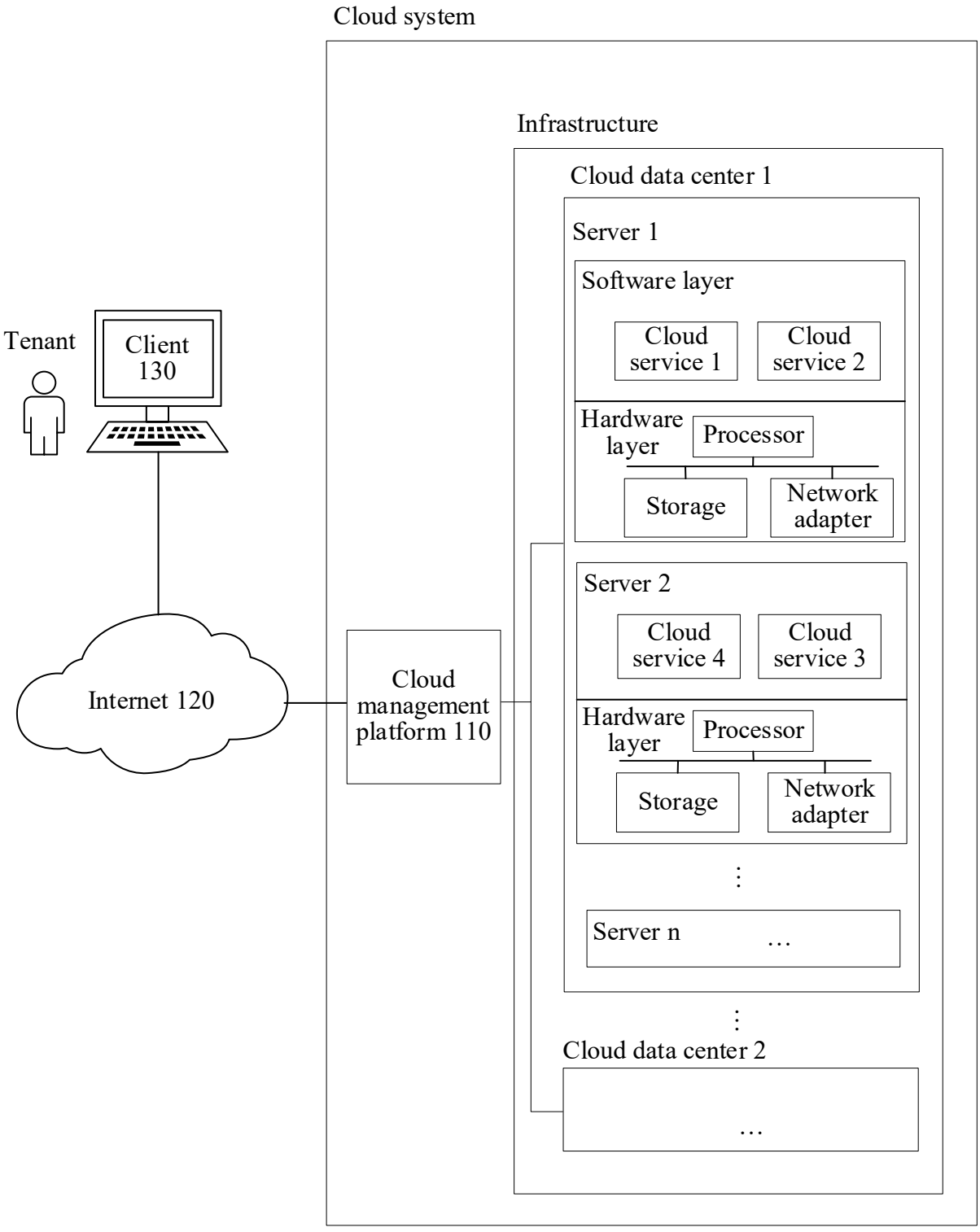
Cloud system
Infrastructure
Cloud data center 1
Server 1
Software layer
Cloud service 1
Cloud service 2
Hardware layer
Processor
Storage
Network adapter
Server 2
Cloud service 4
Cloud service 3
Hardware layer
Processor
Storage
Network adapter
Server n
…
Cloud data center 2
…
Tenant
Client 130
Internet 120
Cloud management platform 110

FIG. 1

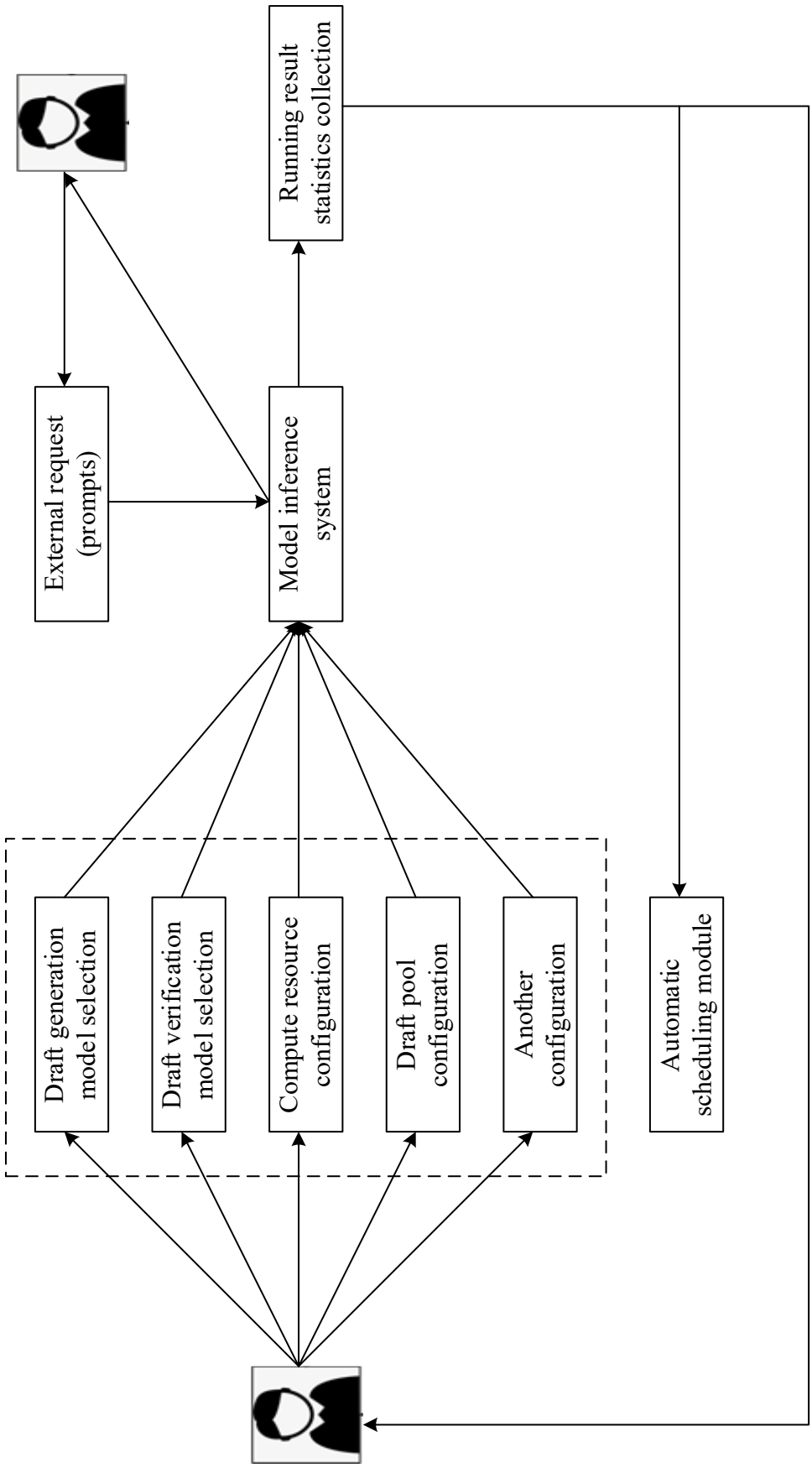
Running result statistics collection
External request (prompts)
Model inference system
Draft generation model selection
Draft verification model selection
Compute resource configuration
Draft pool configuration
Another configuration
Automatic scheduling module

FIG. 2

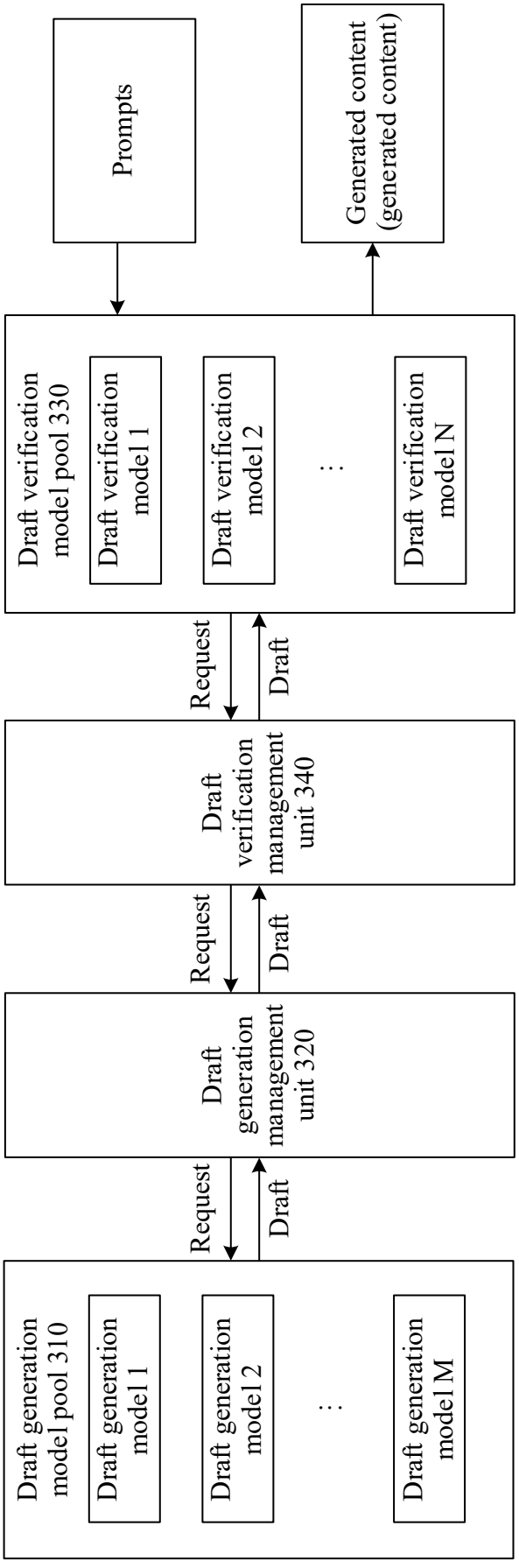
Prompts
Generated content (generated content)
Draft verification model pool 330
Draft verification model 1
Draft verification model 2
Draft verification model N
Request
Draft
Draft verification management unit 340
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model M

FIG. 3

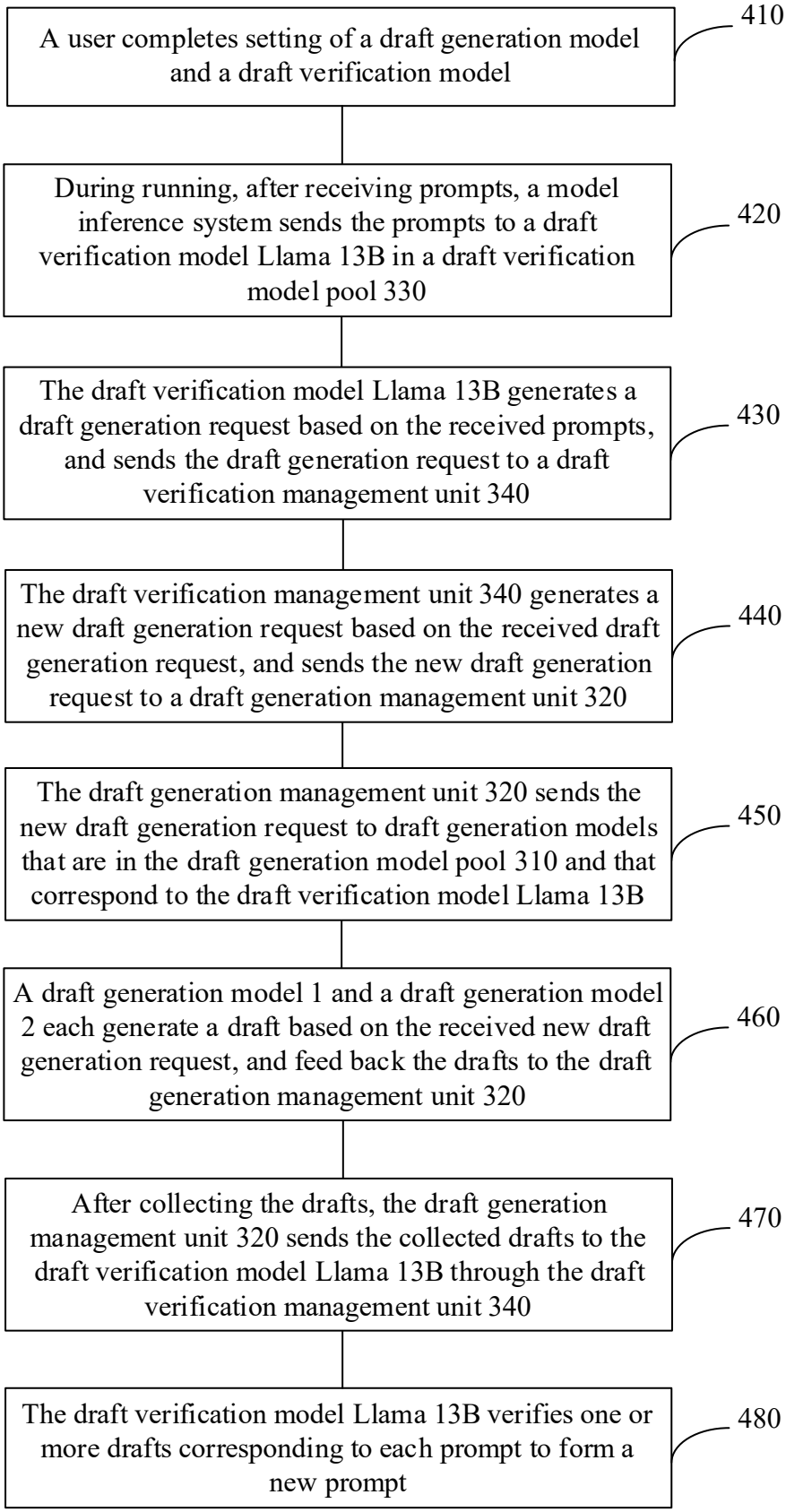
A user completes setting of a draft generation model and a draft verification model
410
During running, after receiving prompts, a model inference system sends the prompts to a draft verification model Llama 13B in a draft verification model pool 330
420
The draft verification model Llama 13B generates a draft generation request based on the received prompts, and sends the draft generation request to a draft verification management unit 340
430
The draft verification management unit 340 generates a new draft generation request based on the received draft generation request, and sends the new draft generation request to a draft generation management unit 320
440
The draft generation management unit 320 sends the new draft generation request to draft generation models that are in the draft generation model pool 310 and that correspond to the draft verification model Llama 13B
450
A draft generation model 1 and a draft generation model 2 each generate a draft based on the received new draft generation request, and feed back the drafts to the draft generation management unit 320
460
After collecting the drafts, the draft generation management unit 320 sends the collected drafts to the draft verification model Llama 13B through the draft verification management unit 340
470
The draft verification model Llama 13B verifies one or more drafts corresponding to each prompt to form a new prompt
480

FIG. 4

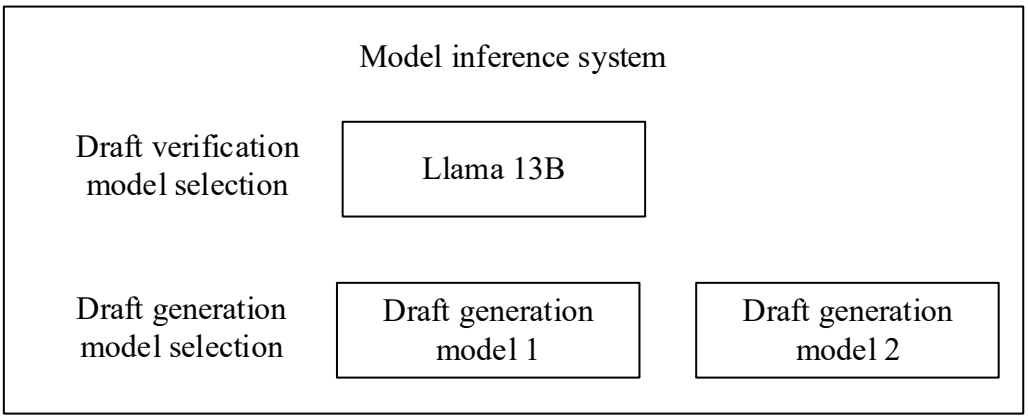
Model inference system
Draft verification model selection
Llama 13B
Draft generation model selection
Draft generation model 1
Draft generation model 2

FIG. 5

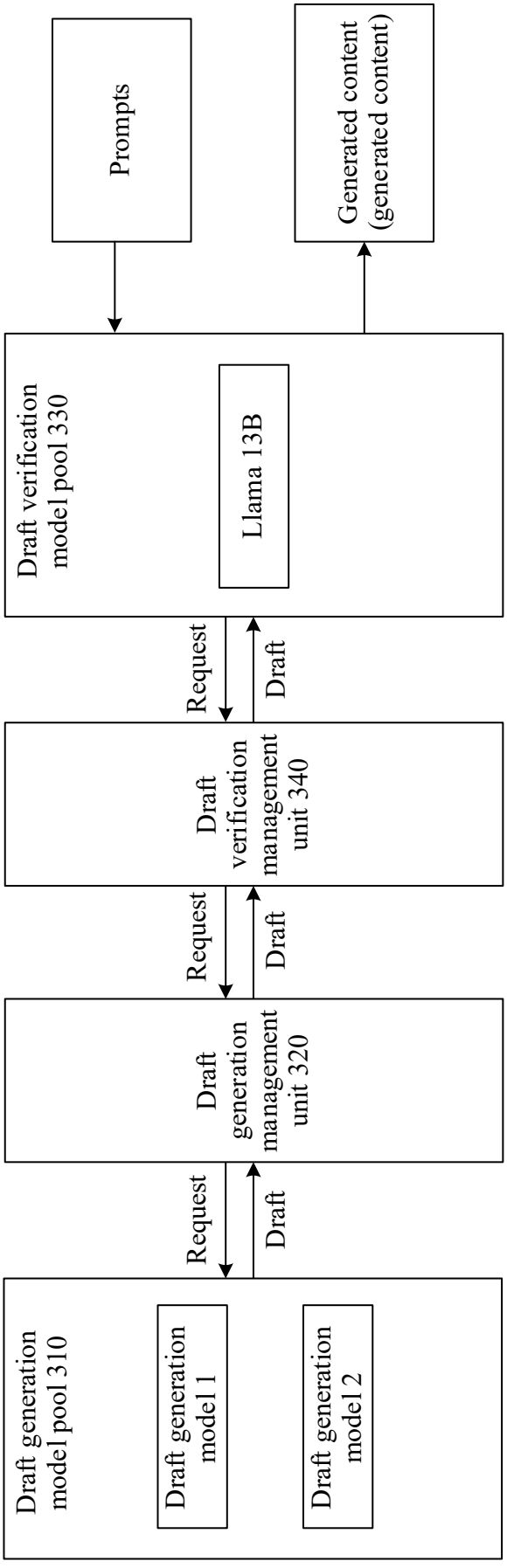
Prompts
Generated content (generated content)
Draft verification model pool 330
Llama 13B
Request
Draft
Draft verification management unit 340
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2

FIG. 6

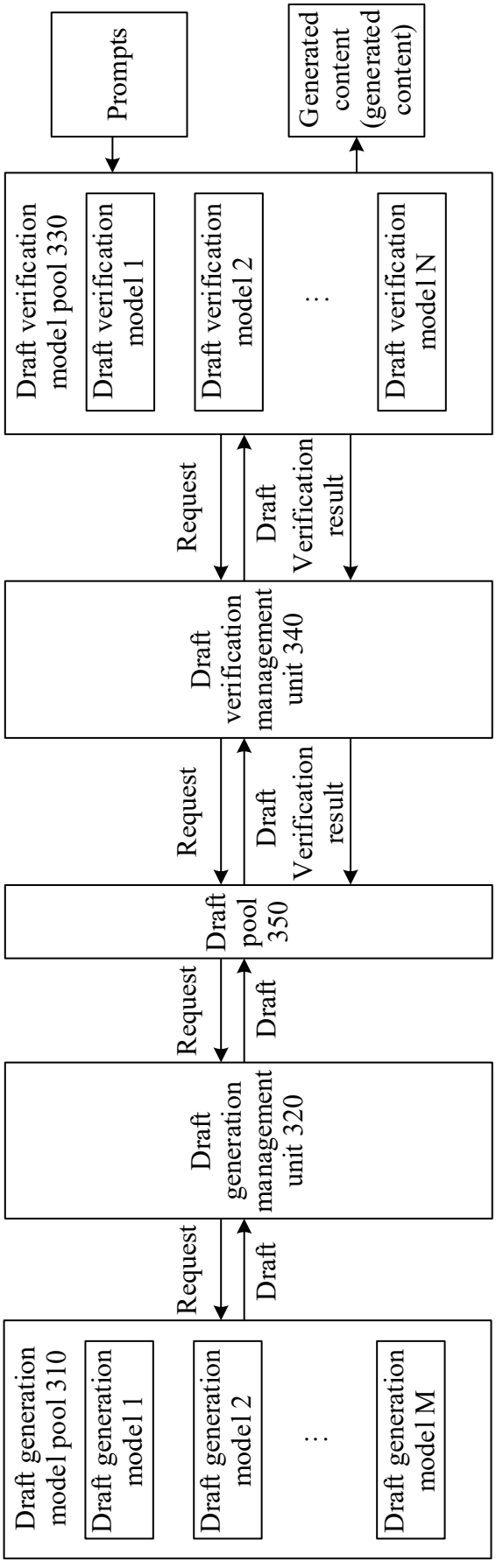
Prompts
Generated content (generated content)
Draft verification model pool 330
Draft verification model 1
Draft verification model 2
Draft verification model N
Request
Draft Verification result
Draft verification management unit 340
Request
Draft Verification result
Draft pool 350
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model M

FIG. 7

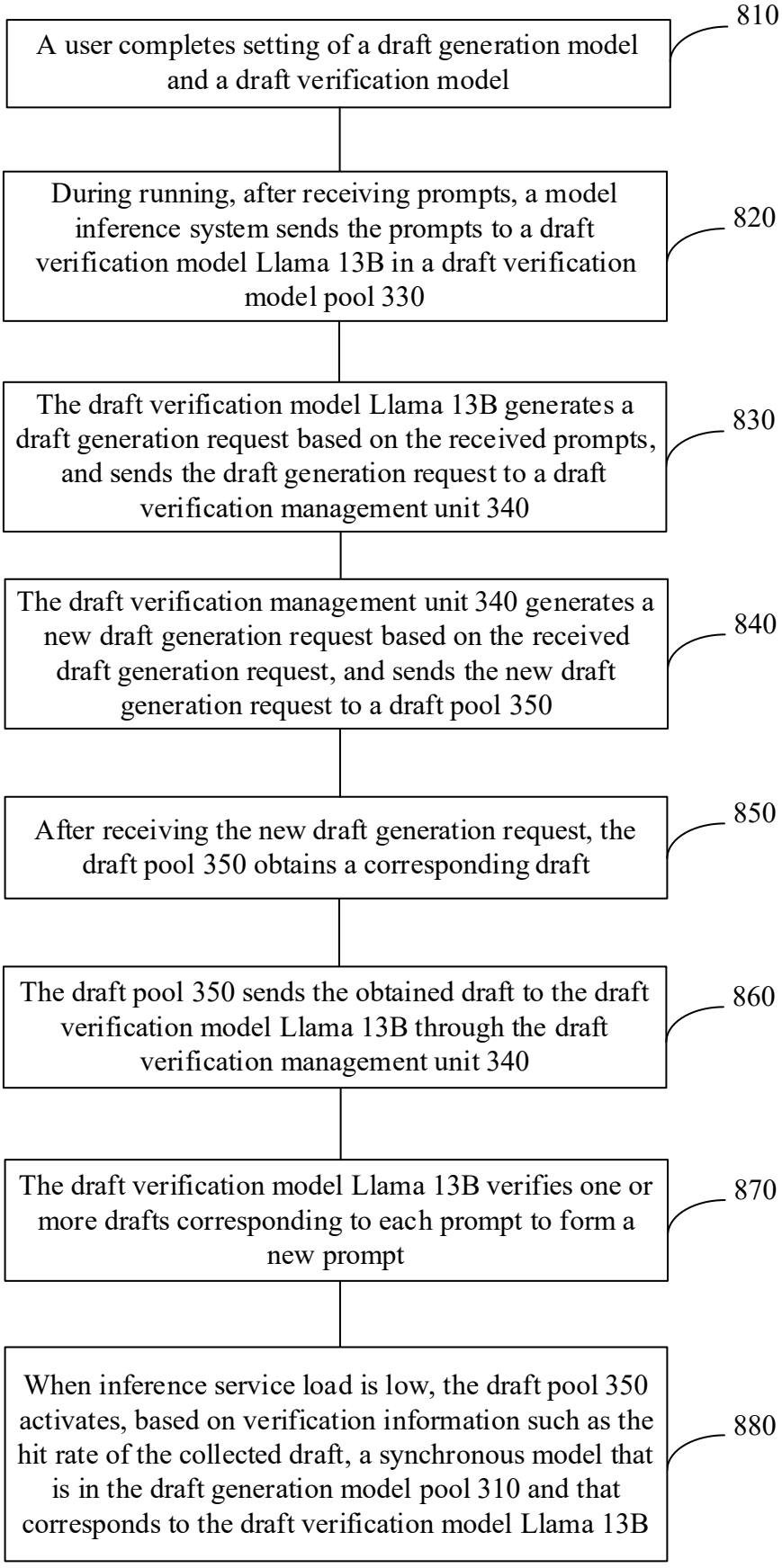
A user completes setting of a draft generation model and a draft verification model
810
During running, after receiving prompts, a model inference system sends the prompts to a draft verification model Llama 13B in a draft verification model pool 330
820
The draft verification model Llama 13B generates a draft generation request based on the received prompts, and sends the draft generation request to a draft verification management unit 340
830
The draft verification management unit 340 generates a new draft generation request based on the received draft generation request, and sends the new draft generation request to a draft pool 350
840
After receiving the new draft generation request, the draft pool 350 obtains a corresponding draft
850
The draft pool 350 sends the obtained draft to the draft verification model Llama 13B through the draft verification management unit 340
860
The draft verification model Llama 13B verifies one or more drafts corresponding to each prompt to form a new prompt
870
When inference service load is low, the draft pool 350 activates, based on verification information such as the hit rate of the collected draft, a synchronous model that is in the draft generation model pool 310 and that corresponds to the draft verification model Llama 13B
880

FIG. 8

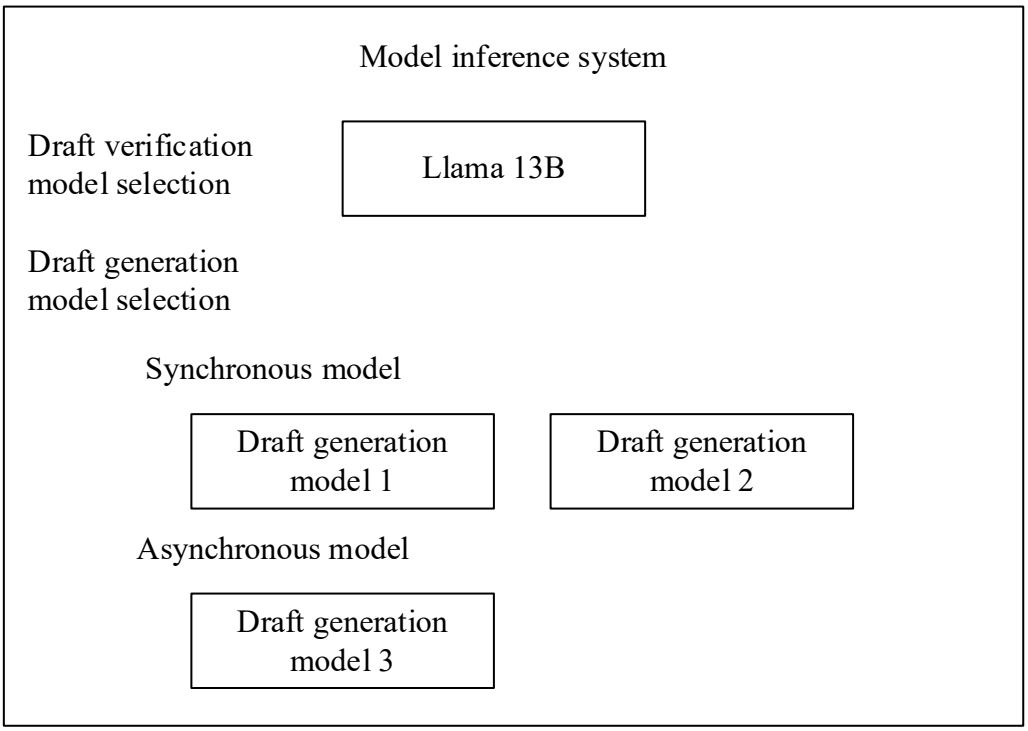
Model inference system
Draft verification model selection
Llama 13B
Draft generation model selection
Synchronous model
Draft generation model 1
Draft generation model 2
Asynchronous model
Draft generation model 3

FIG. 9

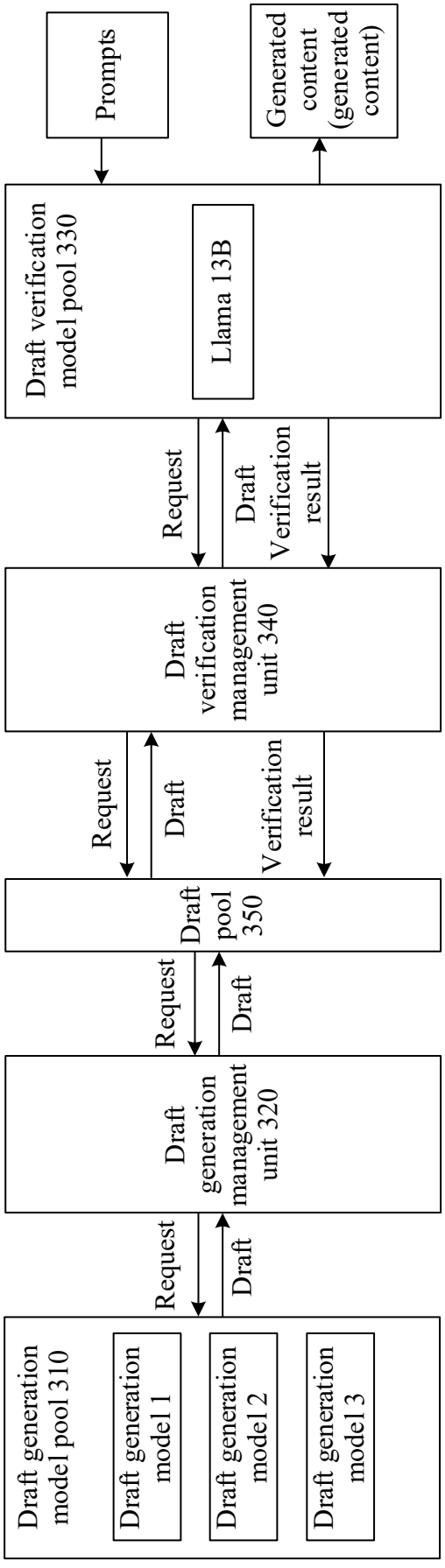
Prompts
Generated content (generated content)
Draft verification model pool 330
Llama 13B
Request
Draft Verification result
Draft verification management unit 340
Request
Draft
Verification result
Draft pool 350
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model 3

FIG. 10

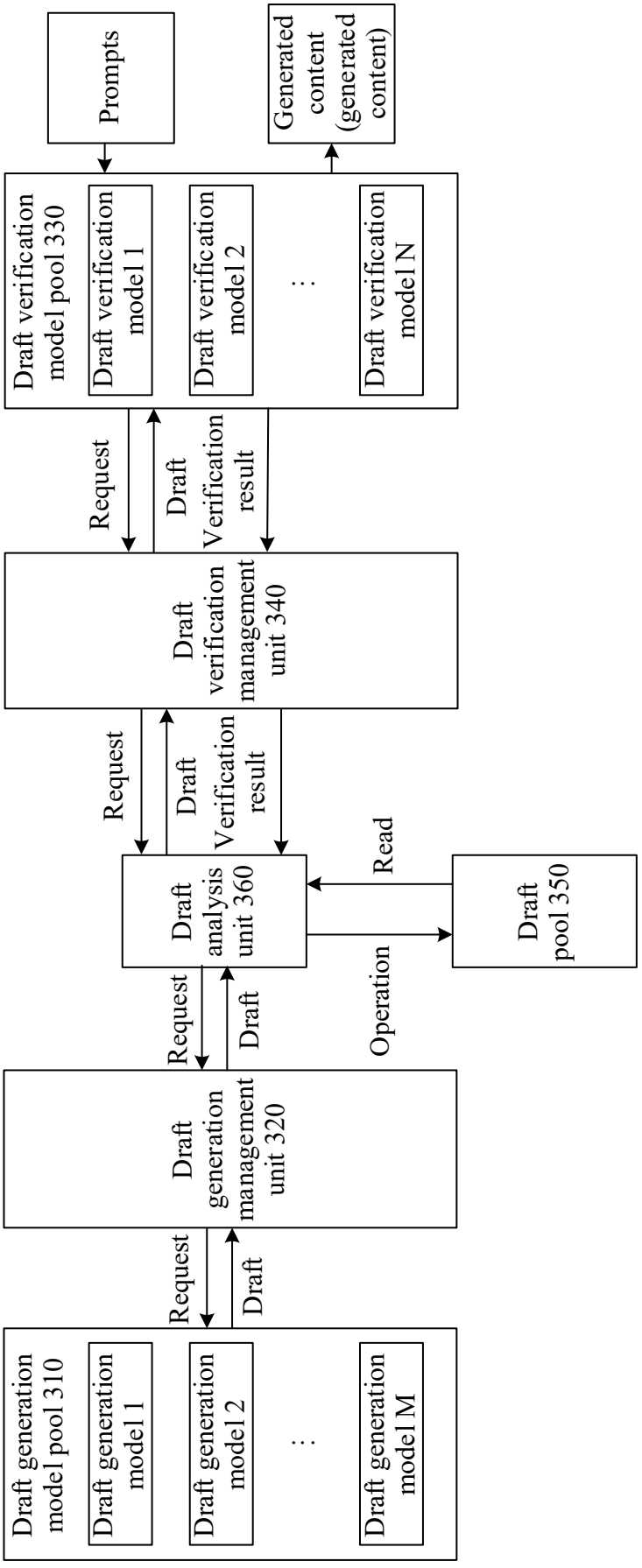
Prompts
Generated content (generated content)
Draft verification model pool 330
Draft verification model 1
Draft verification model 2
Draft verification model N
Request
Draft Verification result
Draft verification management unit 340
Request
Draft Verification result
Draft analysis unit 360
Read
Draft pool 350
Request
Draft
Operation
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model M

FIG. 11

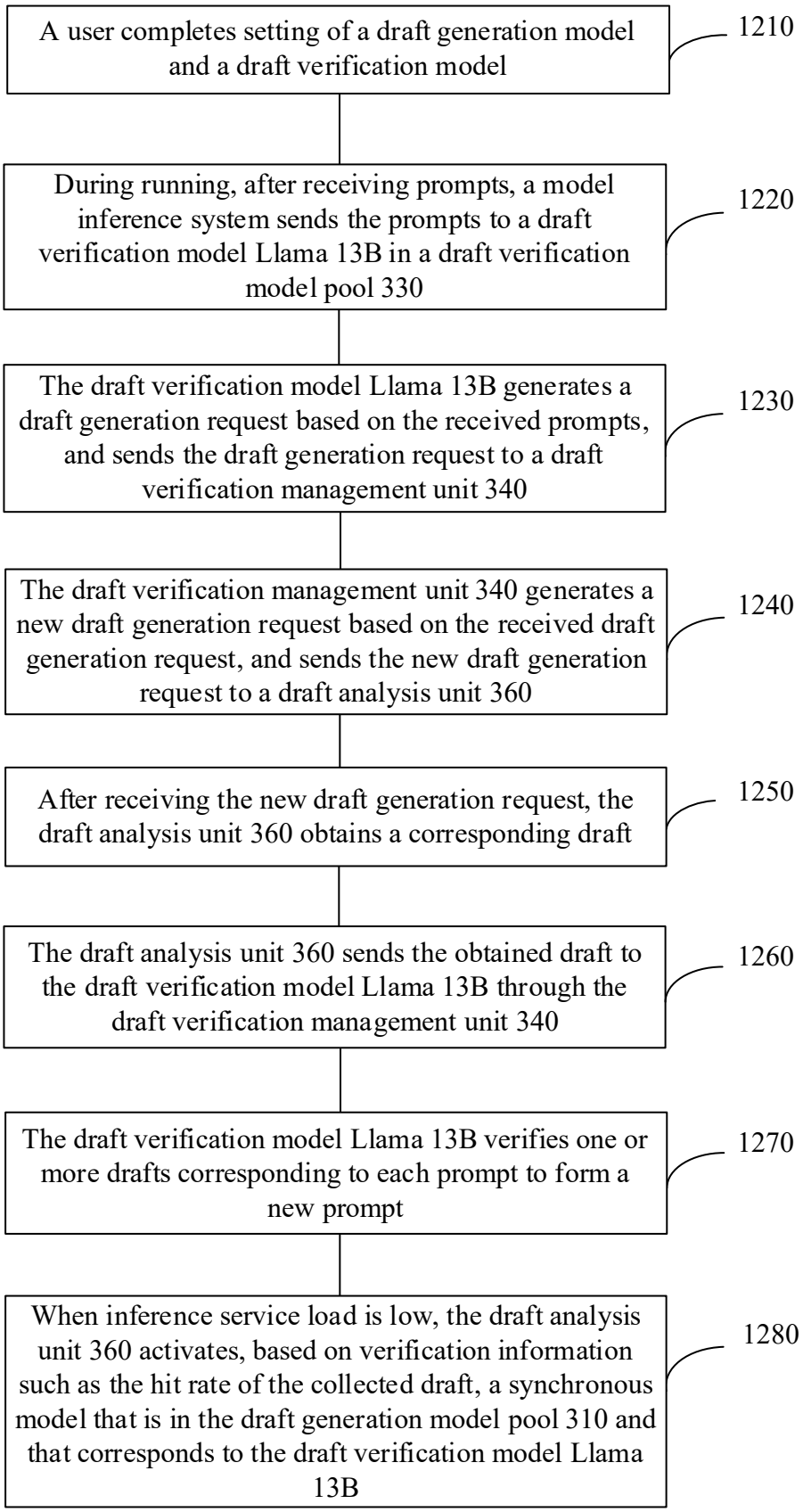
A user completes setting of a draft generation model and a draft verification model
1210
During running, after receiving prompts, a model inference system sends the prompts to a draft verification model Llama 13B in a draft verification model pool 330
1220
The draft verification model Llama 13B generates a draft generation request based on the received prompts, and sends the draft generation request to a draft verification management unit 340
1230
The draft verification management unit 340 generates a new draft generation request based on the received draft generation request, and sends the new draft generation request to a draft analysis unit 360
1240
After receiving the new draft generation request, the draft analysis unit 360 obtains a corresponding draft
1250
The draft analysis unit 360 sends the obtained draft to the draft verification model Llama 13B through the draft verification management unit 340
1260
The draft verification model Llama 13B verifies one or more drafts corresponding to each prompt to form a new prompt
1270
When inference service load is low, the draft analysis unit 360 activates, based on verification information such as the hit rate of the collected draft, a synchronous model that is in the draft generation model pool 310 and that corresponds to the draft verification model Llama 13B
1280

FIG. 12

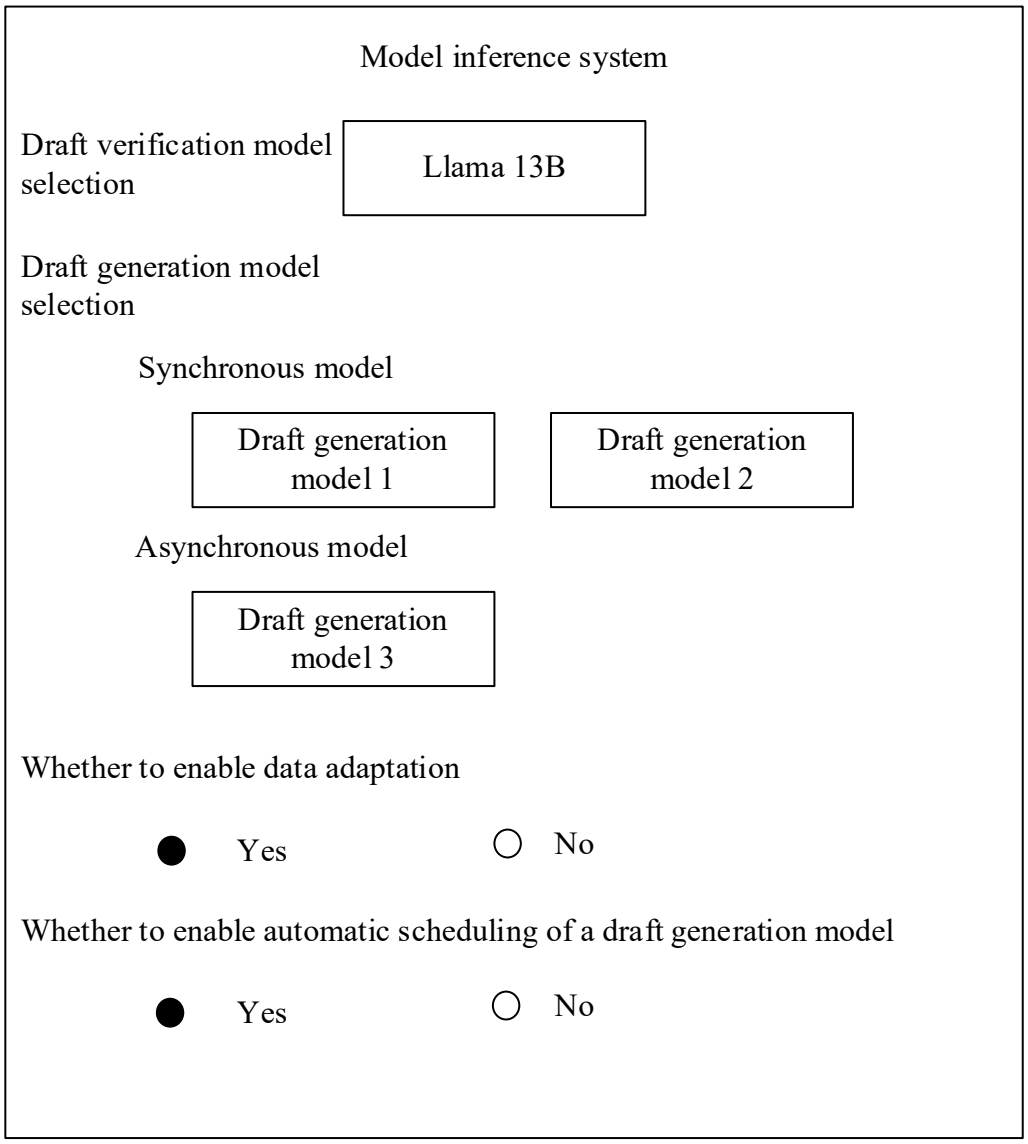
Model inference system
Draft verification model selection
Llama 13B
Draft generation model selection
Synchronous model
Draft generation model 1
Draft generation model 2
Asynchronous model
Draft generation model 3
Whether to enable data adaptation
Yes
No
Whether to enable automatic scheduling of a draft generation model
Yes
No

FIG. 13

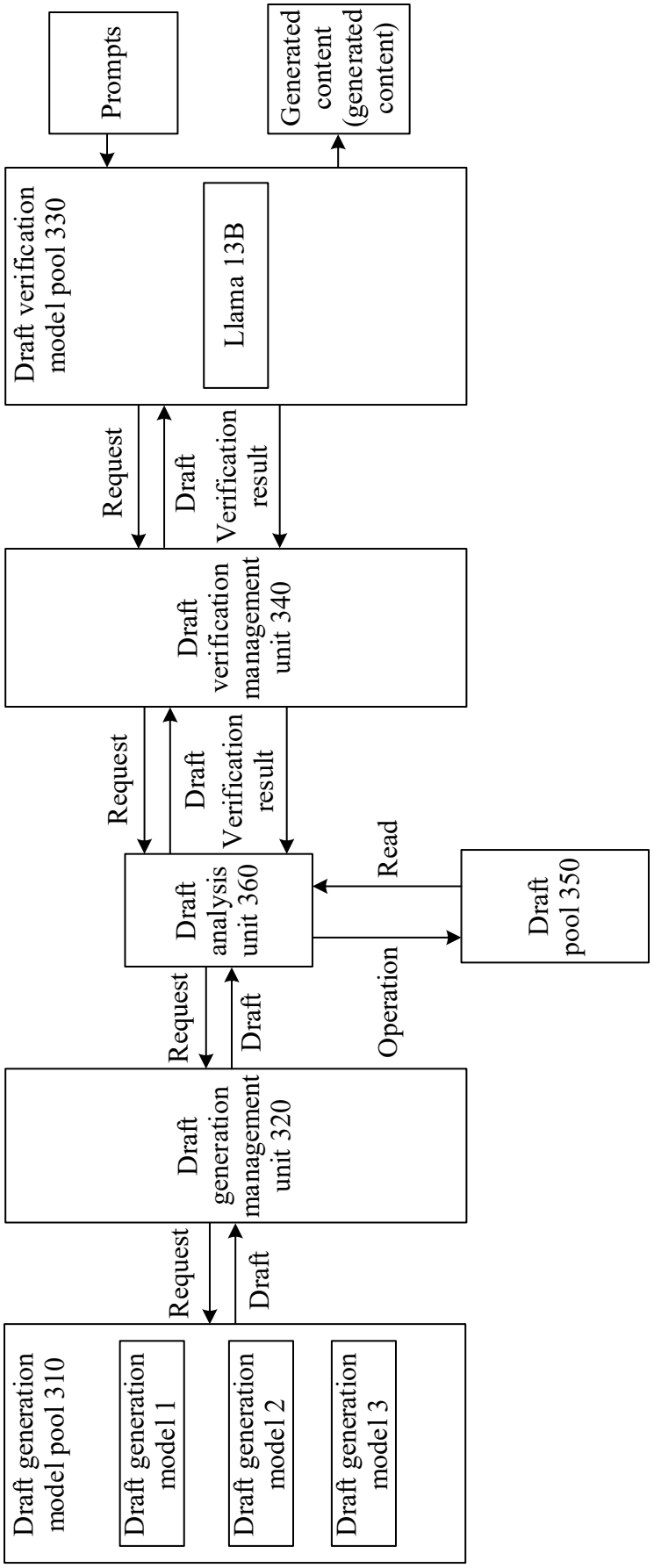
Prompts
Generated content (generated content)
Draft verification model pool 330
Llama 13B
Request
Draft Verification result
Draft verification management unit 340
Request
Draft Verification result
Read
Draft analysis unit 360
Draft pool 350
Request
Draft
Operation
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model 3

FIG. 14

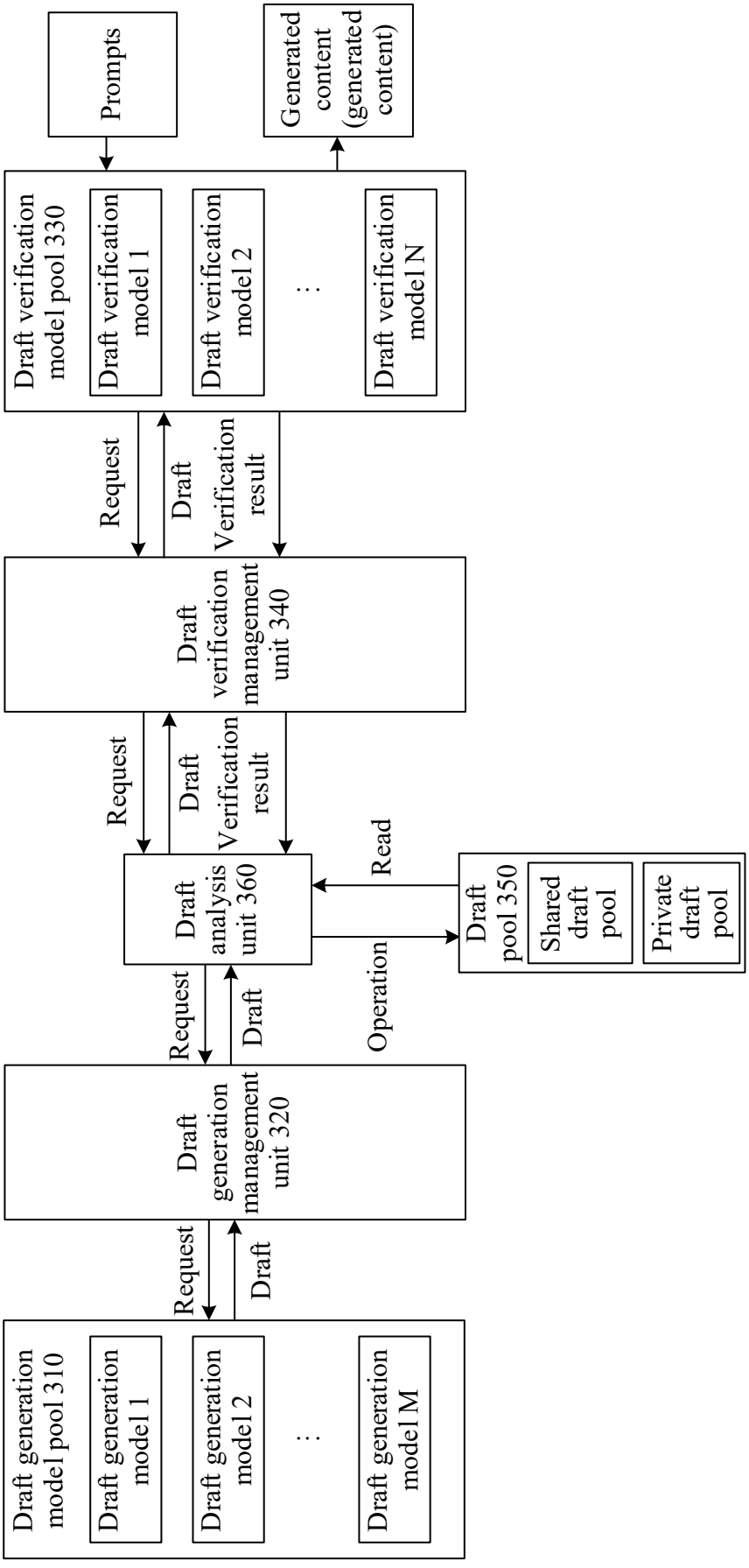
Prompts
Generated content (generated content)
Draft verification model pool 330
Draft verification model 1
Draft verification model 2
Draft verification model N
Request
Draft
Verification result
Draft verification management unit 340
Request
Draft
Verification result
Draft analysis unit 360
Read
Draft pool 350
Shared draft pool
Private draft pool
Request
Draft
Operation
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model M

FIG. 15

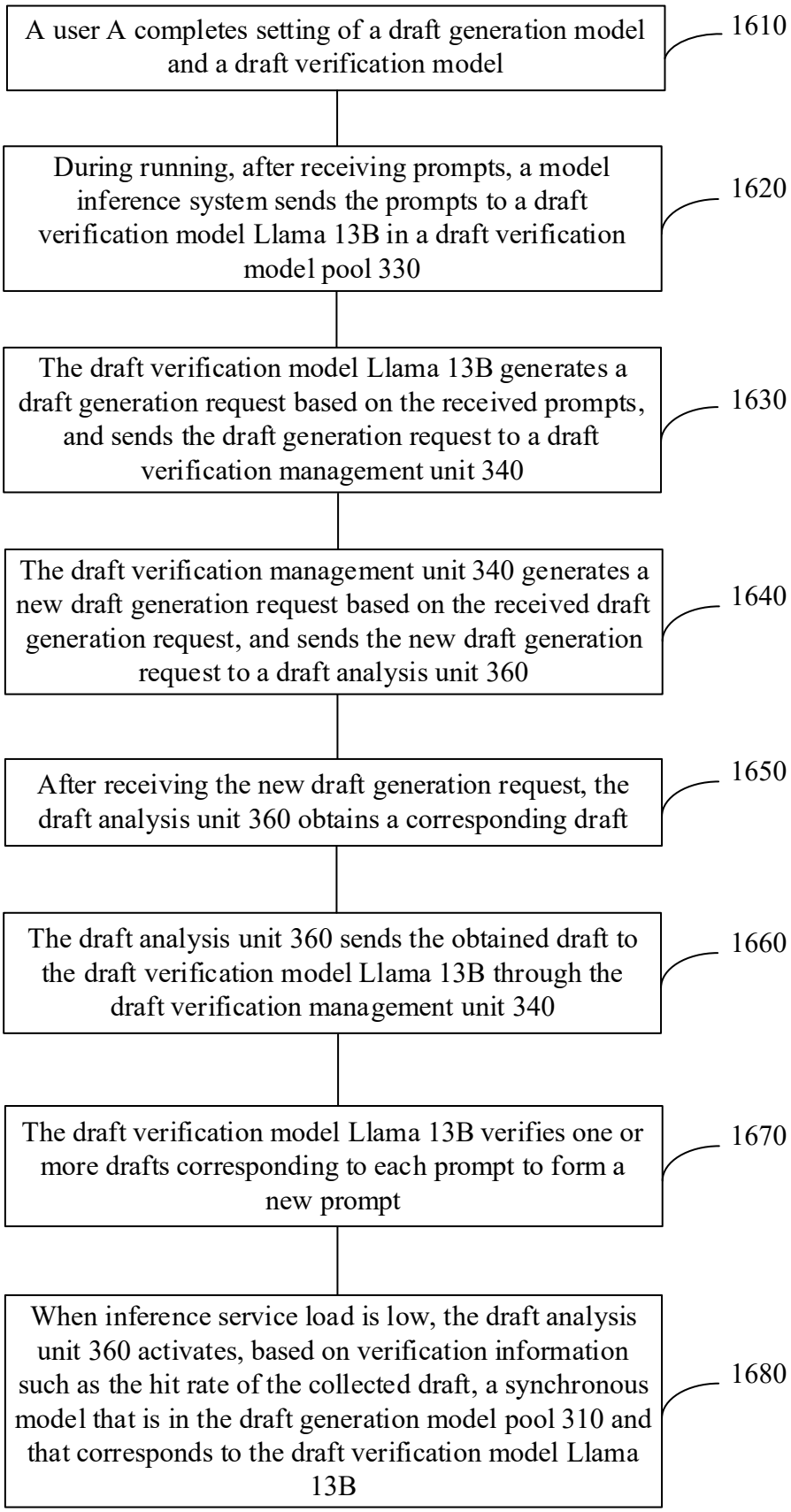
A user A completes setting of a draft generation model and a draft verification model
1610
During running, after receiving prompts, a model inference system sends the prompts to a draft verification model Llama 13B in a draft verification model pool 330
1620
The draft verification model Llama 13B generates a draft generation request based on the received prompts, and sends the draft generation request to a draft verification management unit 340
1630
The draft verification management unit 340 generates a new draft generation request based on the received draft generation request, and sends the new draft generation request to a draft analysis unit 360
1640
After receiving the new draft generation request, the draft analysis unit 360 obtains a corresponding draft
1650
The draft analysis unit 360 sends the obtained draft to the draft verification model Llama 13B through the draft verification management unit 340
1660
The draft verification model Llama 13B verifies one or more drafts corresponding to each prompt to form a new prompt
1670
When inference service load is low, the draft analysis unit 360 activates, based on verification information such as the hit rate of the collected draft, a synchronous model that is in the draft generation model pool 310 and that corresponds to the draft verification model Llama 13B
1680

FIG. 16

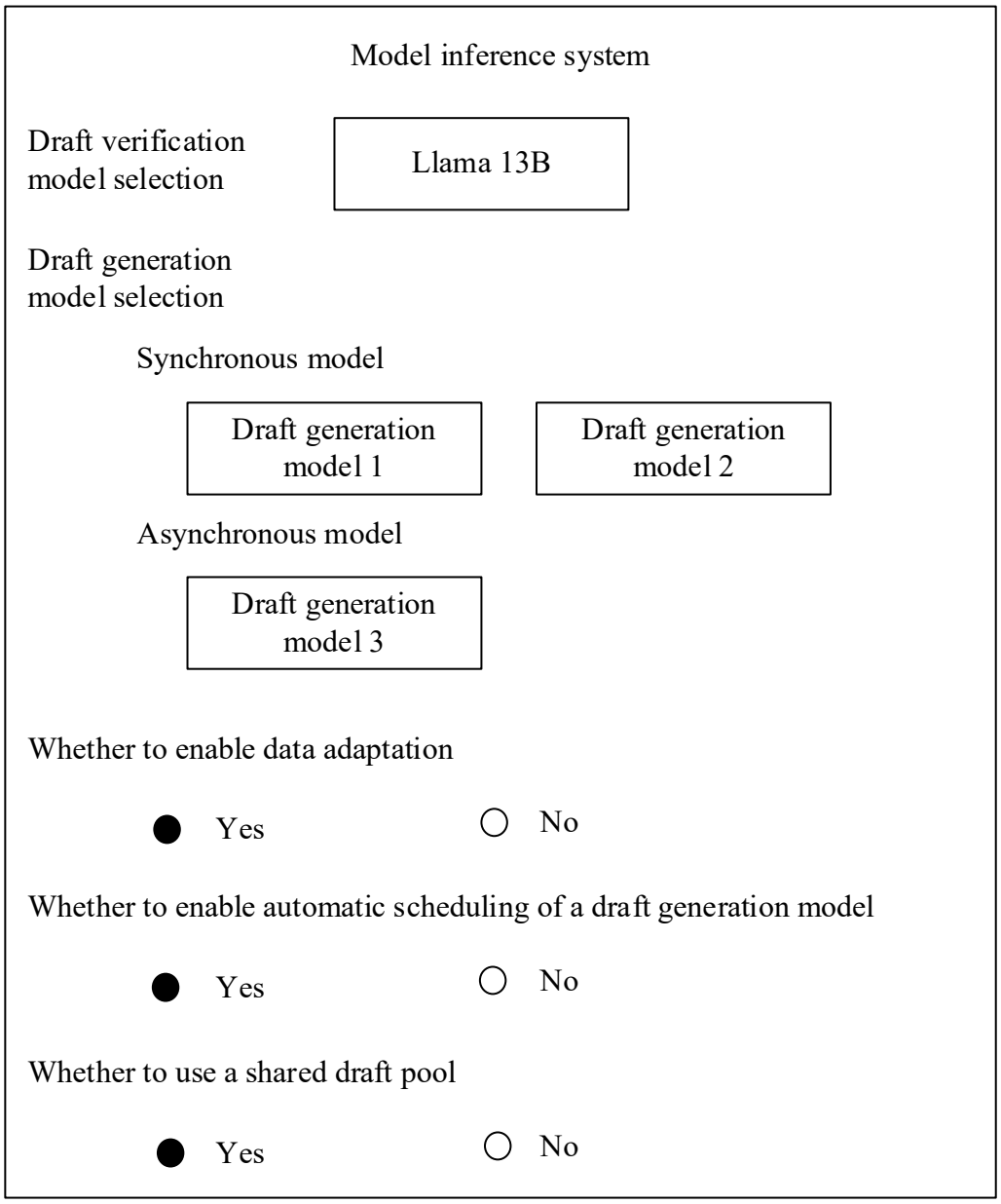
Model inference system
Draft verification model selection
Llama 13B
Draft generation model selection
Synchronous model
Draft generation model 1
Draft generation model 2
Asynchronous model
Draft generation model 3
Whether to enable data adaptation
Yes
No
Whether to enable automatic scheduling of a draft generation model
Yes
No
Whether to use a shared draft pool
Yes
No

FIG. 17

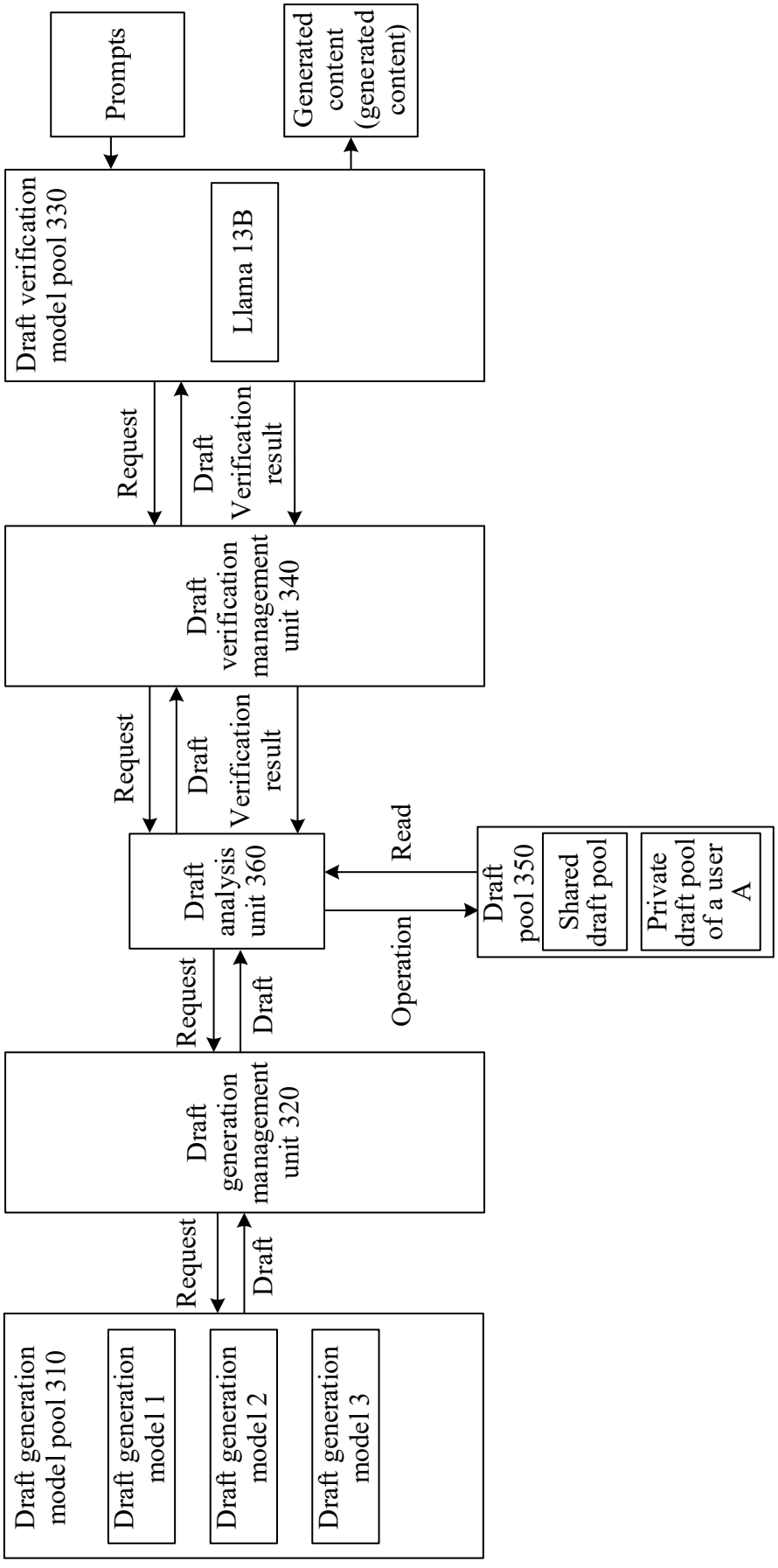
Prompts
Generated content (generated content)
Draft verification model pool 330
Llama 13B
Request
Draft Verification result
Draft verification management unit 340
Request
Draft Verification result
Draft analysis unit 360
Read
Operation
Draft pool 350
Shared draft pool
Private draft pool of a user A
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 2
Draft generation model 3

FIG. 18

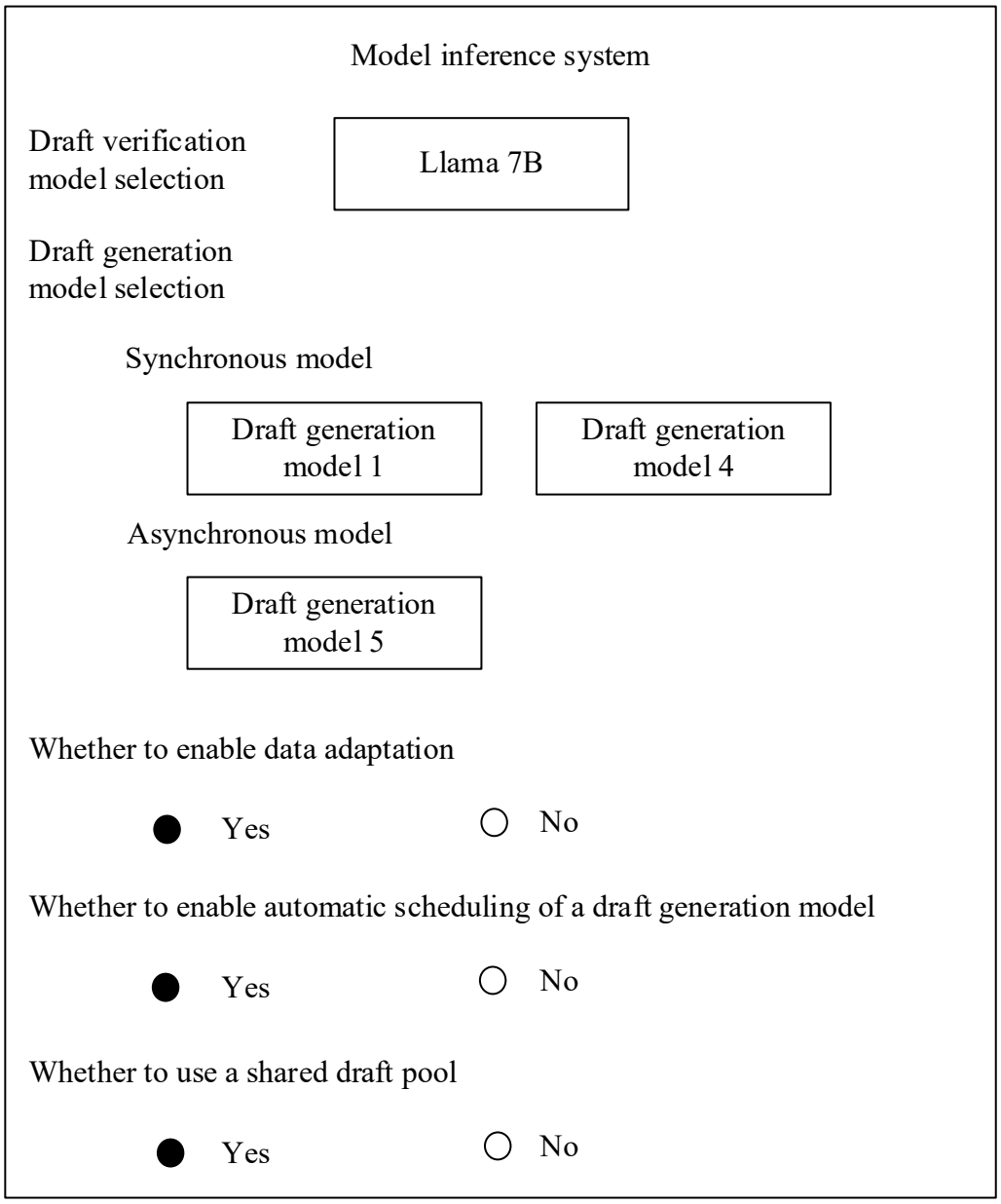
Model inference system
Draft verification model selection
Llama 7B
Draft generation model selection
Synchronous model
Draft generation model 1
Draft generation model 4
Asynchronous model
Draft generation model 5
Whether to enable data adaptation
Yes
No
Whether to enable automatic scheduling of a draft generation model
Yes
No
Whether to use a shared draft pool
Yes
No

FIG. 19

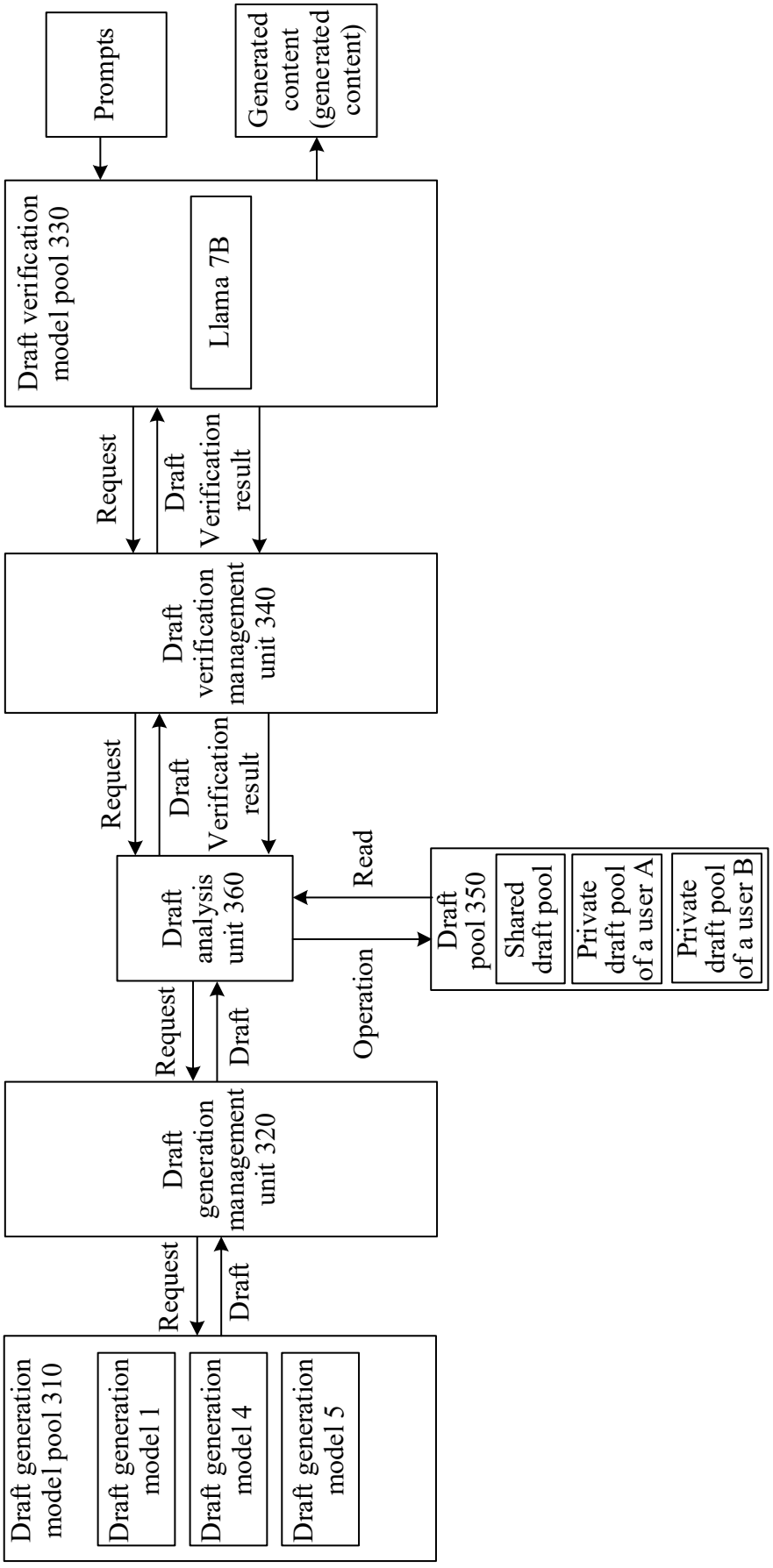
Prompts
Generated content (generated content)
Draft verification model pool 330
Llama 7B
Request
Draft Verification result
Draft verification management unit 340
Request
Draft Verification result
Draft analysis unit 360
Read
Operation
Draft pool 350
Shared draft pool
Private draft pool of a user A
Private draft pool of a user B
Request
Draft
Draft generation management unit 320
Request
Draft
Draft generation model pool 310
Draft generation model 1
Draft generation model 4
Draft generation model 5

FIG. 20

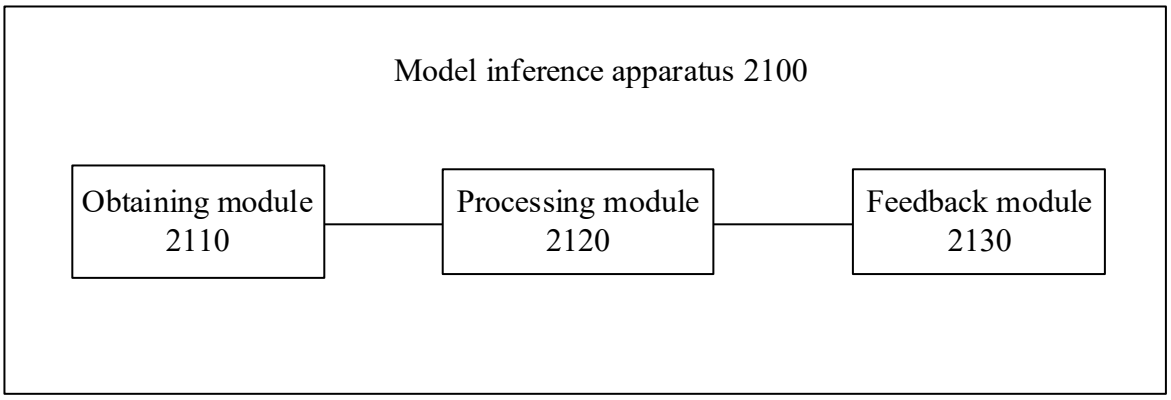
Model inference apparatus 2100
Obtaining module 2110
Processing module 2120
Feedback module 2130

FIG. 21

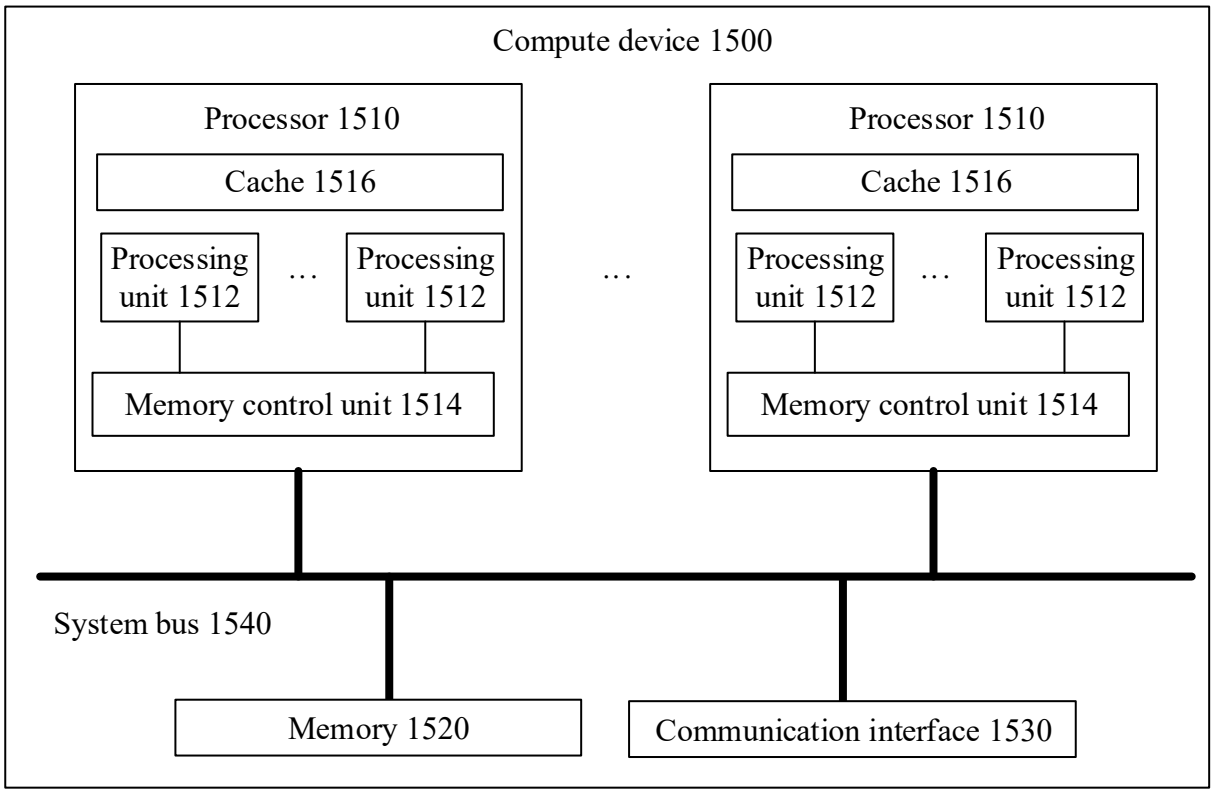
Compute device 1500
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
System bus 1540
Memory 1520
Communication interface 1530

FIG. 22

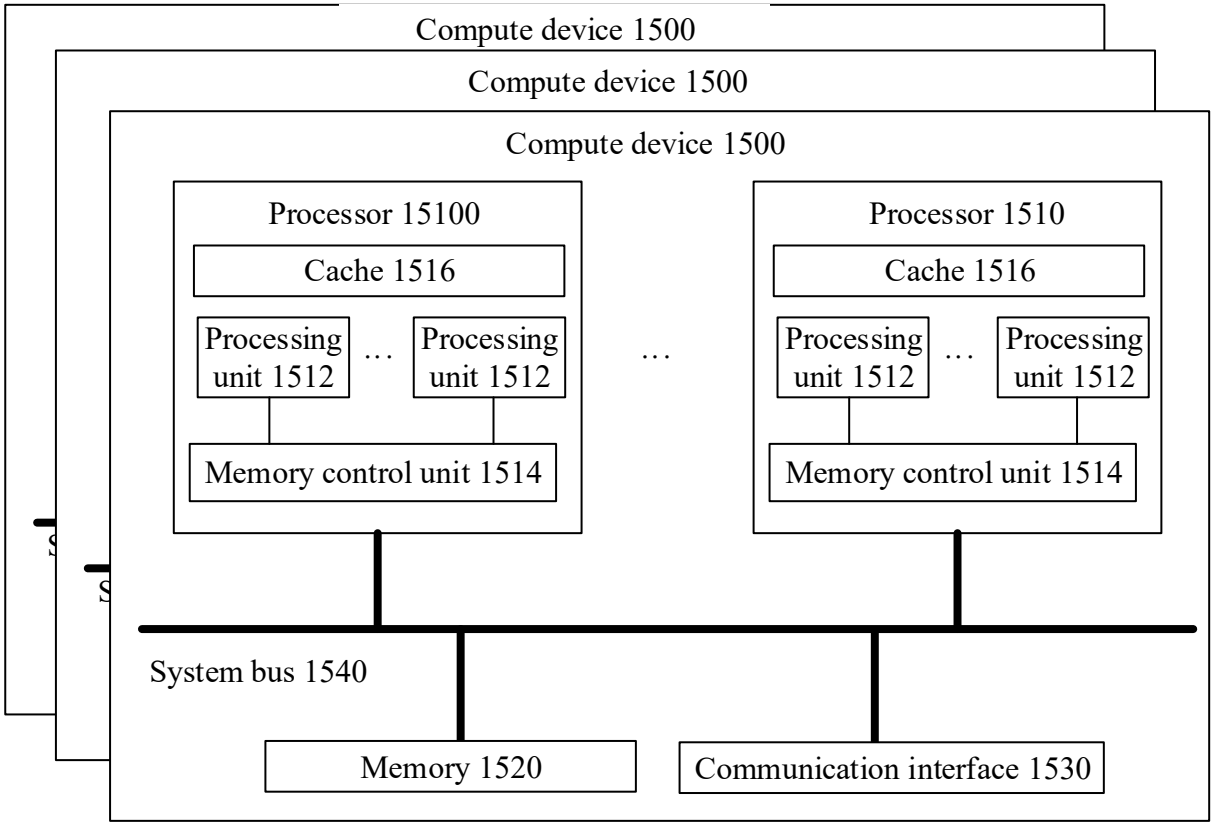
Compute device 1500
Compute device 1500
Compute device 1500
Processor 15100
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
System bus 1540
Memory 1520
Communication interface 1530

FIG. 23

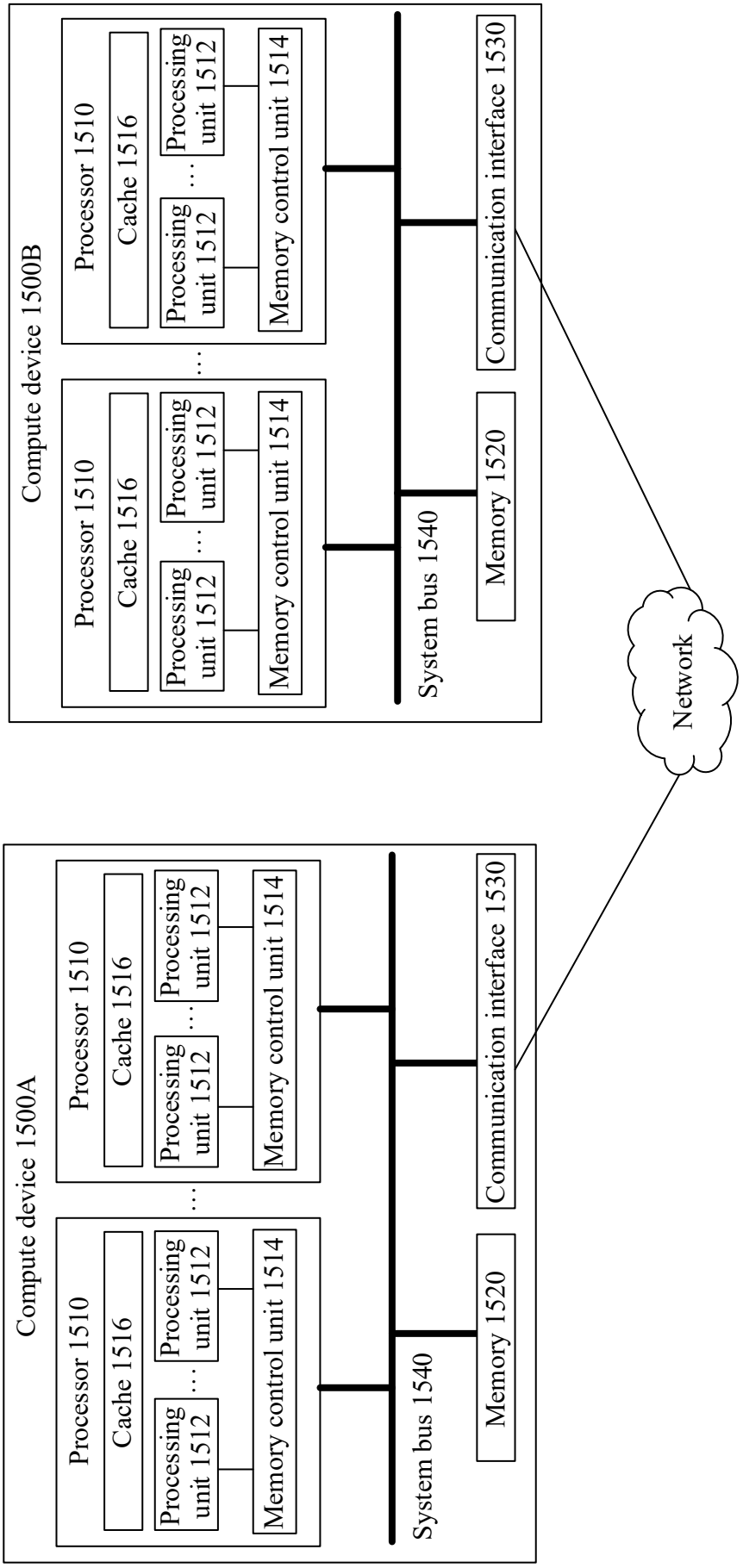
Compute device 1500A
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
System bus 1540
Memory 1520
Communication interface 1530
Network
Compute device 1500B
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
Processor 1510
Cache 1516
Processing unit 1512
Processing unit 1512
Memory control unit 1514
System bus 1540
Memory 1520
Communication interface 1530

FIG. 24