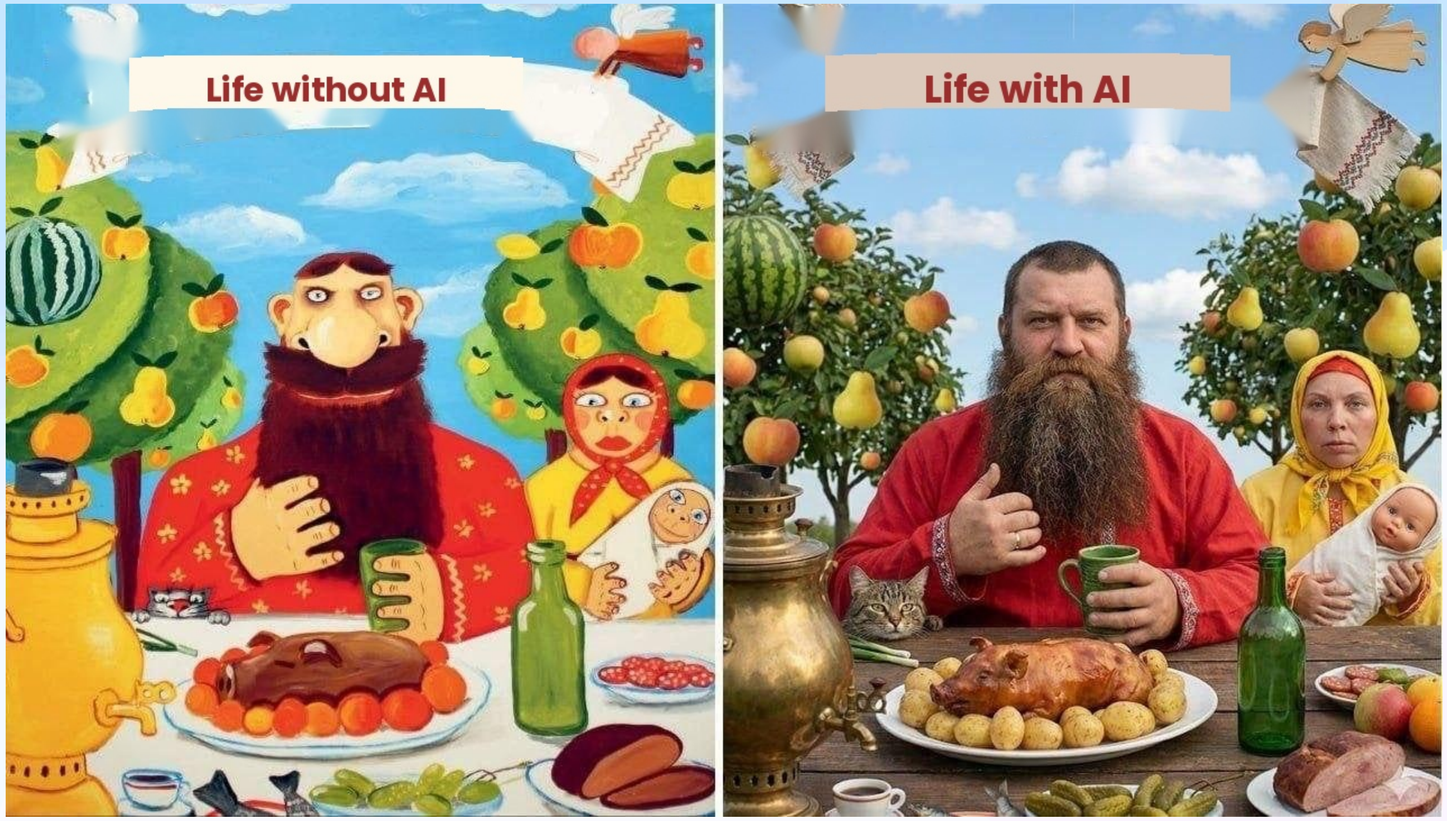
- 2025: Development Orchestrator



From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup

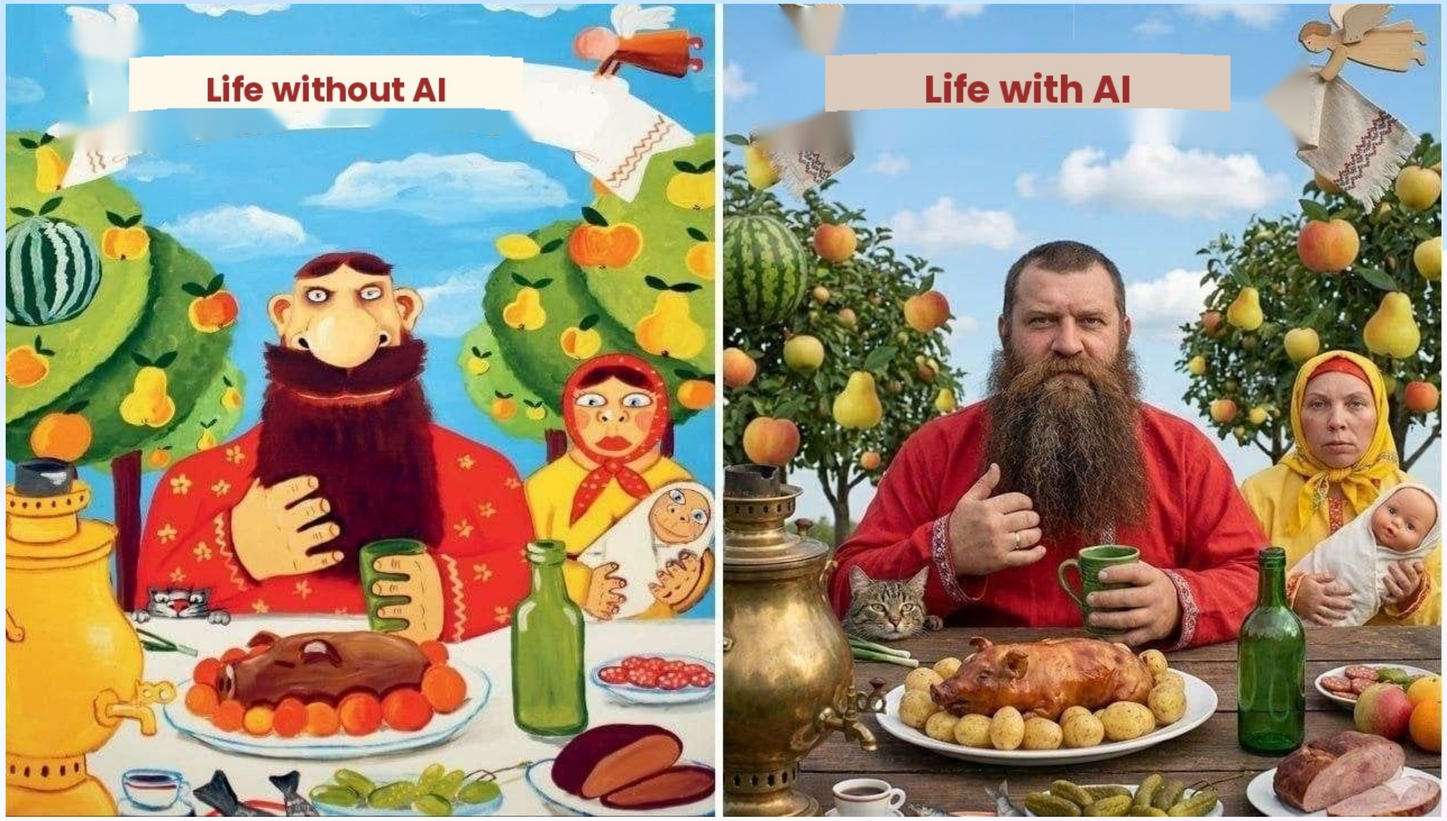
- 2025: Development Orchestrator
- Full architecture design



From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup

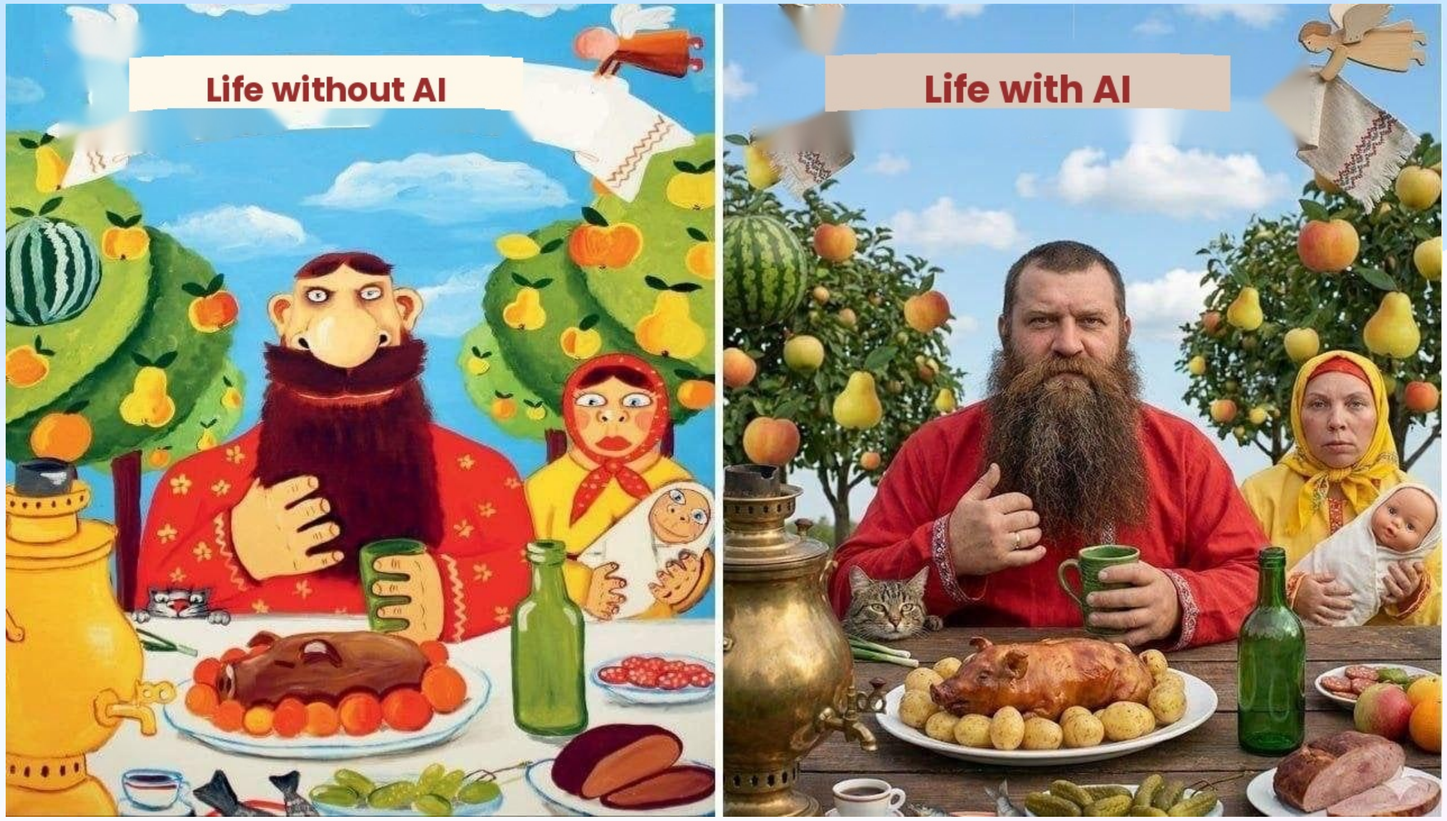
- 2025: Development Orchestrator
- Full architecture design
- Multi-file refactoring



From Little Helper to Big Thing

- 2024: Code Completion
- Autocomplete suggestions
- Single-file edits
- Documentation lookup

- 2025: Development Orchestrator
- Full architecture design
- Multi-file refactoring
- Complete project scaffolding



Multi-LLM Architecture

LLM Development: Balance Sheet

✅ Gains

LLM Development: Balance Sheet

✅ Gains

- 10x prototyping speed

LLM Development: Balance Sheet

✅ Gains

- 10x prototyping speed
- Explore more design options and experiments

LLM Development: Balance Sheet

✅ Gains

- 10x prototyping speed
- Explore more design options and experiments
- Better documentation

LLM Development: Balance Sheet

✅ Gains

- 10x prototyping speed
- Explore more design options and experiments
- Better documentation
- Automated boilerplate: devops, side projects (dashboards, client libraries, rules, analytics, RAG etc)

LLM Development: Balance Sheet

✅ Gains

- 10x prototyping speed
- Explore more design options and experiments
- Better documentation
- Automated boilerplate: devops, side projects (dashboards, client libraries, rules, analytics, RAG etc)
- Bottom line: ~10x for prototyping, ~2x for production

Real Disadvantages

Real Disadvantages

- ⚠️ Context Loss

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.
- ⚠️ Expert Required

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.
- ⚠️ Expert Required
- Domain knowledge essential to validate outputs and guide decisions.

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.
- ⚠️ Expert Required
- Domain knowledge essential to validate outputs and guide decisions.
- ⚠️ Less Focused

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.
- ⚠️ Expert Required
- Domain knowledge essential to validate outputs and guide decisions.
- ⚠️ Less Focused
- Easy to chase tangents. Exploration vs completion tension.

Real Disadvantages

- ⚠️ Context Loss
- Long conversations lose coherence. Requires periodic resets.
- ⚠️ Expert Required
- Domain knowledge essential to validate outputs and guide decisions.
- ⚠️ Less Focused
- Easy to chase tangents. Exploration vs completion tension.
- Less value for our own expert knowledge (not always)

Not Real Problems (Fixable)

Not Real Problems (Fixable)

- Code Bloat

Not Real Problems (Fixable)

- Code Bloat
- → Refactoring passes fix this; LLM can help refactor too

Not Real Problems (Fixable)

- Code Bloat
- → Refactoring passes fix this; LLM can help refactor too
- LLM Slope

Not Real Problems (Fixable)

- Code Bloat
- → Refactoring passes fix this; LLM can help refactor too
- LLM Slope
- → Fresh context windows; summarize and restart

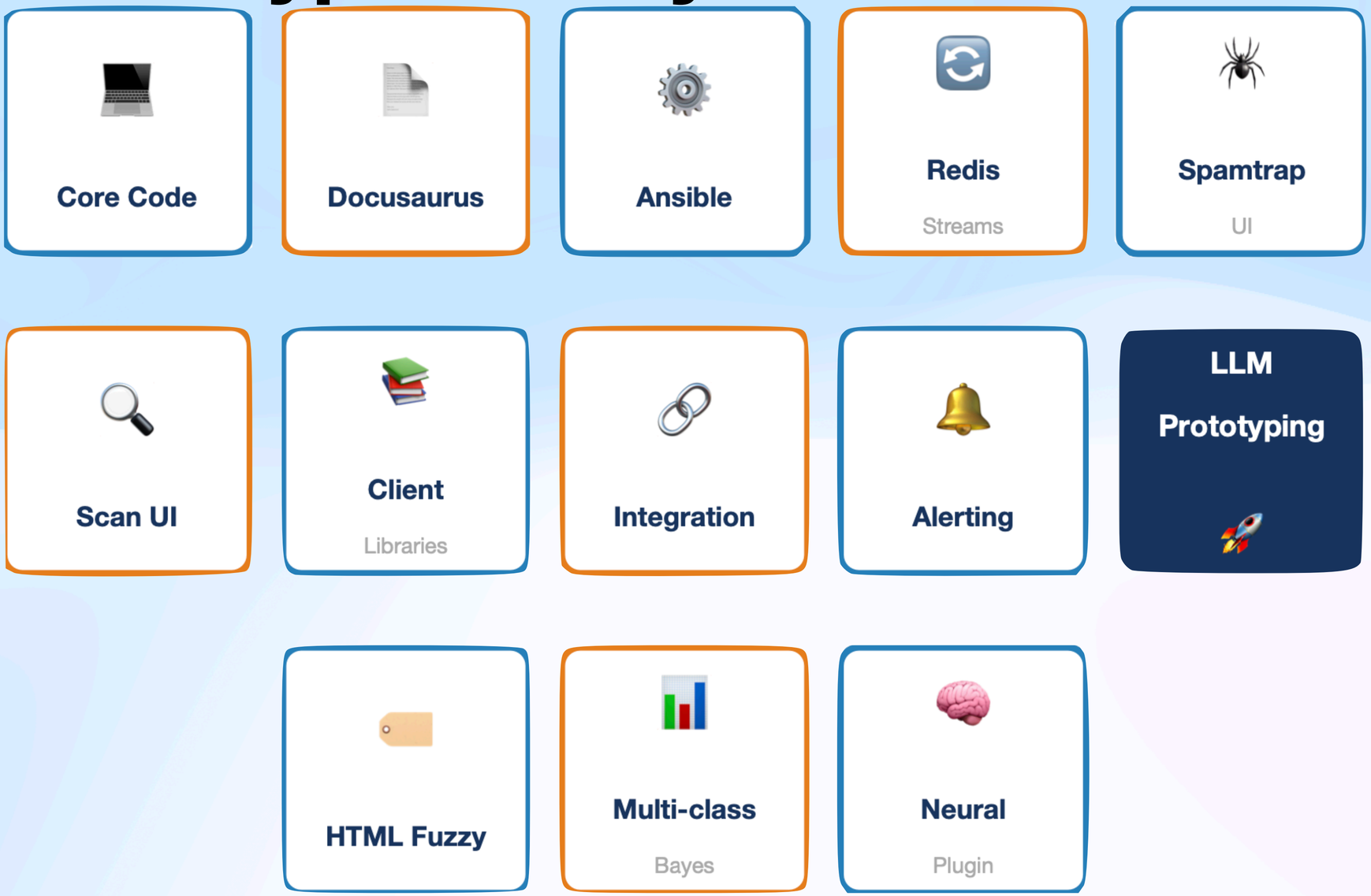
Not Real Problems (Fixable)

- Code Bloat
- → Refactoring passes fix this; LLM can help refactor too
- LLM Slope
- → Fresh context windows; summarize and restart
- Hallucinations

Not Real Problems (Fixable)

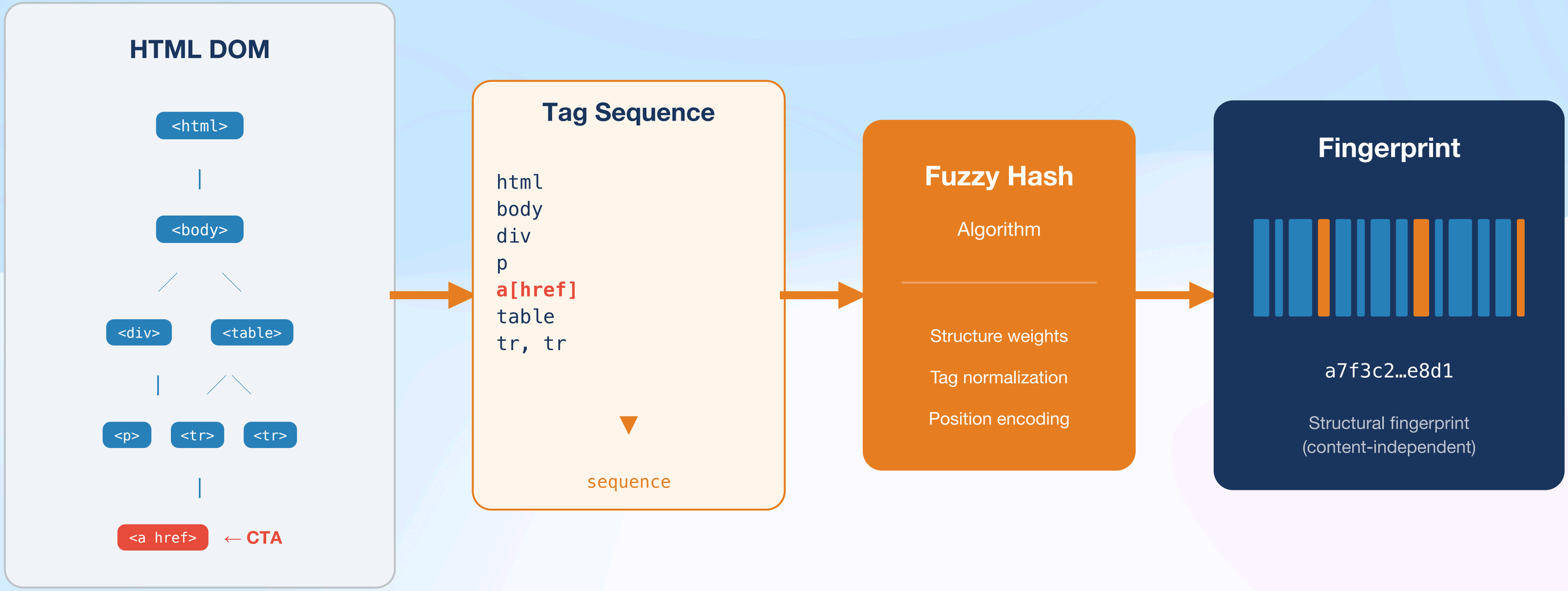
- Code Bloat
- → Refactoring passes fix this; LLM can help refactor too
- LLM Slope
- → Fresh context windows; summarize and restart
- Hallucinations
- → Expert review catches; tests validate; CI enforces

Quick Prototypes Everywhere



Part 2: Technical Projects

HTML Fuzzy Hashing: The Algorithm



HTML Fuzzy: Use Cases

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam
- Lottery scams

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam
- Lottery scams
- Phishing campaigns

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam
- Lottery scams
- Phishing campaigns
- Template-based fraud

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam
- Lottery scams
- Phishing campaigns
- Template-based fraud

HTML Fuzzy: Use Cases

- ✅ Whitelisting Brands
- Shipping notifications
- Order confirmations
- Banking alerts
- 🚫 Blacklisting Spam
- Lottery scams
- Phishing campaigns
- Template-based fraud

- Plus: CTA link analysis distinguishes phishing from legitimate

Multi-class Bayes

Multi-class Bayes

- Before: Binary World

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

- After: Multi-class

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

- After: Multi-class
- SPAM | HAM | NEWSLETTER | TRANSACTIONAL | PROMO

Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

- After: Multi-class
- SPAM | HAM | NEWSLETTER | TRANSACTIONAL | PROMO

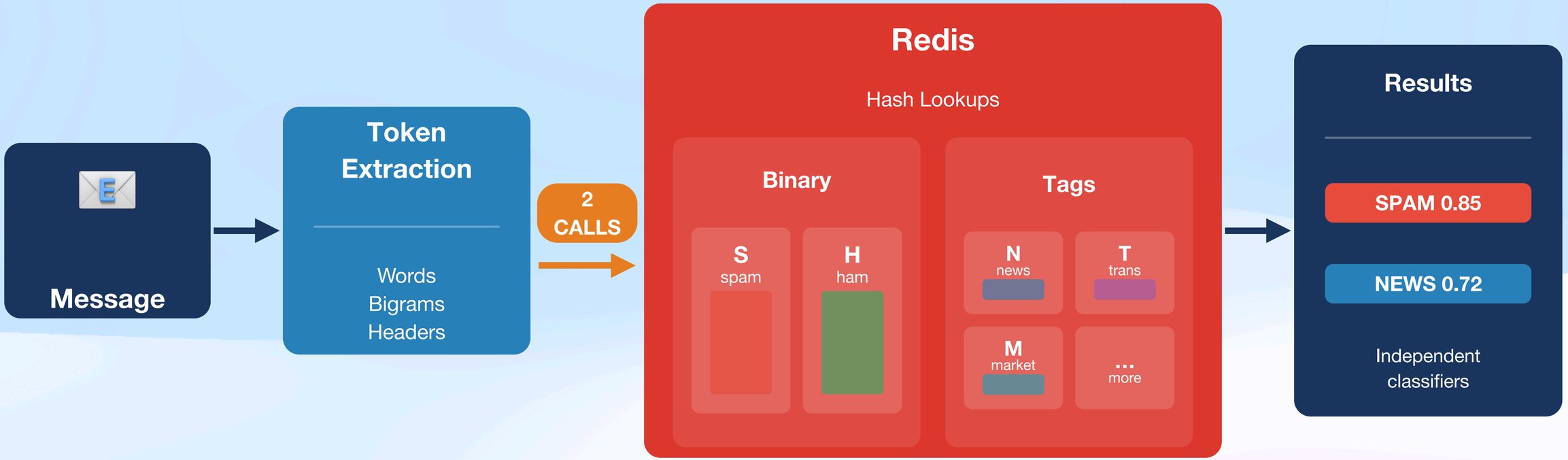
Multi-class Bayes

- Before: Binary World
- SPAM vs HAM only
- Newsletters ≠ spam, but not exactly ham either

- After: Multi-class
- SPAM | HAM | NEWSLETTER | TRANSACTIONAL | PROMO

- Originally a GSoC project — finished with LLM assistance

Multi-class Bayes: Architecture



Architecture:

- Binary classifier: spam vs ham
- Multiclass classifier: content tags
- Independent training & scoring

```
classifiers = {
  "bayes" = { binary }
  "tags" = { multiclass }
}
```

Parallel Scoring

Both classifiers run together

Fuzzy TCP: The Backstory

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

- Solution: Async TCP Protocol

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

- Solution: Async TCP Protocol
- Existing draft PR, never finished

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

- Solution: Async TCP Protocol
- Existing draft PR, never finished
- Completed with LLM assistance in days

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

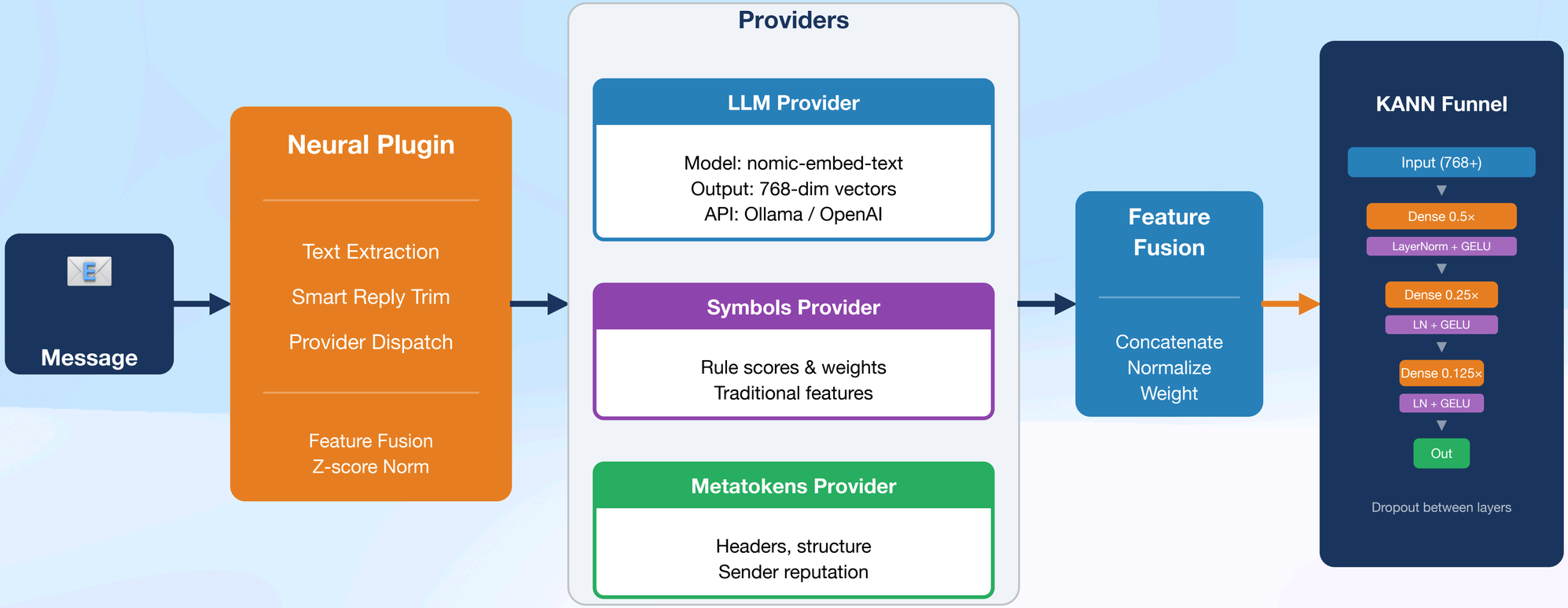
- Solution: Async TCP Protocol
- Existing draft PR, never finished
- Completed with LLM assistance in days
- Connection-based = no port scan appearance

Fuzzy TCP: The Backstory

- October 2025: Hetzner Incident
- 1. Detection: Port scan tool flags UDP traffic on port 11335
- 2. Block: Server suspended — suspected attack
- 3. Reality: Normal fuzzy protocol queries
- 4. Impact: Hundreds of thousands of users affected

- Solution: Async TCP Protocol
- Existing draft PR, never finished
- Completed with LLM assistance in days
- Connection-based = no port scan appearance
- Full backward compatibility

Neural Embeddings: New Architecture



Redis Cache

Embedding storage

Funnel Architecture

✓ Auto-scaling: 3 layers (>512d), 2 layers (256-512d), 1 layer
✓ GELU activation + LayerNorm for embedding inputs
✓ Dropout regularization (0.2 default, 0.1 on final layer)
✓ Redis caching → no repeated embedding API calls

Neural Embeddings: Results

Internal Testing

Metric	Value
True Positives	223
False Positives	18
True Negatives	229
False Negatives	3
Unclassified	93
Accuracy	**95.56%**
Precision	92.53%
Recall	**98.67%**
F1 Score	**95.50%**
Coverage	83.57%

Independent Testing

Metric	Value
True Positives	341
False Positives	20
True Negatives	1025
False Negatives	5
Unclassified	68
Accuracy	**98.20%**
Precision	94.46%
Recall	**98.55%**
F1 Score	**96.46%**
Coverage	95.34%

Neural: Performance and Stability

Neural: Performance and Stability

- GPU vs CPU

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

- Additional Improvements

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

- Additional Improvements
- Smart reply trimming (PR #5845)

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

- Additional Improvements
- Smart reply trimming (PR #5845)
- LLM context enhancement (PR #5732, #5647)

Neural: Performance and Stability

- GPU vs CPU
- GPU (CUDA): 10-100x faster
- CPU: Viable for personal mail servers

- Key Advantage Over Raw LLM
- Stable results — same input → same output
- No API rate limits or costs at inference
- Offline operation possible

- Additional Improvements
- Smart reply trimming (PR #5845)
- LLM context enhancement (PR #5732, #5647)
- Provider architecture for mixing sources

Questions?