PR #33140 完整报告

sgl-project/sglang

[DSV4] Add official DSV4 reasoning effort support

合并时间：2026-08-05 12:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33140

执行摘要

- 一句话：DSV4 官方 reasoning effort 三档 profile 支持
- 推荐动作：值得精读。核心看点在 chat_encoding.py 的‘不执行代码的 checkpoint 元数据探测’设计（AST + literal_eval、大小上限、三级回退链）以及 encoding_dsv4.py 的 profile 映射表——对需要按模型版本切换 prompt 契约的 serving 框架有直接借鉴价值。建议同时阅读配套单测理解各回退分支的覆盖方式。

功能与动机

PR body 明确指出：DeepSeek-V4-Flash-0731 使用官方 reasoning-effort profile，low 不加前缀、high 沿用旧 'Absolute maximum' 前缀、max 使用新 'Beyond maximum' 前缀；而旧版 DeepSeek-V4-Flash 仍走 preview 契约。为避免硬编码 checkpoint 日期，SGLang 需要检查 checkpoint 自带的 encoding/encoding_dsv4.py 来在启动时一次性判定 prompt 契约；转换后的 checkpoint 则应能通过 dsv4_reasoning_effort_profile 显式指定。

实现拆解

1. 双契约文案与映射表（encoding_dsv4.py）：将单一 REASONING_EFFORT_MAX 拆为 REASONING_EFFORT_PREVIEW_MAX 与 REASONING_EFFORT_OFFICIAL_MAX，新增嵌套映射 REASONING_EFFORT_PROFILES（preview 支持 high/max，official 支持 low/high/max）。render_message 与 encode_messages 均新增 reasoning_effort_profile 参数（默认 preview）；effort 缺省时 official 默认 low、preview 默认 high；非法组合从 assert 改为显式抛 ValueError，便于服务端给出可读错误。
2. checkpoint 自描述检测（chat_encoding.py）：新增 _detect_dsv4_reasoning_effort_profile，优先读本地 encoding/encoding_dsv4.py，缺失时用 hf_hub_download 拉取；设 1 MiB 大小上限，用 ast.parse + ast.literal_eval 只读顶层赋值（不执行 checkpoint 代码）；判定优先级为‘official 契约（REASONING_EFFORT_PROMPTS 三档齐全且 DEFAULT_REASONING_EFFORT == low）’>‘preview 契约（存在 REASONING_EFFORT_MAX）’> None。resolve_dsv4_reasoning_effort_profile 按 override > 检测结果 > preview 回退的顺序解析，_validate_dsv4_reasoning_effort_profile 在构造期拦截非法 override。
3. 启动接线（serving_chat.py）：OpenAIServingChat 构造时仅当 chat_encoding_spec == dsv4 才解析一次 _dsv4_reasoning_effort_profile，override 取自 hf_config.to_dict() 的 dsv4_reasoning_effort_profile 键；_apply_jinja_template 按 profile 的

accepted_efforts 过滤 request.reasoning_effort 与 SGLANG_DSV4_REASONING_EFFORT，再透传给 encode_messages，避免编码器内部再做二次判断。

4. 测试与文档配套：test_serving_chat.py 新增 _create_dsv4_checkpoint 工具（临时目录写 encoding/encoding_dsv4.py）与 5 个测试方法，覆盖 profile 检测、resolution 回退、model config override、非法 override 构造失败以及端到端 prompt 前缀断言；environ.py 与 environment_variables.mdx 更新 SGLANG_DSV4_REASONING_EFFORT 语义并给出 --json-model-override-args 示例。

关键文件：

- python/sglang/srt/entrypoints/openai/chat_encoding.py（模块 编码分发；类别 source；类型 core-logic；符号 _detect_dsv4_reasoning_effort_profile, _validate_dsv4_reasoning_effort_profile, resolve_dsv4_reasoning_effort_profile）：核心检测逻辑：新增 _detect_dsv4_reasoning_effort_profile（AST 解析 checkpoint 自带 encoder）、_validate_dsv4_reasoning_effort_profile、resolve_dsv4_reasoning_effort_profile，是判定 preview/official 契约的单一入口。
- test/registered/unit/entrypoints/openai/test_serving_chat.py（模块 聊天服务；类别 test；类型 test-coverage；符号 _create_dsv4_checkpoint, test_dsv4_reasoning_effort_profiles, test_dsv4_reasoning_effort_profile_resolution, test_dsv4_reasoning_effort_profile_from_checkpoint）：新增 5 个测试方法覆盖 profile 检测、解析回退、override、非法值构造失败与端到端前缀断言，是验证本 PR 行为的关键配套。
- python/sglang/srt/entrypoints/openai/encoding_dsv4.py（模块 DSV4 编码器；类别 source；类型 core-logic；符号 render_message, encode_messages）：编码核心：REASONING_EFFORT_PROFILES 定义 preview/official 两套 effort 映射，render_message 与 encode_messages 按 profile 选择前缀并改为显式 ValueError。
- python/sglang/srt/entrypoints/openai/serving_chat.py（模块 聊天服务；类别 source；类型 dependency-wiring；符号 _resolve_chat_encoding_spec, _apply_jinja_template）：服务接线：构造时解析一次 _dsv4_reasoning_effort_profile，并在 _apply_jinja_template 中按 profile 的 accepted_efforts 过滤 effort 后传入编码器。
- python/sglang/srt/environ.py（模块 环境变量；类别 source；类型 configuration）：环境变量语义更新：SGLANG_DSV4_REASONING_EFFORT 不再硬编码 high/max，改为由 checkpoint 解析的 profile 决定。
- docs/docs/references/environment_variables.mdx（模块 文档；类别 docs；类型 documentation）：文档配套：说明两种 profile 可接受的 effort 档位，并给出通过 --json-model-override-args 覆盖 profile 的示例。

关键符号：_detect_dsv4_reasoning_effort_profile, _validate_dsv4_reasoning_effort_profile, resolve_dsv4_reasoning_effort_profile, render_message, encode_messages

关键源码片段

python/sglang/srt/entrypoints/openai/encoding_dsv4.py

编码核心：REASONING_EFFORT_PROFILES 定义 preview/official 两套 effort 映射，render_message 与 encode_messages 按 profile 选择前缀并改为显式 ValueError。

```
# 两代 prompt 契约的前缀文案：
# preview 的 max 沿用 "Absolute maximum"；official 的 max 改用 "Beyond maximum"
REASONING_EFFORT_PREVIEW_MAX = (
    "Reasoning Effort: Absolute maximum with no shortcuts permitted.\n"
    "You MUST be very thorough in your thinking and comprehensively decompose the problem
    to resolve the root cause, rigorously stress-testing your logic against all potential paths,
    edge cases, and adversarial scenarios.\n"
    "Explicitly write out your entire deliberation process, documenting every intermediate step,
    considered alternative, and rejected hypothesis to ensure absolutely no assumption is left
    unchecked.

"
)

REASONING_EFFORT_OFFICIAL_MAX = (
    "Reasoning Effort: Beyond maximum — exhaustive, relentless, and uncompromising.\n"
    "You MUST reason with the utmost depth and rigor, leaving absolutely nothing to chance:
    exhaustively decompose the problem into its most fundamental components, trace every
    causal chain to its root, and resolve the underlying cause rather than any surface symptom.
    \n"
    "Do not stop reasoning until you have independently verified the solution from multiple
    angles and are certain that no assumption remains unchecked and no error remains
    undiscovered.

"
)

# profile -> effort 档位 -> 前缀文案。
# 两个 profile 的 high 档都复用 preview 的 "Absolute maximum" 文案，
# 只有 max 档在 official 下换成 "Beyond maximum"。
REASONING_EFFORT_PROFILES = {
    "preview": {
        "high": "",
        "max": REASONING_EFFORT_PREVIEW_MAX,
    },
    "official": {
        "low": "",
        "high": REASONING_EFFORT_PREVIEW_MAX,
        "max": REASONING_EFFORT_OFFICIAL_MAX,
    },
}


def render_message(
```

```
    index: int,
    messages: List[Dict[str, Any]],
    thinking_mode: str,
    drop_thinking: bool = True,
    reasoning_effort: Optional[str] = None,
    reasoning_effort_profile: str = "preview",
) -> str:
    ...
    # 校验 profile 合法性后取出当前档位映射
    if reasoning_effort_profile not in REASONING_EFFORT_PROFILES:
        raise ValueError(
            f"Invalid reasoning effort profile: {reasoning_effort_profile!r}; "
            f"expected one of {list(REASONING_EFFORT_PROFILES)}"
        )
    effort_prompts = REASONING_EFFORT_PROFILES[reasoning_effort_profile]
    # 未显式指定 effort 时：official 默认 low（无前缀），preview 默认 high（无前缀）
    if reasoning_effort is None:
        reasoning_effort = "low" if reasoning_effort_profile == "official" else "high"
    if reasoning_effort not in effort_prompts:
        raise ValueError(
            f"Invalid reasoning effort {reasoning_effort!r} for profile "
            f"{reasoning_effort_profile!r}; expected one of {list(effort_prompts)}"
        )
    # 前缀只在 thinking 模式的首条消息前插入
    if index == 0 and thinking_mode == "thinking":
        prompt += effort_prompts[reasoning_effort]
    # 后续按 role 渲染 system / developer / user / assistant 消息
```

评论区精华

> JustinTong0323："We are working in parallel 🤔 there are some collisions, may you mind I push mine and then you can continue?"
>
> mmangkad："@JustinTong0323 that should be my last commit, just fixed UT failure (feel free to overwrite)"

两人最初并行开发同一功能，JustinTong0323 在冲突后接管后续提交（检测重构、启动时解析、preview/official 重命名），mmangkad 最终负责端到端验证并 LGTM。

> mmangkad："I verified this end to end on a B300 with both deepseek-ai/DeepSeek-V4-Flash-0731 and deepseek-ai/DeepSeek-V4-Flash. The official profile correctly follows the new low / high / max mapping, while the preview profile preserves the existing high / max behavior and falls back from low to high. The prompt-token counts matched the expected effort prefixes, and all 86 serving-chat unit tests passed."
>
> mmangkad："I like the preview / official rename too, especially with the official Pro release in mind, since it likely will not use the Flash-specific 0731 date."

评审确认了 profile 按能力形态命名（而非按发布时间）的取舍，为后续 DSV4 Pro 预留了扩展空间。

- 并行开发冲突与接力 (other): JustinTong0323 接管后续提交（检测重构、启动期解析、preview/official 重命名），mmangkad 最终完成端到端验证并 LGTM。
- B300 端到端验证两种 profile (testing): 验证通过，确认两代契约行为符合预期且无回归。
- preview/official 命名取舍 (design): 采用 preview/official 命名，为后续 DSV4 Pro 发布预留扩展空间。

风险与影响

- 风险：
  1. chat_encoding.py 的检测依赖 AST 静态解析：若官方 encoder 改为条件赋值、函数返回或文件超过 1 MiB，检测将回退到 preview，Silently 使用旧契约（不崩溃但可能不符合预期）。
  2. 启动期网络依赖：model_path 非本地路径时 hf_hub_download 会阻塞启动流程，Hub 不可达、revision 缺失或仓库未包含该文件都会走回退分支。
  3. 前缀缓存与行为变化：同一对话在 preview（high 无前缀）与 official（low 无前缀但 max 前缀不同）下产生不同 token 前缀，radix cache 跨契约不共享，可能影响 prefix cache 命中率；preview 下显式请求 low 会被过滤为 None 并静默回退到 high。
  4. override 依赖 hf_config.to_dict() 保留 dsv4_reasoning_effort_profile 键，若模型配置转换处理丢失该字段，转换后的 checkpoint 无法显式指定 profile。
  5. serving_chat.py 中使用 assert reasoning_effort_profile is not None，在 python -O 模式下断言失效（当前构造期已保证非 None，风险较低）。 - 影响：用户侧：DeepSeek-V4-Flash-0731 用户自动获得官方 low/high/max 三档 effort 语义，旧版 DeepSeek-V4-Flash 用户行为保持不变（向后兼容）；OpenAI 兼容层新增 dsv4_reasoning_effort_profile 模型配置覆盖点，转换的 checkpoint 可自声明契约。系统侧：启动流程新增一次最多 1 MiB 的 encoder 文件读取与 AST 解析；请求编码路径增加一次按 profile 过滤 effort 的字典查找，开销可忽略。团队侧：为 DSV4 Pro 或后续版本预留了 checkpoint 自描述 + 显式 override 双通道，降低模型发布时间表对 SGLang 代码的耦合。 - 风险标记：启动期 Hugging Face Hub 网络依赖，AST 检测对 checkpoint 结构敏感，profile 切换影响前缀缓存命中，preview 档 low 请求静默回退 high，override 依赖模型配置保留字段

关联脉络

- PR #28040 [Intel GPU] DeepSeek V4 8/N: use sgl-kernel implementation of fused_k_norm_rope_flashmla on XPU: 同一 DeepSeek V4 功能线在 XPU 侧的内核落地，说明 DSV4 支持正跨后端（Blackwell、XPU）铺开，本 PR 则补齐 prompt 契约层。
- PR #33605 [CI] Make B200 base-b suites single-GPU as prep for 1-gpu B200 runners: DeepSeek V4/FP4 相关测试在 B200 CI 上的矩阵调整，与本 PR 同属 DSV4 发布与测试基建演进。

- PR #33611 [Test] Replace NVFP4 MoE runner backend e2e matrix with a layer-level unit test: DeepSeek 量化后端测试从 e2e 矩阵转向层单元测试，与本 PR 一样倾向用单元测试覆盖模型特定行为。
- PR #32541 [Kimi] Support kimi-k3: Kimi K3 Day-0 支持同样扩展了 chat 编码分发路径（resolve_chat_encoding_spec 增加 kimi_k3），与本 PR 共用同一套编码解析框架。